NEW APPROACHES IN ENGINEERING:

THEORY, METHOD, AND PRACTICE

Editors

Assist. Prof. Sıtkı Alper ÖZDEMİR

Assist. Prof. Gülsade KALE

DUVAR

NEW APPROACHES IN ENGINEERING:

THEORY, METHOD, AND PRACTICE

Editors

Assist. Prof. Sıtkı Alper ÖZDEMİR

Assist. Prof. Gülsade KALE

NEW APPROACHES IN ENGINEERING:
THEORY, METHOD, AND PRACTICE
Editors: Assist. Prof. Sıtkı Alper ÖZDEMİR, Assist. Prof. Gülsade KALE

Editor in chief: Berkan Balpetek
Cover and Page Design: Duvar Design
Printing: December 2025
Publisher Certificate No: 49837
ISBN: 978-625-8698-56-5

853 Sokak No:13 P.10 Kemeraltı-Konak/İzmir
Tel: 0 232 484 88 68
www.duvaryayinlari.com
duvarkitabevi@gmail.com

TABLE OF CONTENTS

Chapter 1 1
Fundamentals of Gaussian Process Regression: Theory and
Python Applications
Ali ŞENTÜRK

Chapter 2 26
Electrical Fault Analysis of Total-Cross-Tied Photovoltaic Array System:
Simulation and Performance Analysis
Burcin OZKAYA, Evren ISEN

Chapter 3 40
A Comparison of Hydrometeorological Relationship in the Ceyhan Basın,
Turkiye Using Linear and Non-linear Regression Analysis Based on Long-Term
Observation Data (1980-2020)
Evren TURHAN, Serin DEĞERLİ ŞİMŞEK

Chapter 4 55
Nutritional Composition and Digestibility of Wheat Straw
Hüseyin NURSOY

Chapter 5 63
Use of Wheat Straw in Cattle Rations
Hüseyin NURSOY

Chapter 6 72
The Discrepancy between Finite Element Results and Experimental Tests:
Roots and Solutions
İzel EKİNCİ, Daver ALİ

Chapter 7 94
Legal Transformation and Changes in Implementation Policies in Building
Inspection Legislation
Sıtkı Alper ÖZDEMİR, Mehmet UZUN

Chapter 8 108
Large Language Models in Economics: A Chronological Review and Conceptual Assessment
Aytürk KELEŞ, Ali KELEŞ

Chapter 9 118
Approaches for Lithium Ion Batteries
Mustafa AKKAYA

Chapter 10 132
Evolution of Mobile Communication Technologies from 3G to 6G:
Architectural Transformation, Performance Analysis,
and Simulation-Based Approach
Ali Özhan AKYÜZ, Kazım KUMAŞ, Mustafa AYAN,
Durmuş TEMİZ

Chapter 11 157
K-Fold Cross Validation Approach in Image-Based Cnn Studies
Olcay PALTA, Seda Yetkin YEŞİL

Chapter 12 174
A Comprehensive Overview to Power Quality Issues in Smart Grids
Harun ÖZBAY, Serhat Berat EFE

Chapter 13 186
Artificial Intelligence Pro mpt Writing:Theory and Practice
Rabia KORKMAZ TAN

Chapter 14 211
Synthesis of Copper Incorporated Porous Clay Heterostructures (PCH) and FTIR Study
Funda Turgut BASOGLU

Chapter 15 ....**232**
Evaporation Prediction Using Machine Learning Models Enhanced with
cVAE–cWGAN-GP Based Synthetic Data Generation
Özlem TERZİ, Emine Dilek TAYLAN

Chapter 16 ....**254**
Streamflow Prediction Using Multi-Frequency
Analysis of Time Series: A Wavelet–LSTM Model
Emine Dilek TAYLAN, Özlem TERZİ

Chapter 1

Fundamentals of Gaussian Process Regression: Theory and Python Applications

Ali ŞENTÜRK[1]

Abstract

Gaussian Process Regression (GPR) is a probabilistic regression method based on conditional probability theory and multivariate Gaussian distributions. In GPR, not only an estimate of an unknown function is obtained from observations, but also the uncertainty associated with this estimate can be quantified. GPR is particularly effective in problems where the available data are limited.

In this chapter, the mathematical foundations of Gaussian Process Regression (GPR) are presented, starting from the conceptual principles of the underlying statistical theory. In particular, Gaussian distributions, the covariance matrix and kernel functions used in Gaussian processes, the derivation of GPR prediction equations through joint and conditional distributions, and the use of Cholesky decomposition instead of direct matrix inversion are explained step by step.

Following the theoretical exposition, a Python implementation based on the fundamental GPR equations is developed. Through this implementation, essential numerical operations such as the computation of the predictive mean, uncertainty (variance), and Cholesky decomposition are demonstrated. Subsequently, the same regression problem is revisited using the scikit-learn library. The results obtained from the GPR module provided by this library are compared with those of the custom implementation, and the impact of hyperparameter optimization is explicitly illustrated.

Keywords: Gaussian Process Regression, RBF Kernel, Covariance Matrix, Uncertainty Estimation.

---

[1] Isparta University of Applied Sciences, Faculty of Technology, Department of Electrical and Electronics Engineering, Isparta, Turkiye. alisenturk@isparta.edu.tr, ORCID: 0000-0002-5868-7365.

Introduction

One of the fundamental objectives of machine learning is to produce the most appropriate prediction for an unobserved situation by utilizing available data. When this task involves continuous-valued variables, it is referred to as a regression problem. For example, given two observed data points (x_1, y_1) and (x_2, y_2), the goal is to predict the output y_3 corresponding to a new input x_3. In the simplest approach, these two data points can be connected linearly to obtain a prediction for y_3.

However, real-world data rarely exhibits purely linear behavior. The underlying relationships may be exponential, periodic, or described by high-degree polynomial structures; moreover, the data may be noisy or exhibit highly complex patterns. In such cases, fitting a single curve alone may not be sufficient. It also becomes necessary to quantify the degree of uncertainty associated with the prediction.

Gaussian Process Regression (GPR), unlike classical regression methods, not only provides a point estimate but also offers a quantitative measure of uncertainty associated with this estimate. As a result, predictive uncertainty typically increases in regions where data are sparse, while it tends to decrease in regions where data are dense. Nevertheless, even in densely populated regions, the variability among observations plays a crucial role in determining the level of predictive uncertainty.

Gaussian Process Regression is supported by a strong and comprehensive mathematical framework. For an in-depth treatment of GPR and its theoretical foundations, the seminal work by Rasmussen and Williams [1], which is widely regarded as a fundamental reference in this field, can be consulted. In addition, reference [2] presents an intuitive explanation of GPR, supported by numerous visual examples, and is therefore also recommended. In this chapter, the theoretical expressions underlying GPR are presented as clearly as possible, the fundamental concepts on which the theory is built are explained, and both numerical and visual examples are provided through Python implementations in order to facilitate a deeper understanding of the theoretical and practical aspects of GPR.

Fundamental Concepts

Since GPR is a method grounded in statistical theory, fundamental concepts such as probability distributions, expectation, variance, and standard deviation are briefly reviewed in this section.

A probability distribution is a mathematical function that describes the possible values of a random variable along with the probabilities associated with those values. For instance, when a fair die is rolled, the probability of obtaining any number between 1 and 6 on the top face is equal. Because all outcomes have the same probability, this distribution is referred to as a *uniform distribution*. This distribution is illustrated in Table 1.

x	1	2	3	4	5	6
Pr(X =x)	1/6	1/6	1/6	1/6	1/6	1/6

Table 1. Uniform probability distribution associated with rolling a die

The random variable X may be either discrete or continuous. When the random variable X follows a uniform distribution over the interval $[a, b]$, it is expressed as

$$X \sim \mathcal{U}(a, b) \tag{1}$$

Here, the symbol "∼" indicates that the random variable X is sampled from the specified probability distribution.

For continuous random variables, probabilities are not defined at individual points but over intervals. The function used for this purpose is called the *probability density function* (pdf). The pdf of a random variable X that is uniformly distributed over the interval $[a, b]$ is given by

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & otherwise \end{cases} \tag{2}$$

Using the Python program provided in Program 1, random samples generated from a uniform distribution with parameters $a = 0$ and $b = 10$, along with the corresponding theoretical probability density function (pdf), are visualized. The resulting plot obtained by executing the program is shown in Figure 1.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import uniform

m = 1000       # Number of samples
a, b = 0, 10   # Interval of the uniform distribution

# Uniform random samples
f_random = np.random.uniform(low=a, high=b, size=m)

```

```
# x-axis for the PDF
x = np.linspace(a - 2, b + 2, 400)
pdf = uniform.pdf(x, loc=a, scale=b - a)

# Visualization
plt.figure(figsize=(7, 4))
# Plot histogram
plt.hist(
    f_random,
    bins=10,
    density=True,
    alpha=0.6,
    color="orange",
    edgecolor="black",
    label="Uniform samples")

# Theoretical uniform PDF curve
plt.plot(x, pdf, "r-", linewidth=2, label="Uniform PDF (theoretical)")

plt.xlabel(r"$x$", fontsize=16)
plt.ylabel(r"$P(x)$", fontsize=16)
plt.title("1-D Uniform Distribution — 1000 Random Samples", fontsize=14)
plt.legend()
plt.grid(True)
plt.show()
```

Program 1. Normalized histogram of 1000 random variables sampled from a uniform distribution over the interval [0,10], together with the corresponding theoretical probability density function.

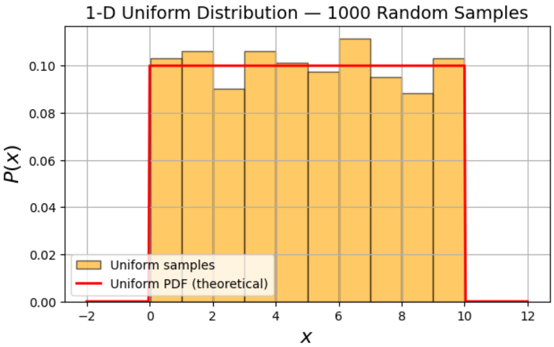


Figure 1. Output obtained from Program 1: the normalized histogram of 1000 random variables sampled from a uniform distribution over the interval [0,10], along with the corresponding theoretical probability density function.

Using the probability density function, the probability that a random variable takes a value within a given interval can be computed as

$$\mathrm{Pr}(\mathrm{a}_1 \le \mathrm{x} \le \mathrm{b}_1) = \int_{a_1}^{b_1} f_X(x)\,\mathrm{dx} \tag{3}$$

For example, for the uniform distribution shown in Figure 1, the probability $P(0 \le x \le 1)$is equal to 0.1.

The mean, or expected value, of a continuous probability distribution is computed using the probability density function $f_X(x)$as

$$\mu_X = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx \tag{4}$$

For example, for the uniform distribution shown in Figure 1, the probability $P(0 \le x \le 1)$is equal to 0.1.

The mean, or expected value, of a continuous probability distribution is computed using the probability density function $f_X(x)$as

$$\mathrm{Var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x) dx \tag{5}$$

The standard deviation σ_X, which is the square root of the variance, expresses the dispersion of the distribution on a more interpretable scale.

Gaussian (Normal) Distribution

The Gaussian distribution is one of the most important distributions in probability theory, as it successfully models many phenomena frequently encountered in real life, such as measurement errors, physical variables, and biological processes. If a random variable X follows a normal distribution with mean μ and variance σ^2, it is denoted as

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{6}$$

This notation indicates that the random variable X is sampled from a normal distribution with mean μ and variance σ^2.

The probability density function of the Gaussian distribution is given by

$$\mathrm{f}_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{7}$$

In Program 2, the probability density functions of two Gaussian distributions with mean zero and standard deviations of 1 and 3 are visualized. The resulting plot obtained from the program execution is shown in Figure 2.

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# x-axis
x = np.linspace(-10, 10, 400)

# Two different standard deviations
y1 = norm.pdf(x, loc=0, scale=1)
y2 = norm.pdf(x, loc=0, scale=3)

plt.figure(figsize=(7,4))
plt.plot(x, y1, label="σ = 1", linewidth=2)
plt.plot(x, y2, label="σ = 3", linewidth=2)

plt.title("Effect of Standard Deviation on the Gaussian Distribution")
plt.xlabel("x")
plt.ylabel("f(x)")
plt.grid(True)
plt.legend()
plt.show()
```

Program 1. Probability density functions of the distributions $\mathcal{N}(0, 1^2)$ and $\mathcal{N}(0, 3^2)$.

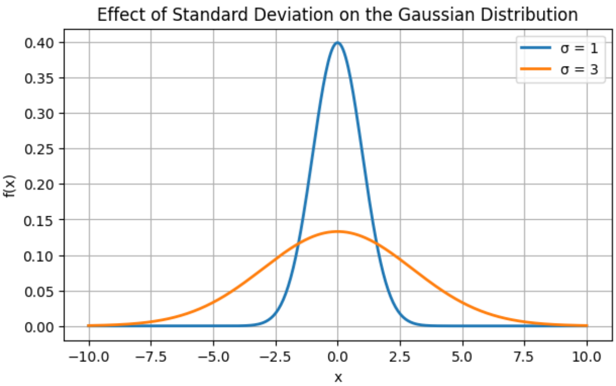


Figure 2. Probability density functions of Gaussian distributions obtained as the output of Program 2.

Since the Gaussian distributions shown in Figure 2 represent probability distributions, the total area under each curve is equal to 1. One of the most important properties of the Gaussian distribution is that a large portion of the probability mass is concentrated around the mean. This property can be approximately expressed as

$$\begin{gathered}\Pr(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68 \\ \Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95 \\ \Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997\end{gathered} \tag{8}$$

Accordingly, the interval extending up to 3σ from the mean covers almost all of the probability mass.

Multivariate Gaussian Distribution

In many practical applications, multiple variables are considered jointly. For example, variables such as height–weight, temperature–humidity, or age–weight are often correlated with each other. When such variables follow a normal distribution and are analyzed jointly, the resulting distribution is referred to as a *multivariate Gaussian (normal) distribution*.

If an n-dimensional random vector $\mathbf{X}$ follows a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, this is denoted as

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{9}$$

Here, $\mathbf{X} \in \mathbb{R}^n$, $\boldsymbol{\mu} \in \mathbb{R}^n$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is a symmetric and positive definite covariance matrix.

The random vector and the mean vector can be written as

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}. \tag{10}$$

One of the key components of the multivariate Gaussian distribution is the covariance matrix. The diagonal elements of this matrix represent the variances of the individual variables, while the off-diagonal elements describe the linear relationships between pairs of variables. For two continuous random variables X_i and X_j, the covariance is defined as

$$\mathrm{Cov}(X_i, X_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x_i - \mu_i)(x_j - \mu_j) f_{X_i X_j}(x_i, x_j)\, dx_i dy_i \tag{11}$$

where $f_{X_i X_j}$denotes the joint probability density function of the variables. Based on this definition, the covariance matrix can be written as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \sigma_2^2 & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \sigma_n^2 \end{bmatrix} \tag{12}$$

As shown in Equation (12), the diagonal elements satisfy $\boldsymbol{\Sigma}_{ii} = \sigma_i^2$, meaning that each diagonal entry corresponds to the variance of the i-th variable. Moreover, since $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ij}$, the covariance matrix is symmetric.

The probability density function of the multivariate Gaussian distribution is given by

$$f_X(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}}\, exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right) \tag{13}$$

Here, $|\mathbf{\Sigma}|$ denotes the determinant of the covariance matrix, $(\mathbf{x} - \boldsymbol{\mu})^\top$ represents the transpose of the vector $\mathbf{x} - \boldsymbol{\mu}$, and $\mathbf{\Sigma}^{-1}$ is the inverse of the covariance matrix.

In Program 3, the probability density function of a two-dimensional Gaussian distribution is computed and visualized using contour plots (iso-density curves). When the off-diagonal elements of the covariance matrix are nonzero, this indicates the presence of a linear relationship between the variables. As a result, the contour lines exhibit an elliptical structure rather than being aligned with the coordinate axes. In Figure 3, the probability density function of the multivariate Gaussian distribution defined for the two-dimensional random vector $\mathbf{X} = [X_1, X_2]^\top$ is illustrated using iso-density contours in the x_1–x_2 plane.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import multivariate_normal

# Mean vector
mu = np.array([0, 0])

# Covariance matrix
Sigma = np.array([[1.0, 0.8],
                  [0.8, 1.5]])

# Points at which the PDF will be evaluated
X_1 = np.linspace(-4, 4, 200)
X_2 = np.linspace(-4, 4, 200)
X1, X2 = np.meshgrid(X_1, X_2)

# Reshape points to (200x200, 2)
pos = np.dstack((X1, X2))

# Compute the multivariate Gaussian PDF
rv = multivariate_normal(mu, Sigma)
Z = rv.pdf(pos)

# Contour plot
plt.figure(figsize=(7, 5))
cont = plt.contour(X1, X2, Z, levels=10, cmap="viridis")
plt.clabel(cont, inline=True, fontsize=8)
plt.title("2D Gaussian Contour Plot")
plt.xlabel(r'$x_1$')
plt.ylabel(r'$x_2$')
plt.grid(True)
plt.show()
```

Program 3. Computation of the probability density function of a two-dimensional multivariate Gaussian distribution and its visualization using contour plots.

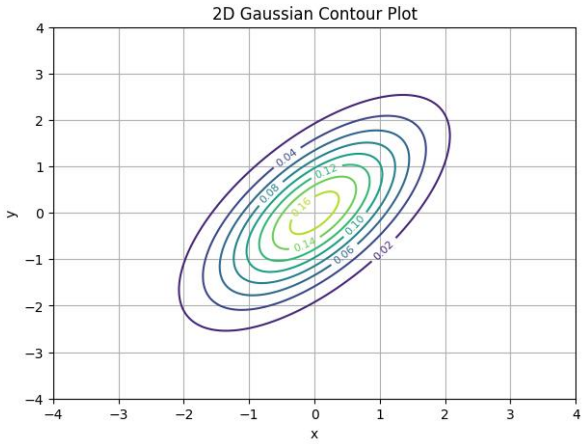


Figure 3. Iso-density contours of the two-dimensional Gaussian distribution obtained as the output of Program 3.

In Gaussian processes, the fact that the joint distribution of any finite number of random variables is multivariate Gaussian constitutes the fundamental theoretical basis of Gaussian Process Regression.

Radial Basis Kernel Function

In a simple regression approach, given two data points (x_1, y_1) and (x_2, y_2), the relationship between them is often assumed to be linear, and the predicted value $\hat{y}$ corresponding to a new input x_* can be obtained under this assumption. However, real-world data are rarely linear, and the underlying function is generally continuous. In other words, abrupt and discontinuous changes are not expected when transitioning to a third data point. Moreover, the closer the input x_* is to the existing data points, the smaller the expected prediction error.

In Gaussian Process Regression (GPR), the mean function $\boldsymbol{\mu}$ is most commonly assumed to be zero [2]. Under this assumption, the behavior of the model is primarily determined by the covariance matrix $\boldsymbol{\Sigma}$. The covariance structure itself is defined through kernel functions. Consequently, kernel functions constitute the most fundamental components that determine the flexibility, smoothness, and overall behavior of a GPR model.

One of the most widely used kernel functions is the squared exponential, also known as the *Radial Basis Function (RBF)* kernel [3]. Other commonly used kernel functions are discussed in reference [4], while more advanced usages of kernel functions are presented in reference [5]. The RBF kernel measures the similarity between two data points and is fundamentally based on the Euclidean distance between them. In its simplest form, the RBF kernel can be expressed as

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\left(x_i - x_j\right)^2}{2}\right) \tag{14}$$

To control the behavior of the kernel function in a more flexible manner, two hyperparameters are introduced: the length scale l and the signal variance σ_f^2. With these parameters, the RBF kernel is defined as

$$k\left(x_i, x_j\right) = \sigma_f^2 \, exp\left(-\frac{\left(x_i - x_j\right)^2}{2l^2}\right) \tag{15}$$

In Equation (15), the length scale l controls how rapidly the function varies in the input space. Small values of l cause even nearby points to be weakly correlated, resulting in functions that vary rapidly and exhibit more pronounced curvature. In contrast, larger values of l lead to smoother and more slowly varying functions. The parameter σ_f^2 determines the vertical scale of the function. Smaller values correspond to a narrower range of outputs, whereas larger values indicate a wider range of uncertainty.

The parameters l and σ_f^2 are also the *hyperparameters* of the GPR model, and their appropriate values are typically learned from data. As a result, the RBF kernel produces high similarity (and thus high covariance) when two data points are close to each other, and low similarity when they are far apart. Using this similarity measure, the relationships among all data points are computed. Accordingly, the covariance matrix used in a GPR model is constructed as

$$Cov\left(x_i, x_j\right) = k\left(x_i, x_j\right) \tag{16}$$

Gaussian Process Regressor

In a Gaussian process, the similarity between points in the input space is defined using a kernel function $k(x, x')$. The kernel function determines the covariance between the function values at these points. When the kernel function is applied to the observed inputs $\boldsymbol{X}$, a covariance matrix **K** is obtained, which represents the covariance structure of the function values associated with these observations.

Rather than selecting a single function, a Gaussian process defines a *prior* over functions by assuming that the values of the function at any finite set of input points jointly follow a multivariate Gaussian distribution. This assumption can be expressed as

$$\boldsymbol{f}(\boldsymbol{X}) \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{X}), \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})\right) \tag{17}$$

Here, $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]^T$ denotes the vector of n observed input values, $\boldsymbol{\mu}(\boldsymbol{X}) = [\mu(x_1), \mu(x_2), \dots, \mu(x_n)]^{\mathrm{T}}$ represents the values of the mean function at these points, and $\mathbf{K} = (k(x_i, x_j))_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$ is the covariance matrix of the function values at these inputs. In practice, the mean function is typically chosen as $\mu(x_i) = 0$ [2]. Under this assumption, the behavior of the model is entirely governed by the covariance matrix constructed via the kernel function.

Let $\boldsymbol{X}_* \in \mathbb{R}^{n_*}$ denote a vector of test input points. For these points, an expression analogous to Equation (17) can be written as

$$\boldsymbol{f}_*(\boldsymbol{X}_*) \sim \mathcal{N}\big(\boldsymbol{0}, \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*)\big) \tag{18}$$

Let $\boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}) \in \mathbb{R}^{n_* \times n}$, $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*) \in \mathbb{R}^{n \times n_*}$ and $\boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})^T = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*)$. Then, the joint Gaussian distribution of the function values at the training and test points can be written as

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{X}) \\ \boldsymbol{f}_*(\boldsymbol{X}) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*) \\ \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix}\right) \tag{19}$$

Using the marginal and conditional distribution properties of the multivariate Gaussian distribution, the conditional distribution of the function values at the test points is obtained as

$$\begin{aligned} \boldsymbol{f}_* | \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{X}_* \sim \mathcal{N}(\boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{y}, \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*) - \\ \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*) \end{aligned} \tag{20}$$

where $\mathbf{y}$ denotes the observed values corresponding to the input vector $\boldsymbol{X}$ [1].

In real-world problems, observations are typically noisy. Each observation is modeled as

$$y_i = f(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_n^2) \tag{21}$$

Under this noise model, Equation (20) becomes

$$\begin{aligned} \boldsymbol{f}_* | \boldsymbol{f}, \boldsymbol{X}, \boldsymbol{X}_* \sim \mathcal{N}(\boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\, \boldsymbol{I}]^{-1}\boldsymbol{y}, \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*) - \\ \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\, \boldsymbol{I}]^{-1}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*) \end{aligned} \tag{22}$$

Accordingly, the predictive mean $\hat{\mathbf{y}}_*$ and the predictive covariance at the test points are given by

$$\hat{\boldsymbol{y}}_* = \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\, \boldsymbol{I}]^{-1}\boldsymbol{y} \tag{23}$$

$$Cov(\boldsymbol{y}_*) = \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X}_*) - \boldsymbol{K}(\boldsymbol{X}_*, \boldsymbol{X})[\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\, \boldsymbol{I}]^{-1}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}_*) \tag{24}$$

Directly computing the inverse of the matrix $\mathbf{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\mathbf{I}$ has a computational complexity of $O(n^3)$and may lead to numerical instability. This becomes particularly problematic when the number of data points is large. Therefore, to compute $\hat{\mathbf{y}}_*$ efficiently and stably, a Cholesky decomposition–based approach is employed.

For notational simplicity, define the noise-augmented covariance matrix as

$$\boldsymbol{K}_n = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2\, \boldsymbol{I} \tag{25}$$

and let

$$\boldsymbol{K}_n^{-1}\boldsymbol{y} = \alpha \tag{26}$$

which can equivalently be written as

$$\boldsymbol{y} = \frac{\alpha}{\boldsymbol{K}_n^{-1}} = \boldsymbol{K}_n\alpha \tag{27}$$

Since $\mathbf{K}_n$ is symmetric and positive definite, it admits a Cholesky decomposition:

$$\boldsymbol{K}_n = \boldsymbol{L}\boldsymbol{L}^{\top} \tag{28}$$

where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{L}^{\top}$ is an upper triangular matrix. Substituting into Equation (27) yields

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{L}^{\top}\alpha \tag{29}$$

Define

$$\mathbf{L}^{\top}\alpha = z \tag{30}$$

which leads to

$$\boldsymbol{y} = \boldsymbol{L}z \tag{31}$$

Rather than explicitly computing the inverse of $\mathbf{L}$, the linear system is solved as

$$z = \boldsymbol{L}\backslash\boldsymbol{y} \tag{32}$$

Once $\mathbf{z}$ is obtained, $\boldsymbol{\alpha}$ is computed by solving

$$\alpha = \boldsymbol{L}^T\backslash z \tag{33}$$

Accordingly, Equation (23) becomes

$$\hat{\boldsymbol{y}}_* = \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\alpha \tag{34}$$

For the predictive covariance in Equation (24), the inverse of $\mathbf{K}_n$ is expressed via the Cholesky factors as

$$\boldsymbol{K}_n^{-1} = \left(\boldsymbol{L}\boldsymbol{L}^{\top}\right)^{-1} = \left(\boldsymbol{L}^{\top}\right)^{-1}\boldsymbol{L}^{-1} \tag{35}$$

For notation simplicity, let $\mathbf{K}(X,X_*) = \mathbf{K}_*$, accordingly $\mathbf{K}(X_*,X) = \mathbf{K}_*^{\top}$. With this notation, the corresponding term in Equation (24) can be written as

$$\begin{gathered}\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\,\boldsymbol{I}]^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*) = \boldsymbol{K}_*^{\top}\boldsymbol{K}_n^{-1}\boldsymbol{K}_* = \\ \left(\boldsymbol{K}_*^{\top}\left(\boldsymbol{L}^{\top}\right)^{-1}\right)\left(\boldsymbol{L}^{-1}\boldsymbol{K}_*\right)\end{gathered} \tag{36}$$

Define $v = \boldsymbol{L}^{-1}\boldsymbol{K}_*$ which implies $\mathbf{L}\mathbf{v} = \mathbf{K}_*$. Hence, the vector $\mathbf{v}$ is obtained by solving the linear system

$$v = \boldsymbol{L}\backslash\boldsymbol{K}_* \tag{37}$$

Using the transpose identity for matrix products,

$$\left(\boldsymbol{K}_*^{\top}\left(\boldsymbol{L}^{\top}\right)^{-1}\right) = \left(\boldsymbol{L}^{-1}\boldsymbol{K}_*\right)^{\top} = v^{\top} \tag{38}$$

Therefore, Equation (36) simplifies to

$$\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\,\boldsymbol{I}]^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*) = \boldsymbol{K}_*^{\top}\boldsymbol{K}_n^{-1}\boldsymbol{K}_* = v^T v \tag{39}$$

As a result, Equation (24) is finally evaluated as

$$Cov(\boldsymbol{f}_*) = \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) - v^T v \tag{40}$$

In this section, the mathematical foundations of Gaussian Process Regression have been presented. In Program 4, a Python implementation is provided to demonstrate how these theoretical expressions are applied in practice. In the program, eight angles randomly selected from the interval 0°–360° are used as training data, while a ninth angle randomly selected from the same interval is used as a test input.

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.linalg import cholesky, solve_triangular

# Generate 8 random angles between 0 and 360 degrees for training
np.random.seed(21)
X_deg = np.sort(np.random.uniform(0.0, 360.0, size=8))
X = np.deg2rad(X_deg).reshape(-1, 1)
y = np.sin(X).reshape(-1)

# Generate 1 random angle for testing
x_star_deg = np.random.uniform(0.0, 360.0)
x_star = np.deg2rad(x_star_deg).reshape(1, 1)
y_true = np.sin(x_star).item()

# RBF kernel function
def rbf(a, b, lengthscale=0.6, variance=1.0):
    # a: (n,1), b: (m,1) -> output: (n,m)
    d2 = (a.reshape(-1, 1) - b.reshape(1, -1))**2
    return variance * np.exp(-0.5 * d2 / (lengthscale**2))

# Hyperparameters
lengthscale = 0.6
variance = 1.0
sigma_n2 = 0.005

# Covariance matrices
K = rbf(X, X, lengthscale, variance)
K_star = rbf(X, x_star, lengthscale, variance)
K_ss = rbf(x_star, x_star, lengthscale, variance)

```

```
# Cholesky decomposition
K_n = K + sigma_n2 * np.eye(len(X))
L = cholesky(K_n, lower=True)

# Compute alpha
z = solve_triangular(L, y, lower=True)
alpha = solve_triangular(L.T, z, lower=False)

# Prediction
y_star_mean = (K_star.T @ alpha).item()

# Predictive variance
v = solve_triangular(L, K_star, lower=True)
f_star_var = (K_ss - v.T @ v).item()

print("Training angles (degrees):", np.round(X_deg, 2))
print("Test angle (degrees):", round(x_star_deg, 2))
print("GPR predictive mean f*(x*) =", round(y_star_mean, 5))
print("True value f(x*) =", round(y_true, 5))
print("GPR predictive variance =", round(f_star_var, 5))
```

Program 4. Construction of a GPR model using eight training points randomly selected from the interval 0°–360°, and computation of the predictive mean and uncertainty (variance) for a test point randomly selected from the same interval.

The output obtained by running Program 4 is as follows:

```
Training angles (degrees): [7.78 17.54 18.28 74.13 104.08 108.82 239.01 259.55]
Test angle (degrees): 110.92
GPR predictive mean f*(x*) = 0.92333
True value f(x*) = 0.93407
GPR predictive variance = 0.00664
```

It may be necessary to examine not only the prediction at a single test point, but also how the GPR model behaves over a specified interval, in order to assess both predictive performance and how the uncertainty varies across the input domain. In Program 5, 300 equally spaced points selected over the interval 0°–360° are treated as test inputs, and the GPR predictions over this range are illustrated. In addition, a confidence band representing the predictive uncertainty is chosen as $\pm 2\sigma$ and visualized accordingly.

```
# Values in the 0–360 degree range for plotting
x_grid_deg = np.linspace(0.0, 360.0, 300)
X_grid = np.deg2rad(x_grid_deg).reshape(-1, 1)

# Covariances on the grid
K_grid = rbf(X, X_grid, lengthscale, variance)
K_gg = rbf(X_grid, X_grid, lengthscale, variance)

# Posterior mean
mu_grid = (K_grid.T @ alpha).reshape(-1)

# Posterior variance
v_grid = solve_triangular(L, K_grid, lower=True)
var_grid = np.diag(K_gg) - np.sum(v_grid**2, axis=0)
var_grid = np.maximum(var_grid, 0.0)
std_grid = np.sqrt(var_grid)

# True sine function
y_true_grid = np.sin(X_grid).reshape(-1)

plt.figure(figsize=(10, 5))
# True sin(x)
plt.plot(
    x_grid_deg,
    y_true_grid,
    linestyle="--",
    linewidth=1.8,
    label="sin(x) (true function)")

# GPR posterior mean
plt.plot(
    x_grid_deg,
    mu_grid,
    linewidth=2.2,
    label="GPR predictive mean")
```

```

# ±2σ confidence interval
plt.fill_between(
    x_grid_deg,
    mu_grid - 2 * std_grid,
    mu_grid + 2 * std_grid,
    alpha=0.25,
    label="Confidence band (±2σ)")

# Training points
plt.scatter(
    X_deg,
    y,
    s=70,
    zorder=3,
    label="Training points")

plt.xlabel("Angle (degrees)")
plt.ylabel(r"$f(x)$")
plt.title("Training Data, GPR Prediction, and Confidence Interval")
plt.grid(True)
plt.legend()
plt.show()
```

Program 5. Visualization of the GPR prediction and confidence interval over the range $0^\circ - 360^\circ$.

The plot obtained by running Program 5 is shown in Figure 4. As observed in Figure 4, GPR produces lower uncertainty in regions close to the training points, whereas the uncertainty increases in regions where the data density is low.

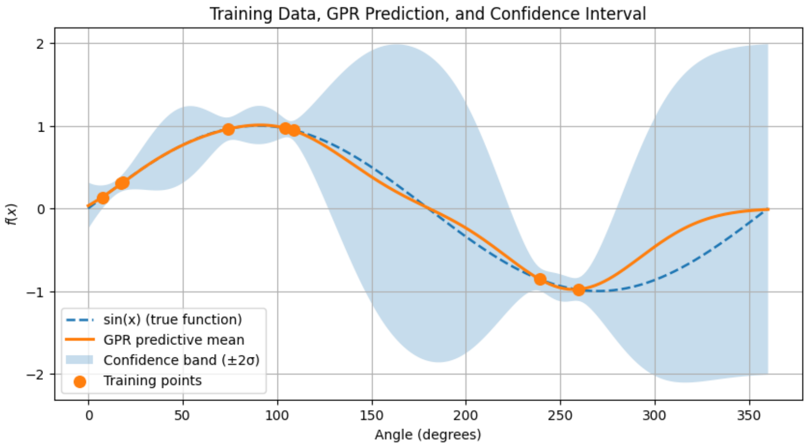


Figure 4. GPR prediction and confidence interval for $0° - 360°$.

Hyperparameter Optimization

The RBF kernel used in GPR has two hyperparameters: the signal variance σ_f^2 and the length scale l, as given in Equation (15). In addition, the noise variance σ_n^2, introduced in Equation (21) and incorporated into the conditional predictive distribution in Equation (22), is also a hyperparameter. Choosing these hyperparameters appropriately is crucial for GPR to model the observed data accurately.

In a GPR model, the *marginal likelihood* is employed to assess how well the model parameters explain the observed data by integrating over all possible function values. Taking the logarithm of this expression yields the *marginal log-likelihood*, which is then used for hyperparameter optimization. The marginal log-likelihood of a GPR model is given by [1]

$$\log p(\mathrm{y} \mid X) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_n^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}_n| - \frac{n}{2}\log 2\pi \tag{41}$$

When the partial derivative of this expression with respect to σ_f^2 is taken and set to zero, the following result is obtained [6]:

$$\sigma_f^2 = \frac{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_n^{-1}\boldsymbol{y}}{n} \tag{42}$$

Here, $\mathbf{K}_n$ denotes the covariance matrix constructed without the σ_f^2 term in Equation (15).

For the terms $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{K}_n^{-1}\boldsymbol{y}$ and $\log|\boldsymbol{K}_n|$ in Equation (41), the optimal values of l or σ_n^2 cannot be obtained by simply taking partial derivatives and setting them to zero. Instead, these hyperparameters are typically determined using numerical optimization methods [6].

GPR Implementation Using the scikit-learn Library

There are numerous libraries developed for Gaussian Process Regression (GPR) applications in the Python programming language. For example, the GPflow library is built on the TensorFlow framework and is particularly suitable for large-scale datasets and Gaussian process models with input dependent variations [7]. Similarly, GPyTorch is a PyTorch-based library [8]. Owing to their GPU support, both libraries can be effectively used for high-dimensional and large-scale problems.

Another GPR library is hetGP, which was originally developed for the R programming language and later released in a Python version. The hetGP library is specifically designed for Gaussian process models with input-dependent (heteroskedastic) variance. In addition, it also supports fixed-variance (homoskedastic) GPR models. Consequently, hetGP is well suited for applications in which measurement uncertainty plays a significant role [6].

The Gaussian Process module included in the scikit-learn library provides a simplified interface for implementing the fundamental principles of GPR [9]. This module is particularly suitable for basic GPR applications in small to medium-scale studies.

In this section, Program 6 follows the same experimental setup as Program 4: eight angles randomly selected from the interval 0°–360° are used as training data, while a ninth angle randomly selected from the same interval is used as a test input. The prediction for the test data is obtained using the Gaussian Process module provided by the scikit-learn library.

Similarly, in Program 7, again using the scikit-learn library, an experiment equivalent to Program 5 is conducted. In this case, 300 equally spaced values over the interval 0°–360° are treated as test inputs, and the GPR predictions over this range are visualized together with a confidence band defined as $\pm 2\sigma$. In this way, the usage of a library is demonstrated for the expressions derived in the previous sections.

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import RBF, WhiteKernel

# Generate 8 random angles between 0 and 360 degrees for training
np.random.seed(21)
```

```
X_deg = np.sort(np.random.uniform(0.0, 360.0, size=8))
X = np.deg2rad(X_deg).reshape(-1, 1)
y = np.sin(X).reshape(-1)

# Generate 1 random angle for testing
x_star_deg = np.random.uniform(0.0, 360.0)
x_star = np.deg2rad(x_star_deg).reshape(1, 1)
y_true = np.sin(x_star).item()

# Hyperparameters
lengthscale = 0.6
variance = 1.0
sigma_n2 = 0.005

kernel = (
    variance * RBF(length_scale=lengthscale, length_scale_bounds=(1e-2, 10.0))
    + WhiteKernel(noise_level=sigma_n2, noise_level_bounds=(1e-6, 1e-1)))

# Train the Gaussian Process Regressor
gpr = GaussianProcessRegressor(
    kernel=kernel,
    alpha=0.0,
    normalize_y=False,
    n_restarts_optimizer=10)
gpr.fit(X, y)

# Prediction
y_star_mean, y_star_std = gpr.predict(x_star, return_std=True)
y_star_var = y_star_std[0]**2

print("Training angles (degrees):", np.round(X_deg, 2))
print("Test angle (degrees):", round(x_star_deg, 2))
print("GPR predictive mean f*(x*) =", round(y_star_mean[0], 5))
print("True value f(x*) =", round(y_true, 5))
print("GPR predictive variance =", round(y_star_var, 7))
print("Learned kernel:", gpr.kernel_)
```

Program 6. Construction of a GPR model in the scikit-learn library using eight training points randomly selected from the interval 0°–360°, and computation of the predictive mean and uncertainty (variance) for a test point randomly selected from the same interval.

In Program 6, setting n_restarts_optimizer=10 causes the algorithm to first evaluate the initial hyperparameter values specified in the code, and then perform 10 additional optimization runs from different starting points. The final result is obtained using the set of hyperparameters that yields the best solution. The ranges explored during this process are specified by length_scale_bounds=(1e-2, 10.0) and noise_level_bounds=(1e-6, 1e-1). For the variance parameter, the default bounds (1e-2, 1e2) are used. However, this range can also be explicitly defined by the user via the ConstantKernel function.

The output obtained by running Program 6 is as follows:

```
Training angles (degrees): [7.78 17.54 18.28 74.13 104.08 108.82 239.01
259.55]
Test angle (degrees): 110.92
GPR predictive mean f*(x*) = 0.9337
True value f(x*) = 0.93407
GPR predictive variance = 1.8e-06
Learned kernel: 1.21**2 * RBF(length_scale=2.15) +
WhiteKernel(noise_level=1e-06)
```

As in Program 5, it can be useful to examine not only the prediction at a single test point but also how the GPR model behaves over a specified interval, in order to evaluate both prediction performance and how the uncertainty varies across the input domain. Therefore, in Program 7, values selected over the interval $0° - 360°$ are treated as test inputs, and the GPR predictions over this range are visualized. In this program, the confidence band representing predictive uncertainty is again chosen as $\pm 2\sigma$.

```
# Values in the 0–360 degree range for plotting
x_grid_deg = np.linspace(0.0, 360.0, 300)
X_grid = np.deg2rad(x_grid_deg).reshape(-1, 1)

# Posterior mean and standard deviation using sklearn GPR
mu_grid, std_grid = gpr.predict(X_grid, return_std=True)

```

```
# True sine function
y_true_grid = np.sin(X_grid).reshape(-1)

plt.figure(figsize=(10, 5))

# True sin(x)
plt.plot(
    x_grid_deg,
    y_true_grid,
    linestyle="--",
    linewidth=1.8,
    label="sin(x) (true function)")

# GPR posterior mean
plt.plot(
    x_grid_deg,
    mu_grid,
    linewidth=2.2,
    label="GPR predictive mean")

# ±2σ confidence band
plt.fill_between(
    x_grid_deg,
    mu_grid - 2 * std_grid,
    mu_grid + 2 * std_grid,
    alpha=0.25,
    label="Confidence band (±2σ)")

# Training points
plt.scatter(
    X_deg,
    y,
    s=70,
    zorder=3,
    label="Training points")

plt.xlabel("Angle (degrees)")
plt.ylabel(r"$f(x)$")
plt.title("Training Data, GPR Prediction, and Confidence Interval")
```

```
plt.grid(True)
plt.legend()
plt.show()
```

Program 7. Visualization of the GPR prediction and confidence interval over the range 0°–360° using the scikit-learn library after model training.

The plot obtained by running Program 7 is shown in Figure 5. When Figures 4 and 5 are compared, the contribution of hyperparameter optimization to the prediction performance can be clearly observed.

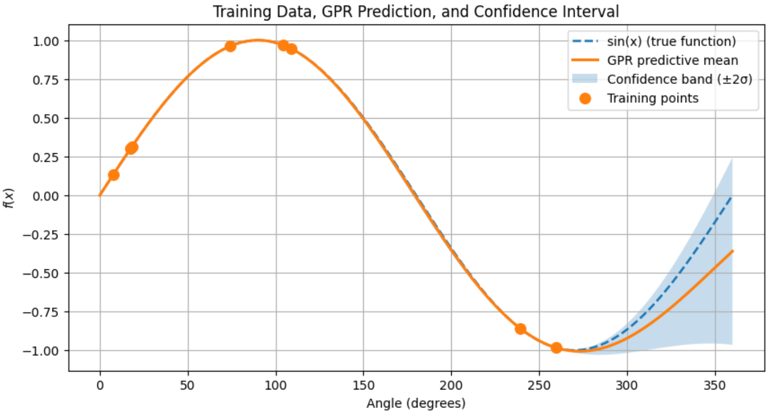


Figure 5. GPR prediction and confidence interval over 0°–360° obtained using the scikit-learn library.

Conclusion

Gaussian Process Regression is a powerful predictive model, particularly well suited for problems in which the available data are limited and prediction uncertainty plays a critical role. One of the most distinctive features of GPR is that it not only produces point estimates, but also provides a statistical quantification of how reliable these estimates are. In this respect, GPR offers significant advantages for the analysis of applications characterized by scarce data and high uncertainty.

As demonstrated in this chapter, GPR is developed using a wide range of concepts from statistical theory, most notably multivariate Gaussian distributions. The performance of the model largely depends on the appropriate selection of kernel hyperparameters. Hyperparameters such as the length scale, signal variance, and noise variance directly influence how well the model fits the data and how uncertainty is interpreted.

The programming implementation of GPR presented in this chapter facilitates a deeper understanding of the theoretical foundations of the method, while libraries such as scikit-learn enable the same process to be carried out in a practical and reliable manner.

In conclusion, owing to its strong theoretical foundation, uncertainty-aware framework, and ability to perform effectively with limited data, Gaussian Process Regression stands out as an important tool in modern machine learning and statistical modeling.

References

[1] C. E. Rasmussen ve C. K. I. Williams, *Gaussian processes for machine learning*, 3. print. içinde Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2008.

[2] J. Wang, "An Intuitive Tutorial to Gaussian Process Regression", *Comput. Sci. Eng.*, c. 25, sy 4, ss. 4-11, Tem. 2023, doi: 10.1109/MCSE.2023.3342149.

[3] Q. Nguyen, *Bayesian Optimization in Action*. Erscheinungsort nicht ermittelbar: Manning Publications, 2023.

[4] M. Liu, G. Chowdhary, B. Castra Da Silva, S.-Y. Liu, ve J. P. How, "Gaussian Processes for Learning and Control: A Tutorial with Examples", *IEEE Control Syst.*, c. 38, sy 5, ss. 53-86, Eki. 2018, doi: 10.1109/MCS.2018.2851010.

[5] D. K. Duvenaud, "Automatic Model Construction with Gaussian Processes".

[6] M. Binois ve R. B. Gramacy, "hetGP : Heteroskedastic Gaussian Process Modeling and Sequential Design in *R*", *J. Stat. Soft.*, c. 98, sy 13, 2021, doi: 10.18637/jss.v098.i13.

[7] A. G. de G. Matthews *vd.*, "GPflow: A Gaussian Process Library using TensorFlow", *Journal of Machine Learning Research*, c. 18, sy 40, ss. 1-6, 2017.

[8] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, ve A. G. Wilson, "GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration".

[9] F. Pedregosa *vd.*, "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, c. 12, ss. 2825-2830, 2011.

Chapter 2

Electrical Fault Analysis of Total-Cross-Tied Photovoltaic Array System: Simulation and Performance Analysis

Burcin OZKAYA[1], Evren ISEN[2]

1. INTRODUCTION

Rising global demand for electric power as a result of rapid economic development, increased population and higher average consumption has resulted in rising concerns about environmental damage and loss of fossil fuel resources (Madeti and Singh, 2017; Osmani et al., 2023). The combination of these factors has resulted in a worldwide shift toward the search for and development of renewable energy sources, which are becoming more and more known as non-polluting and reliable substitutes that can handle the future energy needs and the environmental problems (Bonthagorla and Mikkili, 2020; Ma, Yang, and Lu, 2014). Among renewable energy sources, photovoltaic (PV) solar systems have been considered the most viable and widely accepted option due to their modular design, relatively low operating and maintenance costs, absence of greenhouse gas emissions, and utilization of the sun, an abundant renewable resource (Pendem and Mikkili, 2018; Ma, Yang, and Lu, 2014).

PV systems utilize photovoltaic technology to convert solar energy from the sun into electricity. Since these systems produce power without harmful emissions, they also contribute positively toward protecting the environment by providing a sustainable and green source of energy. Unfortunately, although photovoltaic systems are easy and simple to use, many factors can impact their actual performance over time. The solar irradiance variation , temperature change, cell degradation, and faults are just a few examples of some of the external and internal factors that can affect PV system performance (Pendem and Mikkili, 2018). In particular, partial shading conditions (PSCs) are very detrimental to the overall reliability of PV systems. Partial shading conditions

---

[1] Assoc. Prof. Dr., Bandırma Onyedi Eylül University, Faculty of Engineering and Natural Sciences, Department of Electrical Engineering, bozkaya@bandirma.edu.tr, ORCID No: 0000-0002-9858-3982

[2] Assoc. Prof. Dr., Bandırma Onyedi Eylül University, Faculty of Engineering and Natural Sciences, Department of Electrical Engineering. eisen@bandirma.edu.tr, ORCID No: 0000-0002-3107-9255.

occur when parts of a PV array do not receive the same amount of irradiance because of temporary or localized obstacles (e.g., trees, buildings, clouds, or airborne debris) (Belhachat and Larbes, 2021). In order to minimize the negative effects caused by partial shading, one of the many methods used is the integration of bypass diodes into the photovoltaic modules that serve as a kind of mitigation strategy (Devakirubakaran, Verma, Chokkalingam, Mihet-Popa, 2023; Saiprakash, Mohapatra, Nayak, Ghatak, 2021). The purpose of these diodes is to provide a path for current to flow and bypass shaded or defective parts of a solar module, thus eliminating power bottlenecks by being connected in parallel with solar cells or group of cells. While effective at reducing hotspot formation and improving current flow under mismatched conditions, the use of bypass diodes inadvertently introduces multiple local maximum power points in the system's P–V and I–V curves (Devakirubakaran, Verma, Chokkalingam, Mihet-Popa, 2023). Henceforth, although the functioning of bypass diodes is really important from the protective aspect, the interaction between shading conditions and bypass diodes adds extra complications to the control and optimization of photovoltaic systems.

Numerous techniques proposed in the literature aim to mitigate mismatch losses caused by PSCs, including changing the PV array configurations, MPPT methods, and converter circuits topologies. Of all these methods, PV array layout is a direct means to minimize power losses in a situation of shading, as it affects current distribution in modules under different irradiance conditions (Pendem and Mikkili, 2018). A good configuration allows not only better current distribution but also more energy extraction from the system as a whole. Mismatch losses throughout PV arrays can be almost entirely eliminated if the right interconnection schemes are adopted and the modules are physically arranged for maximum irradiance (Saiprakash, Mohapatra, Nayak, Ghatak, 2021). Different configurations have been tried by researchers such as series-parallel (SP), honeycomb (HC), bridge-linked (BL), and total cross-tied (TCT), which have different levels of shading resistance (Belhachat and Larbes, 2015). The research carried out by the literature on PV array configurations shows that TCT has not only made the current distribution more even but has also provided great tolerance for the faults across the PV array (Bana and Saini, 2017; Belhachat and Larbes, 2015; Saiprakash, Mohapatra, Nayak, Ghatak, 2021; Jha, 2022). The findings of researchs emphasize that the way in which the PV arrays are configured is one of the key determinants of the PV systems' resilience and, thus, efficiency when operated under real-world, non-ideal condition.

Photovoltaic (PV) systems are not only affected by environmental conditions, but they are also subjected to different electrical faults that come to be the most important reasons of their effectiveness, reliability, and safety being undermined tremendously. The last ten years have witnessed a significant amount of research towards enhancing PV system performance during both healthy and faulty conditions considering that the number of field-deployed system failures keeps increasing (Satpathy, Aljafari, Thanikanti, and Madeti, 2023). Such faults are of different types; they can be classified as ground faults (Pillai and Rajasekar, 2018), open-circuit faults (Livera, Florides, Theristis, Makrides, and Georghiou, 2018; Pillai and Rajasekar, 2018; Satpathy, Aljafari, Thanikanti, and Madeti, 2023), short-circuit faults (Livera, Florides, Theristis, Makrides, and Georghiou, 2018; Satpathy, Aljafari, Thanikanti, and Madeti, 2023), line-to-line faults (Zhao, De Palma, Mosesian, Lyons, and Lehman, 2012; Pillai and Rajasekar, 2018; Eskandari, Milimonfared, and Aghaei, 2020; Satpathy, Aljafari, Thanikanti, and Madeti, 2023), line-to-ground faults (Satpathy, Aljafari, Thanikanti, and Madeti, 2023), bypass diode faults and mismatch faults (Pillai and Rajasekar, 2018; Satpathy, Aljafari, Thanikanti, and Madeti, 2023), degradation faults (Satpathy, Aljafari, Thanikanti, and Madeti, 2023), that may start at the module level and progress through the array.

The impact of faults on PV array efficiency has been demonstrated by numerous studies conducted through the years. Amongst them, Satpathy et al. did a lot of work on the possible problem of electrical faults in all commercial PV array configurations. The researchers looked into nine faulty cases where they used to characterize different types of faults such as: open circuit fault, short-circuit fault, bypass diode short-circuit fault, line-ground fault, line-line fault, degradation fault, inverted bypass diode fault, mismatch fault and mismatch with inverter bypass diode fault (Satpathy, Aljafari, Thanikanti, and Madeti, 2023). Eskandari et al. designed an advanced fault detection procedure based on the use of an ensemble learning framework and I-V characteristics to pinpoint and classify line-to-line faults (Eskandari, Milimonfared, and Aghaei, 2020). Zhao et al. presented the behavior and progress of line-to-line faults in PV arrays, among which were those due to the short circuit and double grounding. The experiment showed that the currents produced by the faults in PV system were much influenced by the position of the fault, resistance of the fault, level of solar radiation, and the presence of blocking diodes (Zhao, De Palma, Mosesian, Lyons, and Lehman, 2012). Livera et al. worked on and assessed a fault diagnosis system that is capable of detecting short-circuit and open-circuit failure of modules in PV systems by relying on parameters that are extracted from the existing monitoring infrastructures (Livera, Florides, Theristis, Makrides, and Georghiou, 2018).

The present research delves into the total-cross-tied array configuration as a basis through MATLAB for a 4x4 array. The electrical faults considered in the study were open-circuit and line-to-line. Furthermore, the researchers considered three varying patterns of partial shading: short and narrow, long and narrow, and random. The simulation study involved the creation of three different case studies corresponding to the different shading patterns. The performance of the PV system was analyzed for every shading pattern during the following conditions: uniform shading, partial shading with no fault, partial shading with open-circuit fault, and partial shading with line-to-line fault. The assessment of the case studies was carried out using various parameters such as power-voltage curves, power loss, and fill factor.

2. PHOTOVOLTAIC SYSTEM UNDER INVESTIGATION

In this section, detailed explanations of the PV array's connecting topology, shading examples, and performance evaluation indicators are provided, which was used in the simulation study.

2.1. TCT PV Array Configurations

The PV modules in a total-cross-tied (TCT) configuration are connected electrically in a two-dimensional grid to attain current sharing at the maximum, to reduce the impact of partial shading and mismatch. In TCT, all of the modules in each row are directly connected in parallel to each other, and then the rows are connected to each other in series (Krishna and Moger, 2019; Madhanmohan, Nandakumar, Saleem, 2025). The layout of a 4×4 TCT connected PV array is given in Figure 1. Under standard test conditions, the terminal voltage and current of a PV module are V_g and I_g, respectively. The voltage of each row is the same because all the modules of a row are connected in parallel. The rows are connected in series, hence the output voltage (V_{out}) of the PV array is as follows:

$$V_{out} = \sum_{i=1}^{4} V_{i1} = V_{11} + V_{21} + V_{31} + V_{41} = 4V_g \tag{1}$$

The current delivered by the *i*th row of a PV array can be represented as follows:

$$I_{out} = \sum_{j=1}^{4} I_{ij} = I_{i1} + I_{i2} + I_{i3} + I_{i4} = 4I_g \tag{2}$$

where, the current produced by the photovoltaic module located in the *i*th row and *j*th column of a PV array is denoted as I_{ij}. There is no variation in the current delivered by any row because they are all connected in series with the load.

The current that a photovoltaic module produces, after considering the effect of solar irradiation, is determined by:

$$I = \frac{G}{G_{STC}} I_g \tag{3}$$

Assuming that N_{SH} is the number of modules in a row that are shaded, the row current can be calculated as follows:

$$I_{R_j} = \left[n - \left(1 - \frac{G}{G_{STC}} \right) N_{SH} \right] I_g \tag{4}$$

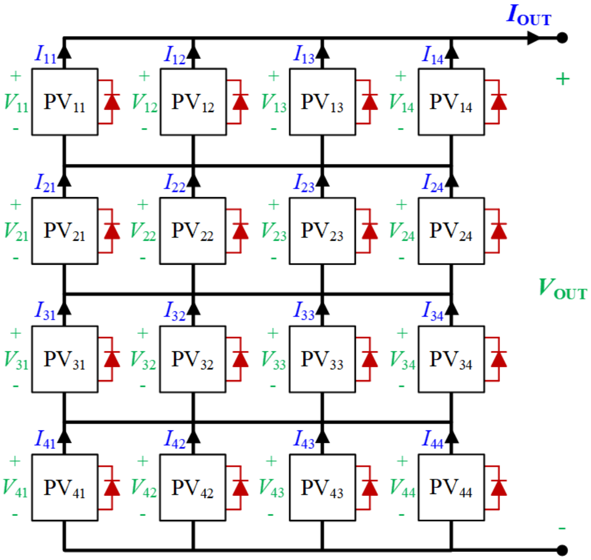
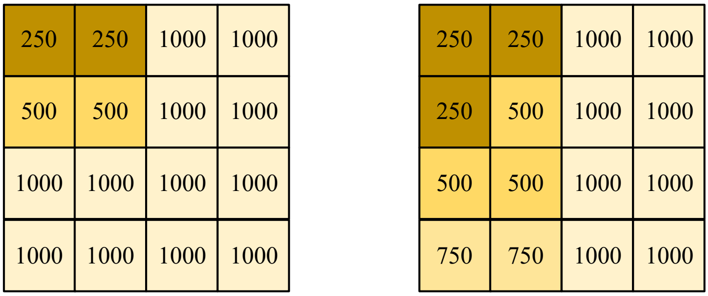


Figure 1. TCT configuration

2.2. Partial Shading Patterns

The power output of the PV array is mainly dependent on partial shading. This section focuses on different shading patterns that were applied to a PV array configured with 4 × 4 TCT. The literature describes different shading patterns based on the number of shaded modules per string (column) and the number of shaded strings, as well as the level of solar irradiation on the PV modules in an array. These three shading patterns shown in Figure 2 are: short and narrow, long and narrow, and random. The various characteristics of these shading patterns are as follows:

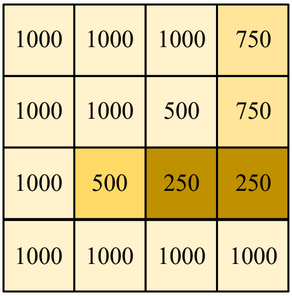


250	250	1000	1000
500	500	1000	1000
1000	1000	1000	1000
1000	1000	1000	1000

(a) Short and narrow

250	250	1000	1000
250	500	1000	1000
500	500	1000	1000
750	750	1000	1000

(b) Long and narrow



1000	1000	1000	750
1000	1000	500	750
1000	500	250	250
1000	1000	1000	1000

(c) Random

Figure 2. The partial shading patterns

2.2.1 Short and narrow shading pattern

In the case of the short and narrow shading pattern, only two out of four strings are shaded, and each shaded string consists of two modules. The solar irradiance is categorized into three classes, as depicted in Figure 2 (a), in order to evaluate how well the PV configurations perform under PSCs. The levels of solar irradiance changing during short and narrow shading are given as follows:

Part 1: solar irradiance - 250 W/m^2, 3 PV modules

Part 2: solar irradiance - 500 W/ m2, 3 PV modules

Part 3: solar irradiance - 1000 W/ m2, 12 PV modules

2.2.2 Long and narrow shading pattern

In the case of the long and narrow shading pattern, all strings are shaded, with each shaded string containing two modules. The solar irradiance levels in a PV array are categorized into four groups, presented in Figure 2 (b) and described as follows:

Part 1: solar irradiance - 250 W/m^2, 3 PV modules

Part 2: solar irradiance - 500 W/ m^2, 3 PV modules

Part 3: solar irradiance - 750 W/ m^2, 2 PV modules

Part 4: solar irradiance - 1000 W/ m2, 8 PV modules

2.2.3 Random shading pattern

Under this shading pattern, the PV modules are randomly shaded to distinct solar irradiation levels, as shown in Figure 2 (c). The modules named PV4 and PV8 are shaded at 750W/m^2; PV10 is shaded at 500W/m^2; PV11 and PV12 are shaded at 250W/m^2, and the remaining modules are shaded at 1000 W/m^2.

2.3. Performance Evaluation Indicators

The global maximum power point (GMPP) is the primary metric that is used when assessing the performance of a PV array. The power loss is then computed for the case of no faults on one hand and for the case of faults on the other hand. The power loss in the no-fault condition (PL_1) is calculated with the help of Equation (5), where $P_{GMPP,NSNF}$ and $P_{GMPP,PSNF}$ are the GMPP at the uniform test condition and the partial shading condition, respectively. The power loss in the fault condition (PL_2) is computed using Equation (6), where $P_{GMPP,PSNF}$ and $P_{GMPP,PSF}$ represent the GMPP under partial shading conditions without and with fault conditions, respectively.

$$PL_1 = P_{GMPP,NSNF} - P_{GMPP,PSNF} \tag{5}$$

$$PL_2 = P_{GMPP,PSNF} - P_{GMPP,PSF} \tag{6}$$

The fill factor (FF) is another performance metric and it is defined as the ratio of the generated maximum power to the product of V_{oc} and I_{sc}, as stated in Equation (7). A value of FF closer to one indicates the superior performance of the PV system (Satpathy, Aljafari, Thanikanti, and Madeti, 2023).

$$FF(\%) = \frac{P_{GMPP}}{V_{oc} \times I_{sc}} \tag{7}$$

3. FAULTS IN PV ARRAYS

The PV arrays can experience a range of electrical faults that can, in turn, greatly effect the performance of the system, cut down the power output and, in the worst cases, lead to total operational failure. The total power of the array is equal to the sum of the individual module outputs under the uniform test conditions; however, in the field, many different fault types will interrupt this predictable behavior. The faults can be temporary, like the aforementioned mismatch faults due to partial shading, or they can be permanent, these include open-circuit faults, line-to-line faults, and bypass diode failures, etc. The faults mentioned above change current and voltage allocation in the array and sometimes cause power imbalance and efficiency loss. Therefore, it becomes necessary to figure out how PV arrays would behave under different faults, and assess their fault tolerance in order to achieve reliable operation, the best performance, and better system resilience, all of which are the main aims of this being.

3.1. Open-Circuit Fault

In an open-circuit fault, electrical continuity in a PV module, string, or interconnection is lost. Thus, the branch where the fault occurred no longer has current flow (Osmani et al., 2023). The open-circuit fault taken into account in this study is illustrated in Figure 3.

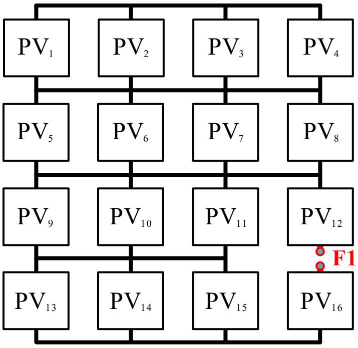


Figure 3. Open-circuit fault in 4×4 TCT configured PV array

3.2. Line-to-Line Fault

A line-to-line (LL) fault is characterized by the accidental contact of the two conductors that are at different potentials, which most often happens as a result of insulation failure, breaking of the cable, or environmental conditions. While photovoltaic modules are designed to limit fault current, low-level failures nevertheless cause significant power loss, string mismatch, bypass diode activation, and distortion of power-voltage characteristics (Osmani et al., 2023). Figure 4 shows the LL fault considered in this study.

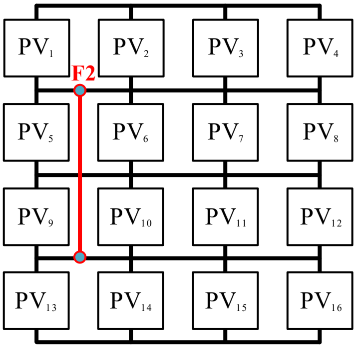


Figure 4. Line-to-line fault in 4×4 TCT configured PV array

4. RESULTS AND DISCUSSIONS

The fault tolerance investigation of TCT configuration was carried out for the 4×4 PV array operating under partial shading conditions. The power generation of the PV array under uniform irradiance (1000 W/m^2 and 25°C) was recorded as 3408.3 W in simulation with 116.033 V as voltage output, 29.3735 A as current output, 145.2 V as V_{oc}, and 31.4562 A as I_{sc}. In this study, the performance of the PV array was examined for fault and non-fault conditions in Case-1 under short and narrow shading pattern, Case-2 under long and narrow shading pattern, and Case-3 under random shading pattern.

4.1. Results of Case-1

In this case, the performance of the PV array under the short and narrow shading pattern for no fault, F1 and F2 fault conditions was investigated. Table 1 shows the maximum power values obtained in the PSNF, F1 and F2 cases and the results of the performance evaluation indicator detailed in section 2.3.

Table 1. The results of the Case-1.

Case(s)	$P_{GMPP,PSNF}$ **(W)**	$P_{GMPP,PSF}$ **(W)**	PL_1 **(W)**	PL_2 **(W)**	FF
PSNF	2376.6	-	1031.7	-	0.52
F1	-	2308.1	-	68.5	0.51
F2	-	1151.7	-	1224.9	0.25

According to Table 1, when the power losses in F1 and F2 were evaluated, it was observed that the power loss was greater in F2. The FF value ranges between 0 and 1, and a higher FF value indicates better PV array performance. Accordingly, since the FF value in case F2 was lower than in case F1, it means that the PV array's operating performance was higher in F1 than in F2. Figure 5 presents the P-V curve of the PV array for NSNF, NF, F1, and F2. When the P-

V curves were examined, it is seen that there were three peaks in the NF case, while there were two peaks in the F1 and F2 faults. Furthermore, the open-circuit voltage in F1 was almost the same as in the NF, while the open-circuit voltage in F2 was halved compared to the NF case.

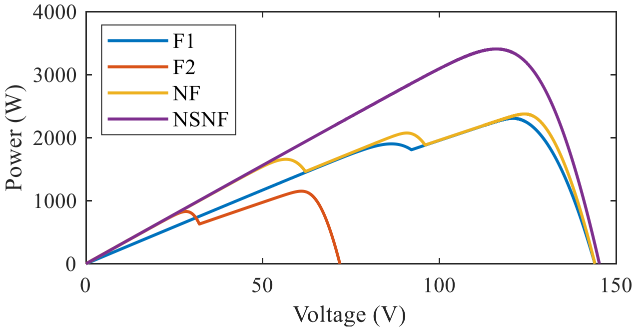


Figure 5. The P-V curves of the Case-1.

4.2. Results of Case-2

The performance of the PV array under long and narrow shading pattern was examined in the absence of faults, as well as under F1 and F2 fault conditions. Table 2 presents the maximum power values achieved in the PSNF, F1, and F2, along with the results of the performance evaluation metrics.

Table 2. The results of Case-2

Case(s)	$P_{GMPP,PSNF}$ **(W)**	$P_{GMPP,PSF}$ **(W)**	PL_1 **(W)**	PL_2 **(W)**	FF
PSNF	2311.5	-	1096.8	-	0.51
F1	-	1874.4	-	437.1	0.41
F2	-	1140.3	-	1171.2	0.25

Table 2 indicates that the power loss in F1 (437.1 W) was lower than the F2 (1171.2 W). When evaluated based on the FF value, it is seen that while the FF value was 0.51 in the PSNF case, it was 0.41 and 0.25 in the F1 and F2 cases, respectively. These results indicate that the operating performance of the PV array was reduced in the fault conditions. Figure 6 shows the P-V curve of the PV array for all cases. Accordingly, the global maximum power value and open-circuit voltage were lowest in case F2.

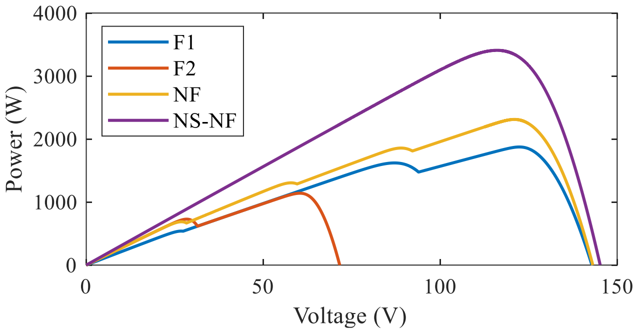


Figure 6. The P-V curves of the Case-2.

4.3. Results of Case-3

In this case, a comparison was made between the performance of the PV array under random shading pattern in the absence of faults, as well as under F1 and F2 faults. The results of the Case-3 are presented in Table 3.

Table 3. The results of Case-3

Case(s)	$P_{GMPP,PSNF}$ **(W)**	$P_{GMPP,PSF}$ **(W)**	PL_1 **(W)**	PL_2 **(W)**	***FF***
PSNF	2204.3	-	1204	-	0.48
F1	-	2017.4	-	186.9	0.44
F2	-	1639.8	-	564.5	0.36

As can be seen in Table 3, the power loss in F1 was found to be 186.9 W, which is a smaller value than the power loss in F2 (564.5 W). When the FF value was evaluated, it was observed that the FF value was 0.48 in the PSNF, whereas it was 0.44 and 0.36 in the F1 and F2 cases, respectively. Based on these results, it can be deduced that the PV array's operational performance was diminished when it was subjected to fault situations. Moreover, the P-V curve of the PV array for all cases presented in Figure 7 show that F2 had the lowest global maximum power value and open-circuit voltage.

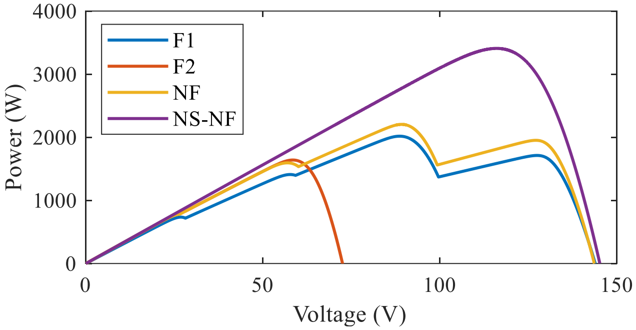


Figure 7. The P-V curves of the Case-3.

5. CONCLUSION

In this study, the effects of partial shading and electrical faults on a TCT-connected PV array were investigated using MATLAB/Simulink. For this purpose, three shading patterns were created, and the global maximum power of the PV array under no-fault, F1-fault, and F2-fault conditions was recorded. Using these values, the power loss and fill factor performance metrics were computed to evaluate the PV array's performance. Moreover, the performance of the PV array was investigated based on the P-V characteristic curve. The global maximum power values and performance indicators clearly show that the power loss occurred and the fill factor value decreased under fault conditions. As a result, it was observed that PV systems experience power losses during partial shading and fault conditions, negatively impacting their performance. It was determined that the fault types and locations affect performance to varying degrees.

REFERENCES

Bana, S., & Saini, R. P. (2017). Experimental investigation on power output of different photovoltaic array configurations under uniform and partial shading scenarios. Energy, 127, 438-453.

Belhachat, F., & Larbes, C. (2021). PV array reconfiguration techniques for maximum power optimization under partial shading conditions: A review. Solar Energy, 230, 558-582.

Bonthagorla, P. K., & Mikkili, S. (2020). Performance analysis of PV array configurations (SP, BL, HC and TT) to enhance maximum power under non-uniform shading conditions. Engineering Reports, 2(8), e12214.

Devakirubakaran, S., Verma, R., Chokkalingam, B., & Mihet-Popa, L. (2023). Performance evaluation of static PV array configurations for mitigating mismatch losses. IEEE Access, 11, 47725-47749.

Eskandari, A., Milimonfared, J., & Aghaei, M. (2020). Line-line fault detection and classification for photovoltaic systems using ensemble learning model based on IV characteristics. Solar Energy, 211, 354-365.

Jha, V. (2022). Comprehensive modeling and simulation of PV module and different PV array configurations under partial shading condition. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 46(2), 503-535.

Krishna, S. G., & Moger, T. (2019). Optimal SuDoKu reconfiguration technique for total-cross-tied PV array to increase power output under non-uniform irradiance. IEEE Transactions on Energy Conversion, 34(4), 1973-1984.

Livera, A., Florides, M., Theristis, M., Makrides, G., & Georghiou, G. E. (2018, June). Failure diagnosis of short-and open-circuit fault conditions in PV systems. IEEE 7th World Conference on Photovoltaic Energy Conversion, 0739-0744.

Ma, T., Yang, H., & Lu, L. (2014). Solar photovoltaic system modeling and performance prediction. Renewable and sustainable energy reviews, 36, 304-315.

Madeti, S. R., & Singh, S. N. (2017). A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. Solar Energy, 158, 161-185.

Madhanmohan, V. P., Nandakumar, M., & Saleem, A. (2025). Enhanced performance of partially shaded photovoltaic arrays using diagonally dispersed total cross tied configuration. Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 47(1), 477-495.

Osmani, K., Haddad, A., Lemenand, T., Castanier, B., Alkhedher, M., & Ramadan, M. (2023). A critical review of PV systems' faults with the relevant detection methods. Energy Nexus, 12, 100257.

Pendem, S. R., & Mikkili, S. (2018). Modeling, simulation and performance analysis of solar PV array configurations (Series, Series–Parallel and Honey-Comb) to extract maximum power under Partial Shading Conditions. Energy Reports, 4, 274-287.

Pillai, D. S., & Rajasekar, N. (2018). A comprehensive review on protection challenges and fault diagnosis in PV systems. Renewable and Sustainable Energy Reviews, 91, 18-40.

Saiprakash, C., Mohapatra, A., Nayak, B., & Ghatak, S. R. (2021). Analysis of partial shading effect on energy output of different solar PV array configurations. Materials Today: Proceedings, 39, 1905-1909.

Satpathy, P. R., Aljafari, B., Thanikanti, S. B., & Madeti, S. R. K. (2023). Electrical fault tolerance of photovoltaic array configurations: Experimental investigation, performance analysis, monitoring and detection. Renewable Energy, 206, 960-981.

Zhao, Y., De Palma, J. F., Mosesian, J., Lyons, R., & Lehman, B. (2012). Line–line fault analysis and protection challenges in solar photovoltaic arrays. IEEE transactions on Industrial Electronics, 60(9), 3784-3795.

Chapter 3

A Comparison of Hydrometeorological Relationship in the Ceyhan Basın, Turkiye Using Linear and Non-linear Regression Analysis Based on Long-Term Observation Data (1980-2020)

Evren TURHAN[1], Serin DEĞERLİ ŞİMŞEK[2]

INTRODUCTION

Accurate planning of water resources is particularly crucial for comprehensive hydrological studies. These planning processes require the use of numerous hydrometeorological data for various purposes. Considering many hydrological studies, the accuracy of variables such as precipitation or rainfall, streamflow, temperature, and so on, is vital for accurate modeling processes. However, it is a fact that these dataset are often incomplete, often due to various reasons. Therefore, conducting analyses using different methods via historically recorded observed data on water resources in drainage basins, and ultimately addressing these shortcomings, will contribute positively to the design phases in many ways. Due to the complexities of the hydrological cycle, considering non-linear modeling studies can be significant for accurate and effective predictions. When examining published literature, it is possible to encounter applications using linear and non-linear regression methods in river flow prediction studies [1-5].

In this study, data from streamflow observation stations numbered 2004, 2006, 2007, and 2008 belonging to the former General Directorate of Electric Power Resources Survey and Development Administration (Turkish: EİEİ) and numbered D20A046 belonging to the General Directorate of State Hydraulic Works (Turkish: DSİ) located within the borders of the Ceyhan Basin were utilized. Precipitation and temperature data from the Ceyhan station of Turkish State Meteorological Services (Turkish: MGM) were assumed as fixed independent variables, and other station data were modelled with linear and

---

[1] Assoc. Prof. Dr.; Adana Alparslan Türkeş Science and Technology University. Faculty of Engineering, Department of Civil Engineering. eturhan@atu.edu.tr ORCID No: 0000-0002-0742-4848

[2] Res. Asst. Adana Alparslan Türkeş Science and Technology University. Faculty of Engineering, Department of Civil Engineering. sdegerli@atu.edu.tr ORCID No: 0000-0003-0208-9152

non-linear regression analyses at the appropriate correlation level. Non-linear regression analysis was performed using a macro software created in MS-Excel. Additionally, Datafit 9.0 was used for non-linear regression modeling. The results were also examined according to different evaluation criteria such as coefficient of determination (R^2), and Root Mean Square Error (RMSE) [6-8]. In addition, it has been tried to determine the most appropriate equations for linear and non-linear analyses.

MATERIALS AND METHODS

Study Area and Data Used

The Ceyhan Basin has a precipitation area of approximately 21.391 km^2, which corresponds to 2.7% of Türkiye's total surface area. The Mediterranean climate prevails in the basin, and heavy rainfall can be observed near the end of the winter season. There are many water sources within the basin; the Ceyhan River, one of the largest, originates in the Pınarbaşı area and flows south into the Mediterranean Sea [9]. The observation stations considered in this study are 2004, 2006, 2007, 2008 and D20A046. The location of the Flow Observation Station (FOS) 2004 is approximately 300 m upstream of the Misis Bridge in the Misis subdistrict of Adana province, and 2006 is on the bridge and the road that turns towards Karaahmet village, at the 5th km of the Göksun-Elbistan highway in Kahramanmaraş province. The location of 2007 FOS is near the bridge in Çukurköprü village of Kadirli district of Adana province. 2008 FOS is on the pedestrian bridge 150 m downstream of the concrete bridge at the entrance of Kadirli district of Adana province. D20A046 FOS is in Sarıdanışmanlı Village, 5 km from the village road that branches left from the 25th km of the Kadirli-Andırın road. Figure 1 shows the locations of the relevant stations in the Ceyhan Basin [10, 11].

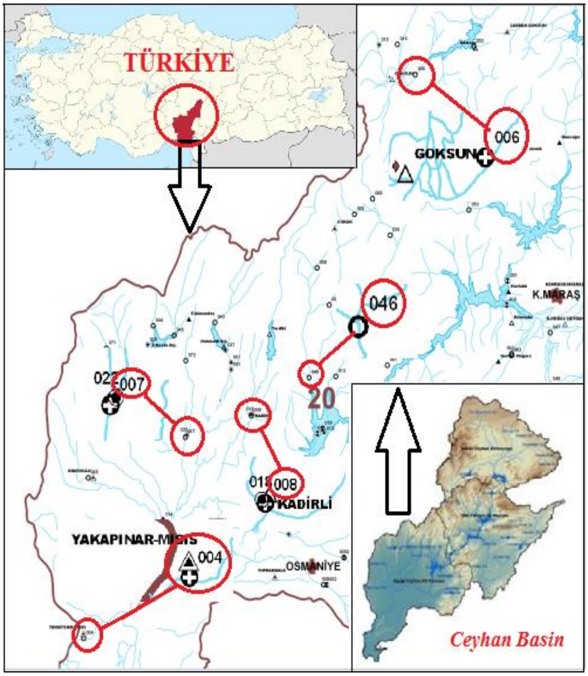


Figure 1. Locations of the streamflow stations considered in the Ceyhan Basin, Türkiye [11]

Additionally, Ceyhan precipitation and temperature observation station data obtained from the MGM were also included in the study (Figure 2). Table 1 provides general information about the stations in the Ceyhan Basin [12].

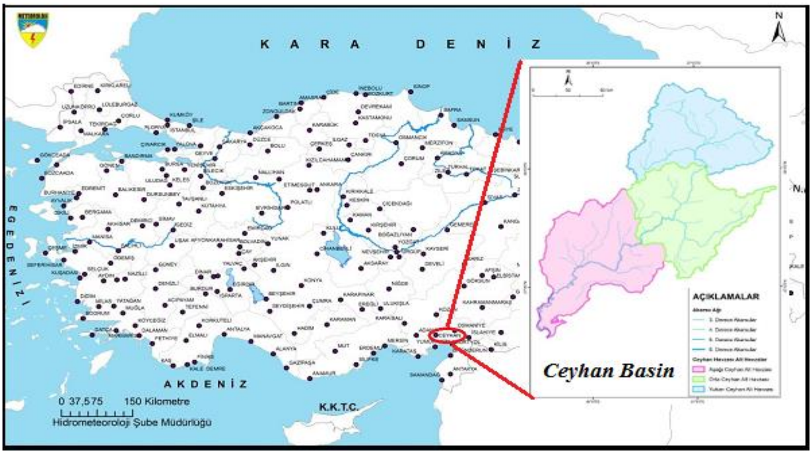


Figure 2. General view of precipitation and temperature stations in the Ceyhan Basin, Türkiye [12]

Table 1. Some information on the stations used in the study [11-12]

Station Number	Station Name	Latitude (N)	Longitude (E)	Rainfall Area (km^2)	Observation time (year)
D20A046	Keşiş Creek-Sarıdanışmanlı	37°26’51”	36°15’16”	420	1980-2020
2004	Ceyhan River-Misis	36°57’28”	35°38’03”	20466	
2006	Göksun Creek-Karaahmet	38°01’55”	36°34’11”	739.20	
2007	Sonboz Creek-Çukurköprü	37°20’29”	35°55’03”	623.00	
2008	Savrun Stream-Kadirli	37°22’19”	36°05’39”	444.00	
17960	Ceyhan Precipitation	37°00’55”	35°47’44”	---	
	Ceyhan Temperature				

Linear, Non-linear Regression Analysis and Evaluation Criteria

In this study, linear and non-linear regression analyses were conducted to constitute the basis of the streamflow estimation approach. Regression analysis aims to identify semantic relationships between the determined variables, to find the equation that formulates this relationship if any exists, and to generate predictions using these equations [13].

To determine whether there is a significant relationship between variables, the correlation or coefficient of determination can be examined. If the value of the correlation coefficient is absolutely close to ‘1’, there is a strong correlation between the two variables, and therefore, a relationship can be established [14]. The simple linear regression relationship is shown in Equation 1; if the dependent variable in linear regression analysis is expressed as 'y' and the independent variables are expressed as x_1, x_2, …, x_k, …, x_s; the multiple regression relationship between them is shown in Equation 2 below [15]:

$$y=a+bx+e \tag{1}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \cdots + \beta_s x_s + \delta \tag{2}$$

In the Equation 1, “a” represents a constant value, and “e” represents the error value. In the Equation 2, the unknowns β_0, β_1, β2, …, β_k, ..., βs are called ‘*Regression Coefficients*’. The coefficient β_k shows the difference in the y variable caused by a unit change in x_k when other variables are constant. The parameters β_k (k=1, 2, …, s) are referred to as '*Partial Regression Coefficients*'. β_0 is considered a constant number and represents the value that occurs when all x_k variables are zero. 'δ' is the error constant [16]. Since the data for regression are real numbers and do not show discontinuity, it was considered suitable for the analysis of the input data.

Non-linear analysis gives the relationship between dependent and independent variables. Logistic, exponential, and polynomial equations are often utilized in non-linear analysis [17]. For example, assuming an exponential function exists between the dependent variable 'N' and the independent variables shown in Equation 3, the relationship in non-linear regression analysis can be considered as follows:

$$N = \lambda\, {a_1}^{\beta_1}\, {a_2}^{\beta_2}\, {a_3}^{\beta_3} \ldots\, {a_v}^{\beta_v} \tag{3}$$

Here, βv represents the vth model parameter, λ represents the multiplicative error term, and v represents the number of variables [18]. In order to evaluate the performance in terms of evaluation, the R^2 coefficient of determination, and RMSE criteria, which are commonly used in the literature, were used [6-8, 19]. The evaluation criterion formulas in question are given in Equations 4, and 5:

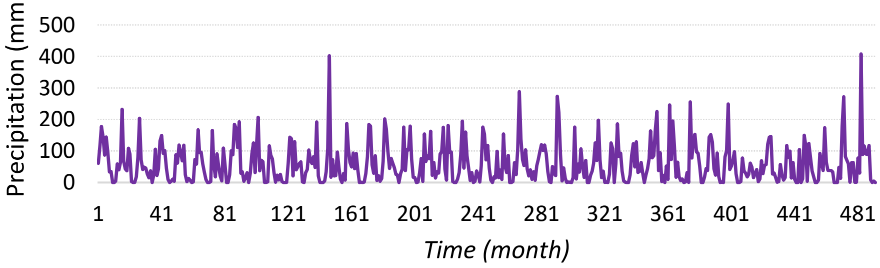
$$R^2 = \frac{\sum_{i=1}^{N}(Q_{obs} - Q_{average})^2 - \sum_{i=1}^{N}(Q_{obs} - Q_{prediction})^2}{\sum_{i=1}^{N}(Q_{obs} - Q_{average})^2} \tag{4}$$

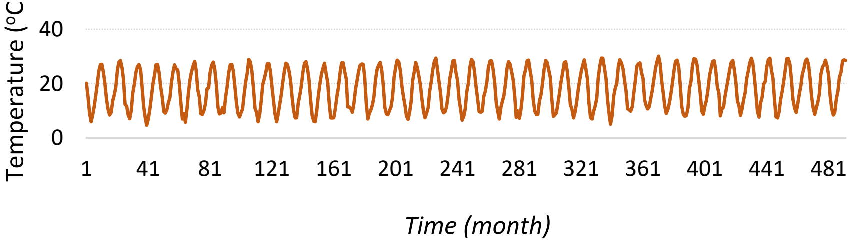
$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Q_{obs} - Q_{prediction})^2} \tag{5}$$

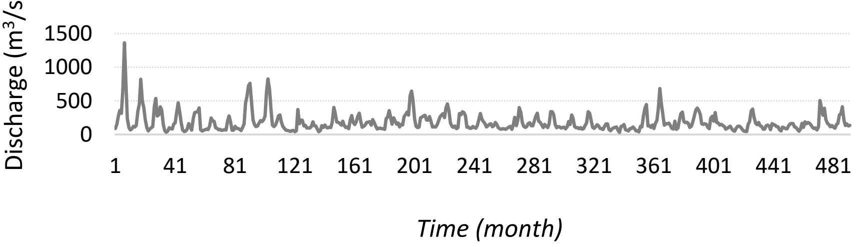
RMSE being close to zero (0) and the R^2 coefficient of determination being close to '1' indicates that there may be a strong correlation in terms of prediction [20]. Macro software was used in the MS-Excel to perform the prediction performances with linear and non-linear regression analyses. Additionally, Datafit 9.0 software was used for non-linear regression modeling [2].

RESULTS AND DISCUSSIONS

In this study, firstly, hydrometeorological data course line graphs were created by using the relevant data and methods (Figure 3a-g). Afterwards, prediction models were carried out in which Ceyhan station precipitation or rainfall data were estimated as independent variables. For non-linear regression analysis, the previous month's streamflow data were also added to the models as an independent variable. The data of the streamflow stations numbered 2004, 2006, 2007, 2008 and D20A046 were entered into the models as dependent variables. The effect of temperature values from hydrometeorological data on linear and non-linear regression analyses was investigated. Temperature data of Ceyhan station were evaluated in the same time period. Finally, models were tested in which both precipitation and temperature data were considered as independent variables (Figure 4-6). As can be seen in the Figure 3, except for a few periods, average or mean precipitation and temperature remain at certain levels. While the highest values in terms of streamflow data were observed at station 2004, the lowest data were obtained from station 2007. The highest value for the specified time period for station no2004 was measured as 182.681 m^3/s. For station no 2007, this value was 12.361 m^3/s.


(a) P_{Ceyhan} Precipitation Course Line


(b) T_{Ceyhan} Temperature Course Line


(c) Q_{2004} Streamflow Course Line

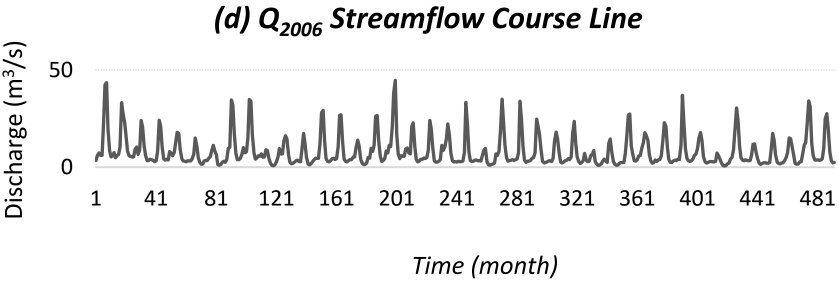


Figure 3. Hydrometeorological data course line graphs

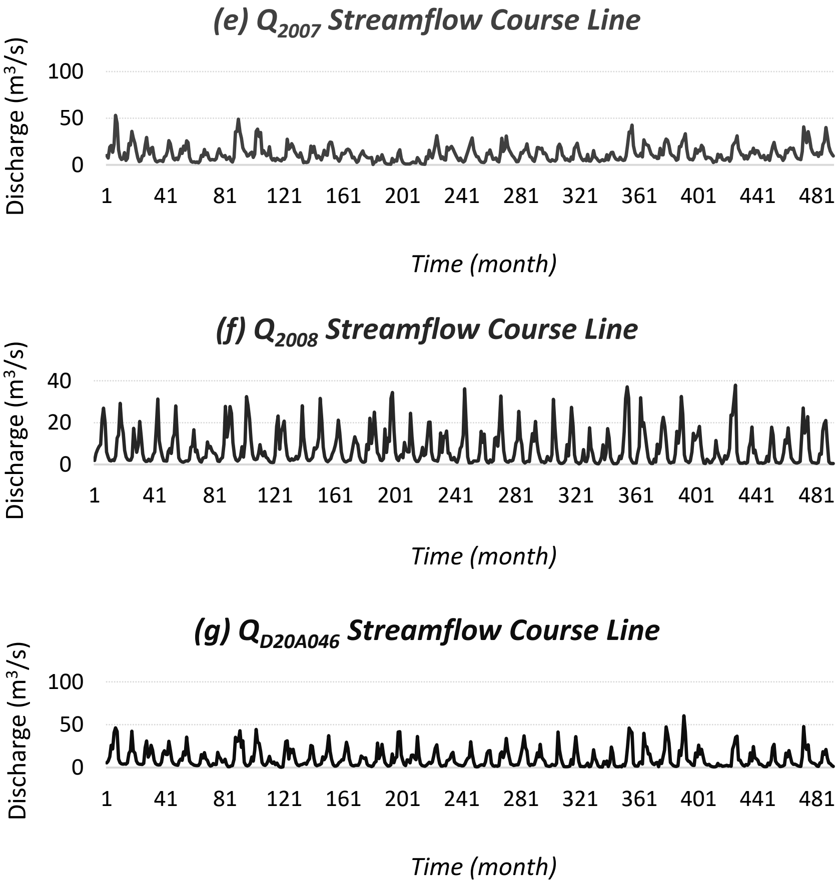


Figure 3 (Continued**)**

Data from the previous month were included in the models, particularly for non-linear analyses. Precipitation or rainfall and temperature were evaluated as fixed independent variables, with precipitation first, temperature second, and both precipitation and temperature last. Data from other stations were processed using a single time lag step in the MS-Excel macro software. Graphs from linear and non-linear regression analyses, along with the appropriate equations, are shown in Figures 4(a-e), 5(a-e), and 6(a-e). Additionally, the obtained outputs from the Datafit program for non-linear analyses are also presented.

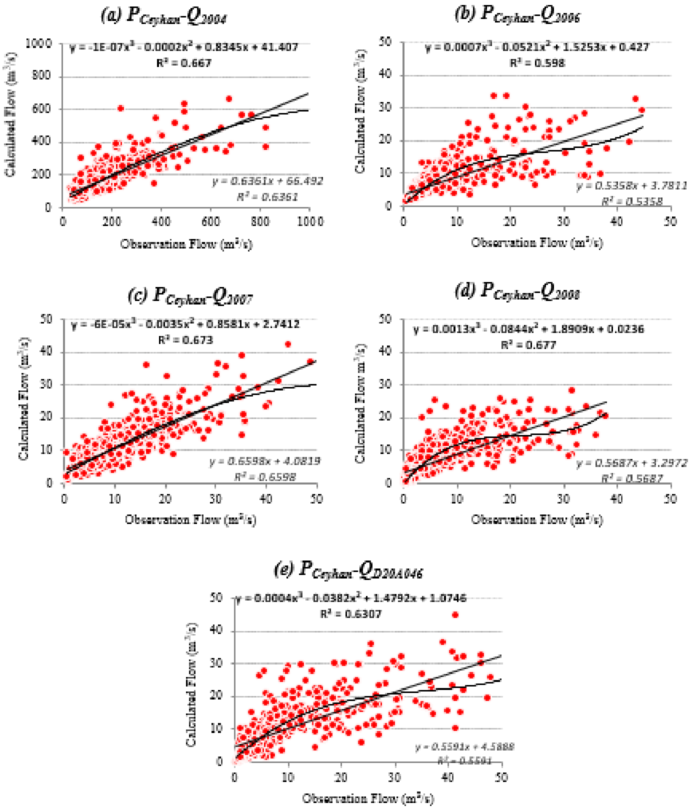


Figure 4. Graphs from linear and non-linear regression analyses for precipitation invariable data (a-e)

As seen in the Figure 4, the highest R^2 value has been achieved with the P-Q_{2007} model. A satisfactory approximation has been achieved with both linear and non-linear input data. The P-Q_{2006} model produced the lowest correlation values. When the Figure 5 is examined, although the T-Q_{2006} model does not give values as high as the precipitation data in the models created with only temperature data, it has produced a suitable coefficient within itself, especially in terms of the non-linear model.

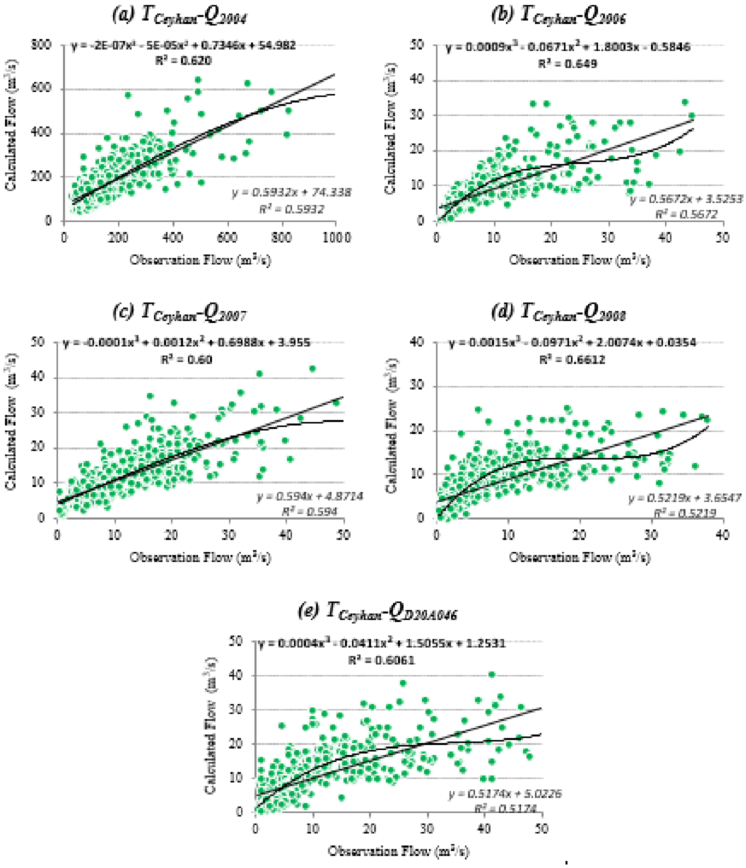


Figure 5. Graphs from linear and non-linear regression analyses for temperature invariable data

T-Q$_{2007}$ model produced the best linear results. Model T-Q$_{D20A046}$ revealed the lowest values for both the regression ones. Figure 6 shows that P-T-Q$_{2007}$ model has been showed up the best linear model when both precipitation and temperature data have been considered. In the non-linear case, there is a convergence to the model trial considered with only precipitation data. The least correlation has been observed in 2006 streamflow input data in the P-T-Q model options.

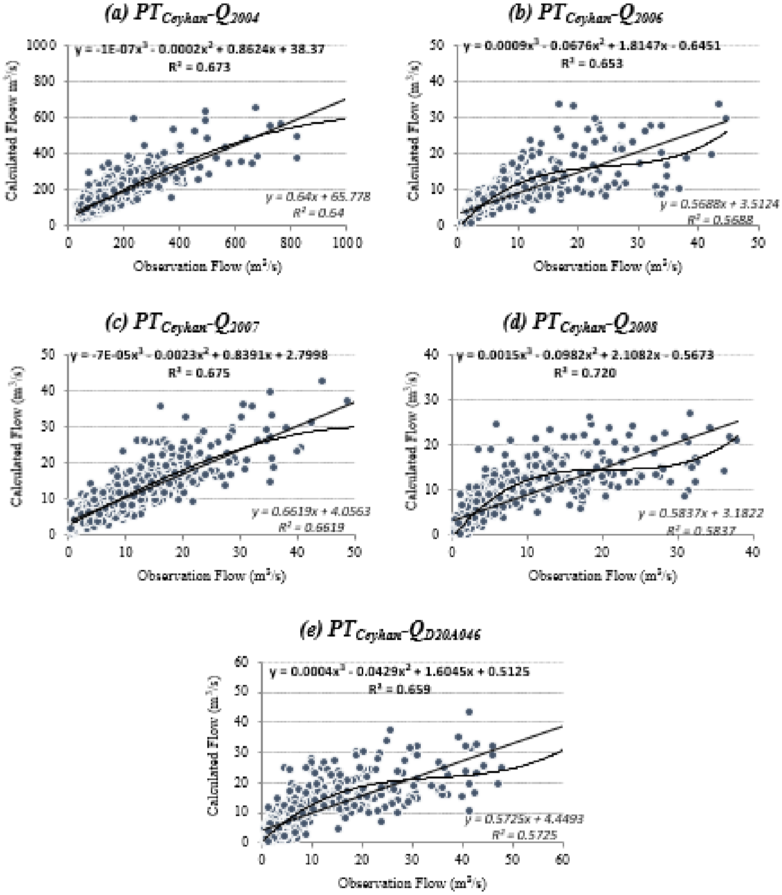


Figure 6. Graphs from linear and non-linear regression analyses for both precipitation and temperature invariable data

Figure 7 presents three-dimensional visualizations of the best model results using the Datafit. These visualizations provide perceptible data to explain the obtained results with the models. In addition, the obtained numerical results within the scope of this study are entirely shown in Table 2. These figures (Figure 7 (a), (b) and (c)) are related to P-Q_{2007}, T-Q_{2006} and P-T-Q_{2007} models that gave the best results. Error rate results also support this sentence.

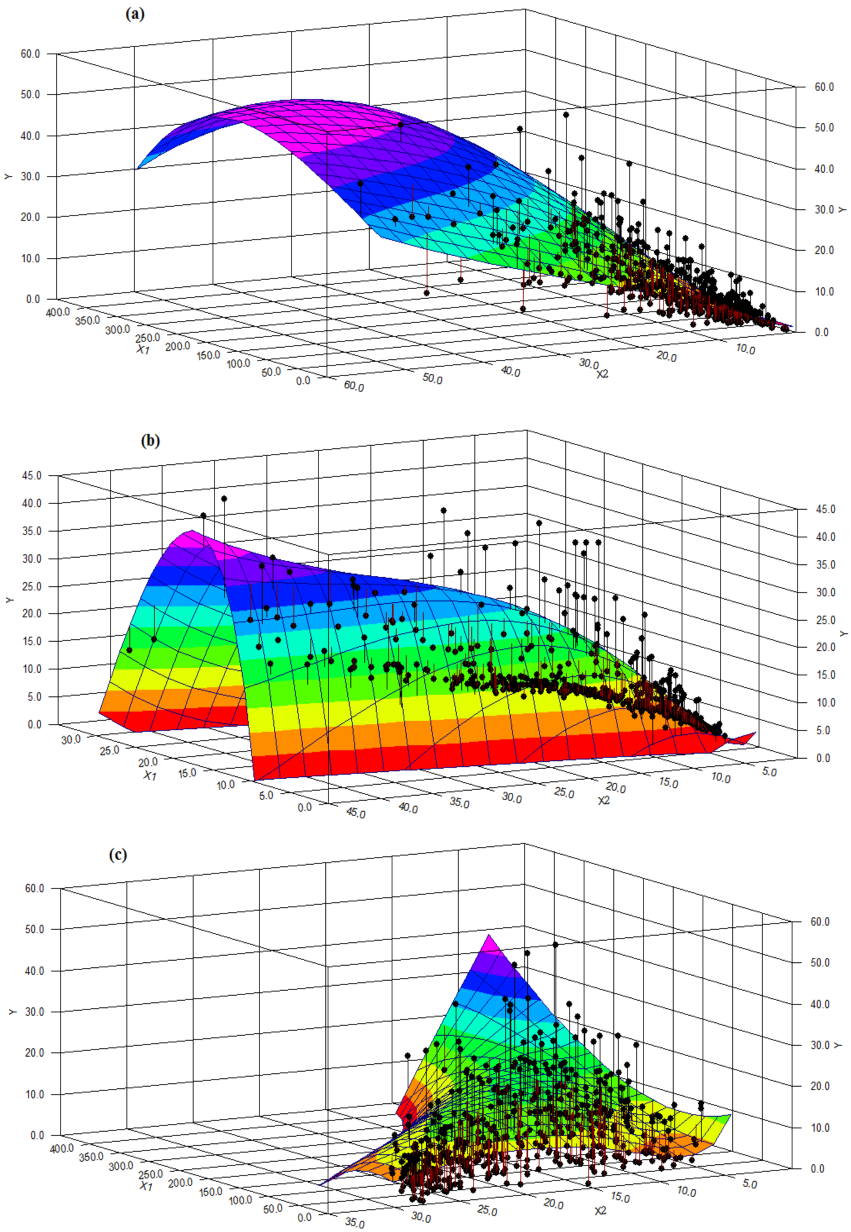


Figure 7. Datafit visualization outputs

Table 2. All numerical results

Model Number	Model Data	*Analysis Results*		
		Linear Regression	*Non-Linear Regression*	*RMSE for Non-Linear Regression*
		R^2		*Evaluation Criteria*
1	P-Q_{2004}	0.636	0.668	83.93150
2	P-Q_{2006}	0.536	0.598	5.43896
3	P-Q_{2007}	0.660	0.674	4.93074
4	P-Q_{2008}	0.569	0.678	5.26998
5	P-$Q_{D20A046}$	0.559	0.631	7.09216
6	T-Q_{2004}	0.593	0.620	88.74535
7	T-Q_{2006}	0.567	0.650	5.25173
8	T-Q_{2007}	0.594	0.607	5.38672
9	T-Q_{2008}	0.522	0.661	5.54866
10	T-$Q_{D20A046}$	0.517	0.606	7.42000
11	P-T-Q_{2004}	0.640	0.674	83.48010
12	P-T-Q_{2006}	0.569	0.654	5.24191
13	P-T-Q_{2007}	0.662	0.676	4.91549
14	P-T-Q_{2008}	0.584	0.721	5.17725
15	P-T-$Q_{D20A046}$	0.572	0.660	6.98365

CONCLUSIONS

In the study, five streamflow observation stations' data no: 2004, 2006, 2007, and 2008 belonging to the former General Directorate of Electric Power Resources Survey and Development Administration (Turkish: EİEİ) and no: D20A046 belonging to the General Directorate of State Hydraulic Works (Turkish: DSİ) located within the borders of the Ceyhan Basin, Türkiye were used. Precipitation and temperature data from the Ceyhan station of Turkish State Meteorological Services (Turkish: MGM) were supposed as fixed independent variables, and other station data were tested with linear and non-linear regression analyses at the suitable correlation level. At first, non-linear regression analysis was executed via a macro software created in the MS-Excel. Further, Datafit 9.0 was utilized for non-linear regression modeling. The obtained results were handled according to different evaluation criteria such as coefficient of determination (R^2), and Root Mean Square Error (RMSE). In the non-linear regression analysis, the previous month's streamflow data were included to the models as an independent variable. The all streamflow stations were added into the models as dependent variables. Temperature data of Ceyhan station were evaluated in the same time period. Eventually, models were tried in which both precipitation and temperature data were noted as independent variables. As a result, the highest R^2 value has been acquired with the P-Q_{2007} model. The P-Q_{2006} model gave the lowest correlation values. T-Q_{2007} model produced the best linear results. Model T-$Q_{D20A046}$ revealed the lowest values for both the regression ones. For the non-linear case, there is a convergence to the model trial considered with only precipitation data. The least correlation has been observed in 2006 streamflow input data in the P-T-Q model options. The error rate values are also quite similar to the R^2 results of the models. For future studies, it can be examined modeling with different input data, including topographic and geographic characteristics, to assess the compatibility of these results with the published literature. Accurately estimating missing streamflow data is crucial for water resources management, and considering the design and implementation of planned water structures will enable both effective engineering solutions and cost-benefit analysis to be achieved at the desired levels. Furthermore, it is thought that integrating different input data within the same basin boundaries, with regard to basin characteristics, can improve the results.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to the former General Directorate of Electric Power Resources Survey and Development Administration (EİEİ), the General Directorate of State Hydraulic Works (DSİ), and Turkish State Meteorological Services (MGM) for providing the current data utilized within the scope of this study.

REFERENCES

[1]. Seçkin, N., Güven, A., & Yurtal, R. (2016). Modelling Flood Discharge Using Artificial Neural Network: Case Study-The Middle Black Sea Watershed. Çukurova University Journal of the Faculty of Engineering, 25(1), 45-56.

[2]. Ercan, B., Yağcı, A. E., & Ünsal, M. (2019). Regression Analysis of the Kahramanmaraş Aksu River Flow Forecasting. International Symposium on Advanced Engineering Technologies (ISADET), 1-3.

[3]. Esha, R., & Imteaz, M. A. (2024). Nonlinear multiple regression analysis for predicting seasonal streamflow using climate indices for New South Wales. Int. J. Hydrology Science and Technology, 17(1), 101-116. https://dx.doi.org/10.1504/IJHST.2024.10060429

[4]. Zaw, W. T., & Naing, T. T. (2008). Empirical Statistical Modeling of Rainfall Prediction over Myanmar. World Academy of Science, Engineering and Technology, 46, 565-568.

[5]. Rezaeianzadeh, M., Tabari, H., Arabi Yazdi, A., Isik, S., & Kalin, L. (2014). Flood flow forecasting using ANN, ANFIS and regression models. Neural Comput & Applic 25, 25–37. https://doi.org/10.1007/s00521-013-1443-6

[6]. Fuladipanah, M., Hazi, M. & Kisi, O (2023). An in-depth comparative analysis of data-driven and classic regression models for scour depth prediction around cylindrical bridge piers. Appl Water Sci, 13, 231. https://doi.org/10.1007/s13201-023-02022-0

[7]. Dehbalaei, F.N., Azari, A. & Akhtari, A.A (2023). Development of a linear–nonlinear hybrid special model to predict monthly runoff in a catchment area and evaluate its performance with novel machine learning methods. Appl Water Sci 13, 118. https://doi.org/10.1007/s13201-023-01917-2

[8]. Londhe S.N., & Gavraskar, S. (2018). Stream flow forecasting using least square support vector regression. J Soft Comput Civ Eng, 2(2), 56–88. https://doi.org/10.22115/scce.2017.96717.1024

[9]. Ministry of Agriculture and Forestry (Turkish: TOB), General Directorate of Water Management (2018). Ceyhan Basin Flood Management Plan, Ankara, Türkiye.

[10]. General Directorate of Electric Power Resources Survey and Development Administration (Turkish: EİEİ) (2011). Flow Observation Yearbooks, Ankara. Access date: 28.02.2023, https://www.dsi.gov.tr/Sayfa/Detay/744

[11]. General Directorate of State Hydraulic Works (Turkish: DSİ) (2020). Flow Observation Yearbooks, Ankara. Access date: 28.02.2023, https://www.dsi.gov.tr/Sayfa/Detay/744

[12]. Turkish State Meteorological Services (Turkish: MGM) (2025). Climate Analysis. Access date: 01.10.2025, https://mgm.gov.tr

[13]. Zengin, H., Özcan, M., Değermenci, A. S., & Çitgez, T. (2023). Multiple linear regression models for the estimation of water flows for forest management and planning in Türkiye. Water SA, 49 (3), 220-229. https://doi.org/10.17159/wsa/2023.v49.i3.4000

[14]. Turhan, E. (2021). A Comparative Evaluation of the Use of Artificial Neural Networks for Modeling the Rainfall–Runoff Relationship in Water Resources Management. Journal of Ecological Engineering, 22(5), 166–178. https://doi.org/10.12911/22998993/135775

[15]. Turhan, E., & Değerli Şimşek, S. (2023). Ceyhan Havzası Akım Gözlem İstasyonu Verileri Kullanılarak Regresyon Analizi. 4th International Symposium of Engineering Applications on Civil Engineering and Earth Sciences (IEACES), Karabük, Türkiye, 520-531.

[16]. Turhan, E. (2022). Streamflow Prediction Study with Sequential Gauging Station Data Using Multiple Nonlinear Regression Method. The Black Sea Journal of Sciences, 12 (2), 931-945. https://doi.org/10.31466/kfbd.1175582

[17]. Esha, R. I., & Imteaz, M. A (2017). Non-linear multiple regression analysis for predicting seasonal streamflow using large scale climate mode. 22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 1586-1592.

[18]. Pandey, G. R., & Nguyen, V. T. V. (1999). A comparative study of regression-based methods in regional flood frequency analysis. Journal of Hydrology, 225, 92-101.

[19]. Uzoukwu, R. A., Agunwamba, J. C., Okoro, B. C., Osuagwu, J. C., & Nwoke, H. U. (2025). Streamflow Modelling of River Niger at Lokoja and Onitsha in Nigeria for Water Resources Development and Management. Saudi J Civ Eng, 9(4), 105-111.

[20]. Ruya M., Günal, A. Y. (2023). A Comparative Study of Using Adaptive Neural Fuzzy Inference System (ANFIS), Gaussian Process Regression (GPR), and SMRGT Models in Flow Coefficient Estimation. 3C Tecnología, Glosas de innovación aplicada a la pyme, 12(2), 125-146. https://doi.org/10.17993/3ctecno.2023.v12n2e44.125-146

Chapter 4

Nutritional Composition and Digestibility of Wheat Straw

Hüseyin NURSOY[1]

INTRODUCTION

Wheat (Triticum spp.) was first cultivated in Mesopotamia approximately 3000 years ago and has spread throughout the world. Anatolia is a very important genetic center harboring the wild relatives of wheat. Today, wild varieties of wheat such as spelt (Triticum monococcum L. subsp. monococcum), emmer wheat (Triticum dicoccum), and miracle wheat (Triticum turgidum var. mirabile) still exist, as well as two main cultivated varieties: bread wheat (Triticum aestivum subsp. aestivum) and durum wheat (Triticum turgidum L. var. Durum Desf.). These varieties have hundreds of different genotypes (Zaharieva and Monneveux, 2014; Işık and Giachno, 2025; Akdağoğlu and Başer, 2025).

Wheat is cultivated on approximately 14 million hectares worldwide, with an annual grain production of 34.5 million tons (Erbaş, 2025). In Turkey, approximately 21 million tons of wheat (Triticum spp.) were produced in 2024. However, in 2025, during the driest summer in the last 55 years, this value decreased by 15% to approximately 16.3 million tons. In 2024, the highest wheat production occurred in Konya, Şanlıurfa, and Diyarbakır. According to 2023 data, wheat yield in Turkey ranges from 278-354 kg/decar. These figures are 586 kg/da in China, 550 kg/da in the European Union, and 359 kg/da on average worldwide (Koç et al. 2022; Erbaş, 2025).

Straw is the cellulose and lignin-rich portion remaining after the seeds are harvested from plants that have completed their vegetation period and are in the generative stage. In general, the straw of leguminous plants is richer in protein and calcium compared to the straw of cereals.

---

[1] Bingöl University, Faculty of Veterinary Medicine, Department of Animal Nutrition and Nutritional Diseases, 12000, Bingöl, Turkey
ORCID: 0000-0002-5524-2459
hnursoy@bingol.edu.tr

Harvest Index (HI) = Grain Weight / Grain + Straw Weight x 100 (Idan and Oleiwi, 2024)

Straw Yield = (1-HI / HI) x Grain Yield (Dai et al. 2016)

Globally, HI ranges from 30-60% (Dai et al. 2015). Studies conducted in Turkey (Koç et al., 2022; Balkan et al., 2024; Çetin and Ayrancı, 2021; Işık and Giachno, 2025) show that the average wheat HI ranges from 22.2-54.39%, or if we assume an average of 35% grain and therefore 55% straw, and if our annual wheat production is approximately 20 million tons, then we roughly have an annual wheat straw production of 30 million tons.

NUTRITIONAL COMPOSITION OF WHEAT STRAW

According to the latest information, the nutrient content of wheat straw is summarized in Table 1.

Table 1. Nutrient composition of wheat straw

Parameters	References		
	(Kumar et al., 2015; Gürsoy 2023; Nazar et al., 2025; Chen et al., 2025)	NASEM, 2021	INRA, 2025
Dry Matter (DM), %	89 - 96.27	89.5	90.3
DM Digestibility, %	43.47	-	25
True DM Digestibility, %	38.54	-	-
Crude Ash, %	4.4 -12.9	8.0	6.6
Insoluble Ash or Silica, %	4.5 - 7.30	-	3
Organic Matter (OM), %	88.96	81.5	83.7
OM Digestibility, %	47.13	-	44.5
True OM Digestibility, %	40.06	-	-
Crude Protein (CP), %	1.8 - 5.45	4.5	3.9
Digestible CP, %	1 - 2	-	-
RUP, % of CP	-	-	48
Crude Cellulose, %	30 - 47	-	-
Cellulose, %	33 - 39.2	-	41.6
NDF, %	74.1-87.4	76.9	78
NDF Digestibility, %	-	-	68
ADF, %	47 - 61.15	53.1	50
ADL or Lignin	8.2 -19.1	8.19	7.5
Hemicellulose, %	23.23 - 32	-	-
Ether Extract (EE), %	1.2-1.8	1.49	1.4
TDN, %	40-48	-	50.6
Nitrogen-free extract (NFE),	28 - 50	-	-

%			
CHO in NFE, %	32	-	-
Pentosanes in NFE, %	54	-	-
Other CHO in NFE, %	14	-	-
Non-Fiber Carbohydrates (NFC), %	7.2		
Sugar, %	4.34	-	1.4
Starch, %	-	1.8	0.8
Gross Energy, kcal/kg DM	-	-	4410
Digestible Energy, kcal/kg DM	-	1960	-
ME, kcal/kg DM	1111-1570	-	1510
NEL, kcal/kg DM	-	-	870
NEg, kcal/kg DM	-	-	670
Ca,%	0.18-0.25	0.38	0.44
P, %	0.05	0.12	0.7
Vitamin A 1000 IU/kg	-	-	1
Vitamin D 1000 IU/kg	-	-	0
Cation-Anion Difference, mEq/kg DM	-	-	78
Cation-Anion Difference, mEq/kg DM	60	-	-

TREATMENTS APPLIED TO WHEAT STRAW TO INCREASE ITS DIGESTABILITY

The processes applied to increase the digestibility of straw are given in Figure 1 below.

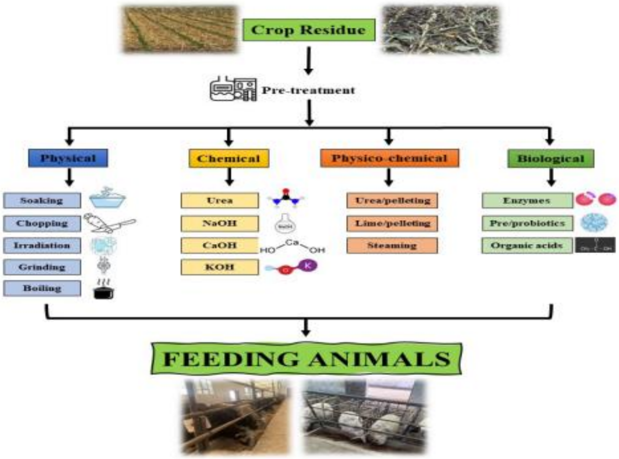


Figure 1. Processes applied to increase the digestibility of straws (Kamal et al., 2025)

Physical Treatments Applied to Wheat Straw

Physical treatments such as soaking, chopping, grinding, and boiling can be applied to increase the digestibility of straws.

Chopping

The most important physical treatment applied to wheat straw is chopping to particles between 2.5-10 cm. The particle size of roughages increases the digestion of NDF in ruminants, leading to longer rumination, higher rumen pH, and increased acetate-propionate ratio (Nursoy, 2023; Zhang et al., 2024). Conversely, reducing the particle size of roughage increases the ability of microorganisms to access and efficiently ferment them, which can lead to a decrease in methane production. The size of fibrous substrate particles affects rumen microbial fermentation but does not alter the fermentation rate (Kamal et al., 2025).

Grinding

Grinding is the crushing of straw to small sizes between 0.6-0.8 mm using milling machines. This structure can then be pelleted or directly mixed with other feeds and given to animals (Kamal et al., 2025).

Chemicals Added to Wheat Straw

Adding solutions of chemicals such as urea, NaOH, CaOH, KOH, NH_3, etc. to straw is intended to break down the lignin found in the cell wall structure. This is because lignin, composed of phenylpropane units, forms bonds with cellulose and hemicellulose. Our aim is actually to break the lignin bonds between the polymers in the cell wall of the straw, freeing the carbohydrates inside the cell and the hemicellulose in the cell wall (Kamal et al., 2025).

Urea, $(CO(NH_3)_2)$

Adding urea to straw can increase its feed value. The use of urea does not have a harmful effect on human or animal health. This process involves dissolving urea in water, promoting the growth of rumen microbes, and releasing ammonia. As a result, the straw becomes softer and more easily digestible. Because urea-treated straw provides the microbes in the rumen with the nitrogen they need to grow, multiply, and break down the straw faster, the straw is digested more quickly, thus increasing appetite and allowing the animal to eat more (Kamal et al., 2025; Malik, 2025).

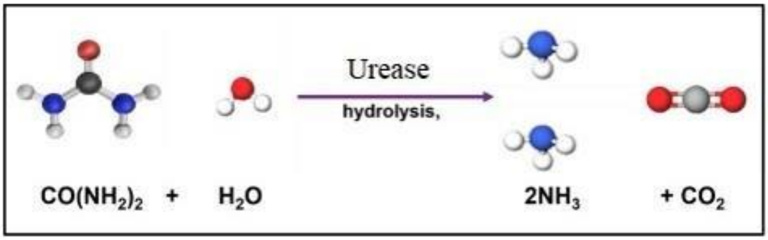


Figure 2. Formation of ammonia as a result of the combination of urea with water (Bruskiewich et al., 2024)

The recommended ratio is 4 kg of urea per 100 kg of straw (Kamal et al. 2025). A 12% urea solution is sprinkled onto wheat or barley straw at a rate of 40 l/100 kg of straw. The mixture is placed in nylon bags and thoroughly mixed. After a good mixture is achieved, the bags are sealed and fermentation is allowed for 20-60 days. At the end of each material and period, the feed is removed from the bags, aired or dried for 1-2 days, and then given to the animals (Malik, 2025).

As with any new feed, it takes 3 to 7 days for animals to get used to eating urea-treated hay, and the smell of this hay will be different from untreated hay. Adding urea can increase the protein level in the hay by 2-14%. However, the metabolizable energy (ME) value of the hay does not improve with urea application. Therefore, feeding with low levels of grains is necessary to provide extra energy (Malik, 2025; Kamal et al. 2025).

Sodium Hydroxide (NaOH), Calcium Hydroxide (CaOH) and Potassium Hydroxide (KOH)

Treating straw with NaOH increases cellulose digestion by 10%, while reducing SE and ME content (Kamal et al. 2025). NaOH treatment of straw is applied as follows: 0.6 kg of dry NaOH is weighed, mixed with 30 liters of water to obtain a 1.5% NaOH solution. 10 kg of straw remains in this solution for 18-20 hours. Then, it is rinsed with clean water for 3-4 hours, then left to mature in the open for 3-6 days, and then fed to ruminants (Malik, 2025).

Biological Agents Added to Wheat Straw

Fungi

Basidiomycetes fungi They are able to break down lignin. White rot fungi can break down plant cell components and convert them into simpler sugars. This process occurs when fungi colonize a suitable substrate and convert it into easily digestible carbohydrates. This is called the fungus's primary metabolism. The fungus completely consumes these sugars, and then secondary metabolism begins. This process breaks down structural polysaccharides and lignin from substrates using extracellular enzymes such as laccase, manganese peroxidase, and peroxidase, separating cellulose from lignin, and increasing the NDF content from 10% to over 70% (Kamal et al. 2025, Figure 3).

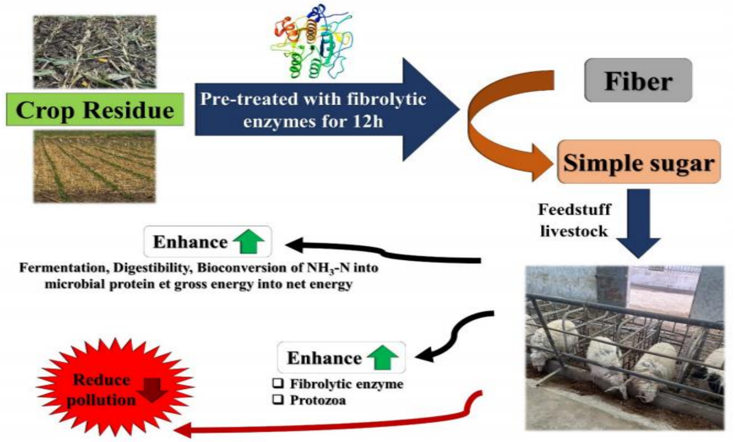


Figure 3. Fungal enzymes break down lignin, releasing cellulose and sugars (Kamal et al. 2025)

Bacteria

Bacteria can be used to improve the digestibility of hay through their extracellular enzymes such as xylanase, cellulase, and pectinase. For example, unique bacterial species such as Megasphaera elsdenii and Prevotella bryantii are used, as well as commonly used bacteria such as Bacillus, Enterococcus, Streptococcus, Bifidobacterium, Propionibacterium, and Lactobacillus. Potentially, bacteria improve feed efficiency, weight gain, and rumen conditions (Kamal et al. 2025).

Organic Acids

Organic acids, which are natural byproducts of microbial fermentation and plants, have been used as food preservatives for thousands of years, making them a natural choice in industry and acceptable for optimizing animal production. Organic acids such as fumaric, malic, and formic acids are replacing antibiotics due to their ability to eliminate harmful microorganisms from the gastrointestinal system, increase farm animal productivity, and protect animal health. Organic acids have a beneficial effect on the nutrition of ruminant animals and can potentially reduce rumen methane production (Kamal et al. 2025).

REFERENCES

Akdağoğlu, M., Başer, İ. Effect of different nitrogen doses and seed amounts on quality characteristics in cultivated (Triticum aestivum L.) andwild wheats (Triticum monococcum L. and Triticum turgidum var. Mirabile). Journal of Tekirdag Agricultural Faculty, 2025, 22, 2, 589-600.

Balkan A, Göçmen D.B., Bilgin O., Başer İ., Özcan K. Farklı Ekim Sıklıklarının Ekmeklik Buğday Çeşitlerinde Tane Verimi ve Verim Unsurları Üzerine Etkisi. ÇOMÜ Zir. Fak. Derg. (COMU J. Agric. Fac.) 2024, 12, 1, 43-54

Bruskiewich P., Zhang R., Zhang C.C. A Simple and Novel Model of the Urease Enzyme. Vancouver Institute for Advanced Studies, 2024. https://www.researchgate.net/ publication/382024371_A_Simple_and_Novel_Model _of_the_Urease_ Enzyme

Chen, B.; Liu, J.; Liu, M.; Zhang, H.; Li, X.; Tian, C.; Chen, Y. Synergistic Effect of Microorganisms and Enzymes on Nutritional Value of Corn Stover and Wheat Straw. *Fermentation* 2025, *11*, 210. https://doi.o,g/10.3390/fermentation11040210

Çetin GN, Ayrancı R. Kırşehir Ekolojik Koşullarında Bazı Makarnalık Buğday Çeşitlerinin Verim ve Verim Bileşenleri Bakımından Değerlendirilmesi. KUZFAD. 2021;1(1):9-20.

Dai J., Bean B., Brown B., Bruening W., Edwards J., Flowers M., Karowf R., Leed C., Morganb G., Ottman M., Ransom J., Wiersma J. Harvest Index and Straw Yield of Five Classes of Wheat Biomass and Bioenergy, 85, 2016, 223-227

Erbaş N. Yozgat ilinde yetiştirilen önemli tarla bitkilerinin üretimi ve karşılaştırmalı maliyet, gelir ve kâr analizi. Tarım Ekonomisi Dergisi**,** 2025, 31, 1, 149-158

Gürsoy E., 2023. Samanların besin değeri ve sindirilebilirliğini artırma yöntemleri. Kadirli Uygulamalı Bilimler Fakültesi Dergisi, 3, 1, 160-169.

Idan M.K., Oleiwi M.S. Effect of Inoculation with Mycorrhizae, Phosphorus Levels, and Humic Acids on some Components of Wheat Yield IOP Conf. Ser.: Earth Environ. Sci. *2024,* 1371 052067

INRA, 2025 Wheat Straw, https://www.feedtables.com/content/wheat-straw

Işık, R.,Giachno R.R. A. Performance of durum wheat landracesconcerning on grain yield and spike characteristics in Bornova ecological conditions. Ege Univ.Ziraat Fak. Derg., 2025, 62, 2, 215-228, https://doi.org/10.20289/zfdergi.1506339

Kamal, M., Aldhalmi, A. K., Abd El-Hack, M. E., Elsherbeni, A. I., Youssef, I. M., Hussein, S., ... & Cheng, Y. Enhancing the Feed Efficiency of Crop Residues in Ruminants–A Comprehensive Review. Annals of Animal Science, 2025, 25, 2, 529-545.

Koç, E., Akın, B., Olgun, M. Modern ve Yerel Buğdayların Bazı Önemli Özelliklerindeki Uzunluk Varyasyonlarının Biyolojik Verim ve Hasat İndeksi Üzerine Etkileri. Biyoloji Bilimleri Araştırma Dergisi, 2022, 14, 2, 87–97.

Kumar D, Datt C, Das LK, Kundu SS. Evaluation of various feedstuffs of ruminants in terms of chemical composition and metabolisable energy content. Vet World. 2015, 8, 5, 605-609. doi: 10.14202/vetworld.2015.

Malik P.K. Facts about feed and fodder. 2025, https://www.slideshare.net/slideshow/feed-and-fodder-status-in-india/1868989 #5

NASEM, 2021. National Academies of Sciences, Engineering, and Medicine. 2021. Nutrient Requirements of Dairy Cattle: Eighth Revised Edition. Washington, DC: The National Academies Press.https://doi.org/10.17226/12474.

Nazar, Mudasir, et al. "Biological delignification and anaerobic fermentation of wheat straw: A promising approach for sustainable utilization of crop straw bioresources." Industrial Crops and Products 2025, 227, 120839.

Nursoy H. Health and Enviroment in Veterinary Medicine, Editors: Gültekin Yıldız, Murat Sedat Baran, Oktay Kaplan, Özlem Durna Aydın, CHAPTER 6: TMR AND PARTICLE SIZE FOR HEIFERS AND COWS, 2023, Pages: 69-85, Iksad Publishing House, ISBN: 978-625-367-200-3, Ankara, Türkiye.

Zhang K, Yan Y, Zhao R, Song X, Du L, Zhang B, Yang C, Tang X. Effects of Substitution of Wheat Straw by Giant Reed on Growth Performance, Serum Biochemical Parameters, Nutrient Digestibility, and Antioxidant Properties of Sheep. Animals (Basel). 2024 Dec 20, 14, 24, 3678. doi: 10.3390/ani14243678.

Zaharieva, M., Monneveux, P. Cultivated einkorn wheat (Triticum monococcum L. subsp. monococcum): the long life of a founder crop of agriculture. Genetic Resources and Crop Evolution. (Netherlands). 2014, 61, 3, 677-706.

Chapter 5

Use of Wheat Straw in Cattle Rations

Hüseyin NURSOY[1]

INTRODUCTION

Wheat (Triticum spp.) was first cultivated in Mesopotamia approximately 3000 years ago and has spread throughout the world. Anatolia is a very important genetic center harboring the wild relatives of wheat. Today, wild varieties of wheat such as spelt (Triticum monococcum L. subsp. monococcum), emmer wheat (Triticum dicoccum), and miracle wheat (Triticum turgidum var. mirabile) still exist, as well as two main cultivated varieties: bread wheat (Triticum aestivum subsp. aestivum) and durum wheat (Triticum turgidum L. var. Durum Desf.). These varieties have hundreds of different genotypes (Zaharieva and Monneveux, 2014; Işık and Giachno, 2025; Akdağoğlu and Başer, 2025).

Wheat is cultivated on approximately 14 million hectares worldwide, with an annual grain production of 34.5 million tons (Erbaş, 2025). In Turkey, approximately 21 million tons of wheat (Triticum spp.) were produced in 2024. However, in 2025, during the driest summer in the last 55 years, this value decreased by 15% to approximately 16.3 million tons. In 2024, the highest wheat production occurred in Konya, Şanlıurfa, and Diyarbakır. According to 2023 data, wheat yield in Turkey ranges from 278-354 kg/decar. These figures are 586 kg/da in China, 550 kg/da in the European Union, and 359 kg/da on average worldwide (Koç et al. 2022; Erbaş, 2025).

Straw is the cellulose and lignin-rich portion remaining after the seeds are harvested from plants that have completed their vegetation period and are in the generative stage. In general, the straw of leguminous plants is richer in protein and calcium compared to the straw of cereals.

Harvest Index (HI) = Grain Weight / Grain + Straw Weight x 100 (Idan and Oleiwi, 2024)

Straw Yield = (1-HI / HI) x Grain Yield (Dai et al. 2016)

---

[1] Bingöl University, Faculty of Veterinary Medicine, Department of Animal Nutrition and Nutritional Diseases, 12000, Bingöl, Turkey
ORCID: 0000-0002-5524-2459
hnursoy@bingol.edu.tr

Globally, HI ranges from 30-60% (Dai et al. 2015). Studies conducted in Turkey (Koç et al., 2022; Balkan et al., 2024; Çetin and Ayrancı, 2021; Işık and Giachno, 2025) show that the average wheat HI ranges from 22.2-54.39%, or if we assume an average of 35% grain and therefore 55% straw, and if our annual wheat production is approximately 20 million tons, then we roughly have an annual wheat straw production of 30 million tons.

NUTRITIONAL COMPOSITION OF WHEAT STRAW

According to the latest information, the nutrient content of wheat straw is summarized in Table 1.

Table 1. Nutrient composition of wheat straw

Parameters	References		
	(Kumar et al., 2015; Gürsoy 2023; Nazar et al., 2025; Chen et al., 2025)	NASEM, 2021	INRA, 2025
Dry Matter (DM), %	89 - 96.27	89.5	90.3
DM Digestibility, %	43.47	-	25
True DM Digestibility, %	38.54	-	-
Crude Ash, %	4.4 -12.9	8.0	6.6
Insoluble Ash or Silica, %	4.5 - 7.30	-	3
Organic Matter (OM), %	88.96	81.5	83.7
OM Digestibility, %	47.13	-	44.5
True OM Digestibility, %	40.06	-	-
Crude Protein (CP), %	1.8 - 5.45	4.5	3.9
Digestible CP, %	1 - 2	-	-
RUP, % of CP	-	-	48
Crude Cellulose, %	30 - 47	-	-
Cellulose, %	33 - 39.2	-	41.6
NDF, %	74.1-87.4	76.9	78
NDF Digestibility, %	-	-	68
ADF, %	47 - 61.15	53.1	50
ADL or Lignin	8.2 -19.1	8.19	7.5
Hemicellulose, %	23.23 - 32	-	-
Ether Extract (EE), %	1.2-1.8	1.49	1.4
TDN, %	40-48	-	50.6
Nitrogen-free extract (NFE), %	28 - 50	-	-
CHO in NFE, %	32	-	-
Pentosanes in NFE, %	54	-	-
Other CHO in NFE, %	14	-	-
Non-Fiber Carbohydrates (NFC), %	7.2		

Sugar, %	4.34	-	1.4
Starch, %	-	1.8	0.8
Gross Energy, kcal/kg DM	-	-	4410
Digestible Energy, kcal/kg DM	-	1960	-
ME, kcal/kg DM	1111-1570	-	1510
NEL, kcal/kg DM	-	-	870
NEg, kcal/kg DM	-	-	670
Ca,%	0.18-0.25	0.38	0.44
P, %	0.05	0.12	0.7
Vitamin A 1000 IU/kg	-	-	1
Vitamin D 1000 IU/kg	-	-	0
Cation-Anion Difference, mEq/kg DM	-	-	78
Cation-Anion Difference, mEq/kg DM	60	-	-

USE OF WHEAT STRAW IN CATTLE RATIONS

Use of Wheat Straw for Calves

Wheat straw, even if finely chopped, is not added to the starter feeds for the first 15 calves (Bagheri et al. 2021).

It is reported that the use of wheat straw in calf starter feeds causes a decrease in calf diarrhea, keeps rumen pH stable, and leads to significant improvements in feed consumption and live weight gain. In calves given calf starter feed and straw ad libitum, straw consumption in weeks 5, 6, 7 and 8 is reported to be 15, 29, 57 and 87 g, respectively. When cereal stalks are used as bedding, it is observed that calves consume a certain amount of bedding (Çoşkun, 2021).

Bagheri et al. (2021) provided newborn Holstein calves with freely available chopped wheat straw in 1 mm, 4 mm, and 7 mm sizes alongside their starter feed from day 15 to day 90. The researchers weaned the calves on day 56 and found that 1 mm chopped wheat straw was consumed more (151 g/day), but there was no difference in live weight gain. On days 49 and 63 of the study, calves supplemented with wheat straw spent more time eating starter and roughage, lying down, and ruminating compared to the control group, and less time standing and engaging in non-feeding behaviors. Furthermore, the rumen pH of calves given wheat straw was higher than that of the control group calves 4 and 8 hours after starting feed was given on day 35. In conclusion, free-choice wheat straw supplementation increased solid feed intake, rumen pH, and calf welfare. However, they reported that the particle size of the wheat straw had no effect.

Rashid et al. (2025), in their studies on calves with an average live weight of 70 kg, investigated the replacement of the group given TMR rations with 30% wheat straw with wheat straw treated with 0.5% urea + 5% molasses. Feed consumption increased in the urea + molasses groups compared to the untreated wheat straw group. Studies have reported that increasing the proportion of wheat straw treated with urea from 0% to 100% led to significant improvements in dry matter intake (from 2931 g/day to 4034 g/day) and organic matter intake (from 2596 g/day to 3623 g/day). Digestibility of dry matter, organic matter, crude protein, and crude fiber also showed an increasing trend, reaching 77.42%, 81.21%, 87.10%, and 60.22%, respectively, at the highest urea-treated wheat straw levels. Furthermore, calves fed with 100% urea-treated wheat straw showed significantly higher weight gain and feed utilization rates compared to other groups. In conclusion, researchers report that adding higher levels of urea + molasses-treated wheat straw to the TMR increases nutrient digestibility and supports growth performance and feed efficiency in Azikheli breed buffalo calves.

Use of Wheat Straw for Lactating Cows

In dairy cows, during the first month of fresh milk production, a maximum of 0.5-1 kg/day of straw can be given to the cows. It is used to make the DM content of the TMR ration 50%, to prevent a decrease in the number of ruminations, to reduce the risk of acidosis and to prevent decreases in milk fat percentage, to prevent feed selection, and to ensure the stability of rumen pH (Havekes et al. 2020; Coşkun, 2021).

Coon et al. (2018) reported that Holstein cows fed short straw at 21% of the TMR during the dry period and 9% of the TMR during the first 28 days of lactation, with particle sizes of 5.08 cm and 2.54 cm, consumed similarly (19 kg/day), and that milk yields were similar at an average of 37 kg/day. However, cows fed short straw tended to produce 75 kg more milk cumulatively during the first 28 days of lactation compared to cows fed long straw. The researchers emphasize that the wheat straw feed particle sizes consumed in early lactation rations (1.71 kg/day) should be neither too long (greater than 5.08 cm) nor too short (less than 2.54 cm).

Use of Wheat Straw for Dry Cows and Heifers

The fact that high-yielding dairy cows still maintain high milk yields even when it's time to dry off makes drying them difficult. In this case, coercive practices such as giving the cows very little or no feed, or restricting water, may be resorted to. Although it has been reported that using only straw for 3-5 days still causes stress in cows and increases cortisol, NEFA, and BHBA levels, this method is applied in most dairy farms (Çoşkun, 2021).

While the pre-partum dry period ration is bulky and low in energy, the post-partum ration is generally more energy-dense and has a smaller average particle size. Feed selection within the TMR during the dry period and early lactation is a potential behavioral problem in cows. To address this problem, Havekes et al. (2020) used wheat straw in both long (10.16 cm) and short (2.54 cm) forms in their study where corn silage was 36% and wheat straw was 29%. In their study, they reported that there was no difference between long and short straws in dry period average dry matter consumption of 15.3 kg/day, rumination time of 498 minutes/day, and feeding time of 205 minutes/day. Similarly, they stated that there was no difference between long and short straws in the first 4 weeks when wheat straw was used in the TMR at a rate of 2% during the lactation period, with an average milk yield of 39.1 kg/day, dry matter consumption of 19.3 kg/day, rumination time of 512 minutes/day, and feeding time of 187 minutes/day. As a result, it can be calculated that the consumption of long and short-cut wheat straw during the dry period is 15.3 x 29% = 4.43 kg/day, and the wheat straw consumption during lactation is 19.3 x 2% = 386 g/day.

Coon et al. (2018) used wheat straw up to 21% in the TMR ration during the dry period in Holstein cows.

Wheat Straw Use for Beef Cattle

In Turkey, wheat straw is the most commonly used roughage in traditionally practiced cattle fattening. In fattening, which is mostly based on concentrate feeds, the cellulose and ballast material requirements are met if 15-20% wheat straw is used (Çoşkun, 2021).

Montenegro et al. (2025) reported that increasing wheat straw by 5% and 10% in finishing rations with an average starting live weight of 366 kg negatively affected some carcass characteristics and live performance measures; however, the overall performance of calves fed with a low 5% wheat straw ration was similar to that of calves fed with a control ration.

Zhang et al. (2022) reported that rations with 14.65% corn silage and 9% wheat straw in the TMR had an in situ rumen degradability of 74.3%, and that live weight gain, dry matter intake, carcass characteristics, meat quality, amino acids in meat, and most blood parameters were similar to the group with 24.24% corn silage in the TMR. They also reported that GCAA was lower in the straw group, YYO was higher in the straw group, and that the cost decreased with straw feeding.

Use of Wheat Straw for Feeding Breeding Female and Male Cattle

The goal for beef breed female cattle is to obtain one calf per year. For this purpose, quality pastures are used during the spring and summer months. In winter months, economical feeding is achieved by using wheat straw as much as possible. The daily nutrient requirements of beef cattle are lower than those of dairy breeds. For example, a female beef cattle can be fed at a maintenance level for half the year, except for the last two months of pregnancy and the calf nursing period which lasts approximately 4-6 months. If the hay is treated with ammonia or urea, the protein needs of breeding beef cattle can be largely met (Coşkun, 2021).

In their study, Males et al. (1982) gave Angus cows aged 3 years and older rations containing 100% alfalfa hay and 75% wheat straw + 25% barley-urea as winter feed, and it is also recommended to add at least 0.65 kg of crude protein source feed to cows weighing 460-480 kg that overwintered with a high proportion of wheat straw.

REFERENCES

Akdağoğlu, M., Başer, İ. Effect of different nitrogen doses and seed amounts on quality characteristics in cultivated (Triticum aestivum L.) andwild wheats (Triticum monococcum L. and Triticum turgidum var. Mirabile). Journal of Tekirdag Agricultural Faculty, 2025, 22, 2, 589-600.

Bagheri, A.A. Alamouti, M.A. Norouzian, M. Mirzaei, M.H. Ghaffari. Effects of wheat straw particle size as a free-choice provision on growth performance and feeding behaviors of dairy calves. Animal, 2021, 15, 2, 100128, 1751-7311, https://doi.org/10.1016 /j. animal.2020.100128.

Balkan A, Göçmen D.B., Bilgin O., Başer İ., Özcan K. Farklı Ekim Sıklıklarının Ekmeklik Buğday Çeşitlerinde Tane Verimi ve Verim Unsurları Üzerine Etkisi. ÇOMÜ Zir. Fak. Derg. (COMU J. Agric. Fac.) 2024, 12, 1, 43-54

Chen, B.; Liu, J.; Liu, M.; Zhang, H.; Li, X.; Tian, C.; Chen, Y. Synergistic Effect of Microorganisms and Enzymes on Nutritional Value of Corn Stover and Wheat Straw. Fermentation 2025, 11, 210. https://doi.o,g/10.3390/fermentation11040210

Coon R.E., T.F. Duffield, DeVries T.J. Effect of straw particle size on the behavior, health, and production of early-lactation dairy cows, Journal of Dairy Science, 2018, 101, 7, 6375-6387, https://doi.org/10.3168/jds.2017-13920.

Coşkun, B. 2021. Hayvan Beslemede Hiç Saman Kullanmayalım Mı? Türkiye Hayvancılığında Kaba Yem Sorunu ve Çözüm Yolları Çalıştayı, Muş Alparslan Üniversitesi, 21-22 Haziran 2021, Çalıştay Kitabı, 116-135.

Çetin GN, Ayrancı R. Kırşehir Ekolojik Koşullarında Bazı Makarnalık Buğday Çeşitlerinin Verim ve Verim Bileşenleri Bakımından Değerlendirilmesi. KUZFAD. 2021;1(1):9-20.

Dai J., Bean B., Brown B., Bruening W., Edwards J., Flowers M., Karowf R., Leed C., Morganb G., Ottman M., Ransom J., Wiersma J. Harvest Index and Straw Yield of Five Classes of Wheat Biomass and Bioenergy, 85, 2016, 223-227

Erbaş N. Yozgat ilinde yetiştirilen önemli tarla bitkilerinin üretimi ve karşılaştırmalı maliyet, gelir ve kâr analizi. Tarım Ekonomisi Dergisi, 2025, 31, 1, 149-158

Gürsoy E., 2023. Samanların besin değeri ve sindirilebilirliğini artırma yöntemleri. Kadirli Uygulamalı Bilimler Fakültesi Dergisi, 3, 1, 160-169.

Havekes, C. D., et al. Effects of wheat straw chop length in high-straw dry cow diets on intake, health, and performance of dairy cows across the transition period. Journal of Dairy Science 2020, 103, 1, 254-271.

Idan M.K., Oleiwi M.S. Effect of Inoculation with Mycorrhizae, Phosphorus Levels, and Humic Acids on some Components of Wheat Yield IOP Conf. Ser.: Earth Environ. Sci. 2024, 1371 052067

INRA, 2025 Wheat Straw, https://www.feedtables.com/content/wheat-straw

Işık, R.,Giachno R.R. A. Performance of durum wheat landracesconcerning on grain yield and spike characteristics in Bornova ecological conditions. Ege Univ.Ziraat Fak. Derg., 2025, 62, 2, 215-228, https://doi.org/10.20289/zfdergi.1506339

Koç, E., Akın, B., Olgun, M. Modern ve Yerel Buğdayların Bazı Önemli Özelliklerindeki Uzunluk Varyasyonlarının Biyolojik Verim ve Hasat İndeksi Üzerine Etkileri. Biyoloji Bilimleri Araştırma Dergisi, 2022, 14, 2, 87–97.

Kumar D, Datt C, Das LK, Kundu SS. Evaluation of various feedstuffs of ruminants in terms of chemical composition and metabolisable energy content. Vet World. 2015, 8, 5:605-9. doi: 10.14202/vetworld.2015, 605-609.

Males J. R., McReynolds W. E., Gaskins C. T., Preston R. L., Supplementation of Wheat Straw Diets to Optimize Performance of Wintering Beef Cows, Journal of Animal Science, 1982, 54, 2, 384–390, https://doi.org/10.2527/jas1982.542384x

Montenegro B.J., Penner G.B., Lardner H.A., Larson K.A., McKinnon J.J., Gibb D., McAllister T.A., Junior G.O.R. Maximizing the utilization of wheat straw in finishing beef cattle diets with canola or flax screenings supplementation: Growth performance, carcass characteristics and economic analysis., Journal of Animal Science, Volume 103, Issue Suppl. 3, 2025, 122–123, https://doi.org/10.1093/jas/skaf300.150

NASEM, 2021.National Academies of Sciences, Engineering, and Medicine. 2021. Nutrient Requirements of Dairy Cattle: Eighth Revised Edition. Washington, DC: The National Academies Press.https://doi.org/10.17226/12474.

Nazar, Mudasir, et al. "Biological delignification and anaerobic fermentation of wheat straw: A promising approach for sustainable utilization of crop straw bioresources." Industrial Crops and Products 2025, 227, 120839.

Rashid, H.U., Khan, T., Turi, A.A. et al. Impact of urea-molasses treated wheat straw levels in total mixed rations on growth and nutrient digestibility in Azikheli buffalo calves. Trop Anim Health Prod. 2025, 57, 110, https://doi.org/10.1007/s11250-025-04363-1

Zhang H, Zhang L, Xue X, Zhang X, Wang H, Gao T, Phillips C. Effect of feeding a diet comprised of various corn silages inclusion with peanut vine or wheat straw on performance, digestion, serum parameters and meat nutrients in finishing beef cattle. Anim Biosci. 2022, 35,1, 29-38. doi: 10.5713/ab.21.0088.

Zaharieva, M., Monneveux, P. Cultivated einkorn wheat (Triticum monococcum L. subsp. monococcum): the long life of a founder crop of agriculture. Genetic Resources and Crop Evolution. (Netherlands). 2014, 61, 3, 677-706.

Chapter 6

The Dıscrepancy Between Fınıte Element Results and Experımental Tests: Roots and Solutıons

İzel EKİNCİ[1], Daver ALİ[2]

Abstract

Finite Element Analysis (FEA) has become one of the most widely used computational tools for predicting the mechanical behavior of engineering systems across various disciplines, including mechanical, civil, aerospace, and biomedical engineering. Despite its extensive use, discrepancies between finite element predictions and experimental test results remain a persistent and often unavoidable issue. These discrepancies may stem from simplifications in numerical modeling, uncertainties in experimental measurements, limitations of constitutive material models, and numerical approximations inherent to the finite element method itself. This chapter presents a comprehensive and critical examination of the origins of discrepancies between finite element results and experimental observations. The discussion systematically addresses modeling assumptions, material idealizations, boundary condition definitions, geometrical representations, numerical errors, and experimental uncertainties. Furthermore, state-of-the-art strategies for reducing these discrepancies are explored, including advanced material modeling, improved experimental characterization, verification and validation frameworks, and uncertainty quantification techniques. By clarifying the underlying causes of mismatch and proposing practical solutions, this chapter aims to enhance the reliability, credibility, and predictive power of finite element models in engineering applications.

Keywords: Finite Element Analysis; Experimental Validation; Material Characterization; Boundary Conditions; Uncertainty Quantification; Verification and Validation

---

[1] Student; Karabük University Faculty of Engineering and Natural Sciences Department of Biomedical Engineering. izelbagriyanik@gmail.com ORCID No: 0009-0006-9198-0324

[2] Assoc.Prof.Dr; Karabük University Faculty of Engineering and Natural Sciences Department of Biomedical Engineering. daverali@karabuk.edu.tr ORCID No: 0000-0002-8500-7820

1. Introduction

FEA has evolved from a specialized research method into an indispensable cornerstone of modern engineering analysis and design (Zienkiewicz & Taylor, 2005). In the contemporary engineering landscape, the ability to simulate complex geometries, characterize heterogeneous materials, and predict nonlinear physical phenomena is no longer a luxury but a necessity (Bathe, 2006). This computational capability enables engineers and researchers to significantly mitigate the financial burden of extensive physical prototyping, minimize experimental effort, and accelerate the innovation cycle (Oberkampf & Trucano, 2002a).

In high-stakes applications, particularly within safety-critical fields such as biomedical engineering, aerospace, and nuclear technology, the role of FEA extends beyond merely supporting experimental observations (Viceconti et al., 2005). It is increasingly expected to serve as a predictive tool capable of forecasting mechanical behavior, failure mechanisms, and fatigue life before any physical testing. For instance, in the design of patient-specific implants or porous tissue scaffolds, in silico models are relied upon to optimize micro-architectures before the manufacturing process begins (Hollister, 2005a; Oberkampf & Trucano, 2002a). Recent studies on metamaterial-based and lattice scaffold architectures further demonstrate the increasing reliance on FEA-driven design in biomedical applications (Ekinci et al., 2025; Kavakli & Ali, 2025).

However, despite the widespread adoption and sophistication of commercial FEA solvers, a fundamental challenge persists: the discrepancy between finite element predictions and experimental test results (Roache, 1998). These discrepancies often manifest as quantifiable deviations in global stiffness, yield strength, and deformation patterns, or as qualitative differences in failure modes and crack propagation paths (Oberkampf & Trucano, 2002a). While it is common for practitioners to attribute such mismatches to simple modeling errors or user mistakes, they frequently stem from deeper, systemic issues. These issues arise from the inherent limitations of numerical discretization, the idealization of physical reality, and the uncertainties intrinsic to experimental methodologies (Ereiz et al., 2022).

The challenge is particularly exacerbated in biomedical engineering applications (Zhang et al., 2025). Constructs such as porous scaffolds, polymeric implants, and tissue-engineered matrices exhibit significant material heterogeneity (Rasheed et al., 2023). Furthermore, manufacturing processes such as Additive Manufacturing (AM) often introduce stochastic defects, surface roughness, and microporosity that are rarely captured in nominal CAD models (Mondal et al., 2022). Consequently, even under well-controlled laboratory conditions (in vitro), achieving an exact agreement between numerical simulations and experimental outputs is practically unattainable (Shetty et al., 2025).

Therefore, the objective of this chapter is not to propose methods to eliminate discrepancies, as some level of error is unavoidable in any approximation of physical reality. Rather, the goal is to provide a framework for understanding their origins, quantifying their significance, and developing strategies to manage them. By adopting a systematic and critical perspective, this chapter aims to shift the engineering paradigm from a pursuit of perfect agreement toward the establishment of reliable and validated computational models.

2. Fundamentals of Finite Element Modeling

2.1 Governing Equations and Numerical Approximation

FEA is fundamentally rooted in the numerical solution of Partial Differential Equations (PDEs) derived from the laws of continuum mechanics (Shetty et al., 2025). In structural mechanics, these governing equations typically encompass:

1. **Equilibrium Equations:** Ensuring the balance of forces and moments within the body.
2. **Constitutive Relations:** Describing the stress-strain behavior of the material (e.g., Hooke's Law for linear elasticity).
3. **Kinematic Compatibility Conditions:** Ensuring that the displacement field is continuous and physically admissible.

For simple geometries and idealized loading conditions, exact analytical solutions (closed-form solutions) can be derived. However, for the complex, irregular geometries and mixed boundary conditions encountered in real-world engineering problems, analytical solutions are intractable. This necessitates the use of numerical approximation techniques (Shetty et al., 2025; Zienkiewicz & Taylor, 2005).

The Finite Element Method (FEM) addresses this by discretizing the continuous domain into a finite number of sub-domains called "elements," connected at discrete points known as "nodes." The method approximates the continuous displacement field within each element using predefined polynomial interpolation functions, often referred to as shape functions. Mathematically, this converts the system of PDEs into a system of algebraic equations (typically represented as [K][u] = [F], where [K] is the stiffness matrix, [u] is the displacement vector, and [F] is the force vector) (Bathe, 2006; Zienkiewicz & Taylor, 2005).

It is crucial to recognize that this discretization process introduces approximation errors that are intrinsic to the method (Babuška & Suri, 1994). The solution obtained is a piecewise approximation of the true field variable. While mesh refinement (increasing mesh density) or p-refinement (increasing the polynomial order of shape functions) can reduce these errors, the numerical solution remains an approximation of the true physical behavior, never a perfect alternative (Malvern, 1969).

2.2 Assumptions of Continuum Mechanics

The validity of standard finite element models relies heavily on the axioms of continuum mechanics (Malvern, 1969). These include the assumptions of material homogeneity (properties are constant throughout the volume), isotropy (properties are direction-independent), and the principle of scale separation .

The principle of scale separation assumes that the characteristic length scale of the material's microstructure is significantly smaller than the size of the finite elements and the macroscopic dimensions of the specimen (Hill, 1963). Under this assumption, the material can be treated as a continuous medium with an averaged effective property (Belytschko et al., 2014a). However, real engineering materials often exhibit complex microstructural features, such as voids, inclusions, grain boundaries, and manufacturing-induced anisotropy, which violate these assumptions. This is particularly critical in:

- **Porous Materials:** Where the pore size may be comparable to the element size.
- **Composite Materials:** Where distinct fiber and matrix phases exist.
- **Additively Manufactured Lattice Structures:** Where the strut diameter is small relative to the unit cell.

When the characteristic length scale of these microstructural features approaches the element size or the specimen dimensions, the concept of a Representative Volume Element (RVE) may become invalid. In such cases, the continuum assumption breaks down, leading to significant discrepancies between the homogenized numerical predictions and the heterogeneous experimental observations (Oberkampf & Trucano, 2002a).

3. Experimental Testing and Its Limitations

In the validation of computational models, experimental data is frequently regarded as the ground truth against which numerical predictions must be benchmarked (Oberkampf & Trucano, 2002a). However, this assumption relies on the premise that experimental results are error-free and perfectly representative of the physical phenomenon. In reality, experimental testing is subject to a wide array of uncertainties and limitations. Understanding these limitations is crucial, as discrepancies are often erroneously attributed to the FEA when, in fact, the experimental benchmark itself may be flawed or misinterpreted (Oberkampf & Trucano, 2002b).

3.1 Variability in Material Properties

Experimental testing is inherently subject to statistical variability. Even when specimens are produced using identical manufacturing protocols and source materials, stochastic variations in microstructure, local density, and defect distribution are unavoidable (L. J. Gibson, 2003). Consequently, no two physical specimens are truly similar. This variability is particularly pronounced in advanced manufacturing techniques such as AM (DebRoy et al., 2018). In the context of 3D-printed constructs (e.g., scaffolds or lattice structures), process parameters such as nozzle temperature, layer height, and print speed have a significant influence on the resulting mechanical properties (I. Gibson et al., 2021). For instance, thermal history differences between layers can lead to variations in interlayer bonding strength, resulting in anisotropic behavior that differs from specimen to specimen (Ahn et al., 2002). Experimental studies on additively manufactured and porous structures have repeatedly highlighted that specimen-to-specimen variability can substantially affect the agreement between experimental results and numerical predictions (Ali, 2021).

3.2 Measurement and Instrumentation Uncertainty

Every measurement chain, from the physical sensor to the data acquisition system, introduces an uncertainty. Finite accuracy limits, linearity errors, and hysteresis govern load cells, extensometers, and strain gauges. For example, in the testing of soft biomedical tissues or low-stiffness scaffolds, the force response may be close to the noise floor of a high-capacity load cell, leading to a low signal-to-noise ratio that obscures the true mechanical response. Furthermore, the method of strain measurement has a significant impact on the accuracy of the results. Traditional methods, such as crosshead displacement, often incorporate the compliance of the testing machine itself, leading to an overestimation of the specimen's strain (Luecke et al., 2010). While video extensometers or Digital Image Correlation offer non-contact alternatives, they too are subject to optical resolution limits and calibration errors (Sutton et al., 2009b).

4. Material Modeling as a Major Source of Discrepancy

The heart of any FEA lies in the constitutive model, which is the mathematical description of how a material responds to physical forces. While geometry and boundary conditions define the "where" and "how" of the problem, the material model represents the "what." Consequently, the selection of an inappropriate constitutive law or the inaccurate calibration of material parameters is frequently the dominant source of error in numerical simulations (Oberkampf & Roy, 2010).

4.1 Linear Elastic Idealization

The most pervasive simplification in engineering simulation is the assumption of linear elastic isotropic behavior (Zienkiewicz & Taylor, 2005). Governed by Hooke's Law, this model assumes a constant relationship between stress and strain, characterized solely by Young's Modulus (E) and Poisson's ratio. Its popularity stems from its computational efficiency and the ease of obtaining these two parameters from standard tensile tests. While the linear elastic assumption is mathematically convenient, it is physically valid only for infinitesimal strains and stress levels that are well below the proportional limit (Bathe, 2006). However, many engineering materials, particularly polymers used in biomedical scaffolds (e.g., PLA, PCL) and biological tissues, exhibit fundamentally nonlinear behavior even at low load levels (Fung, 2013).

In experimental testing, even materials conventionally treated as elastic often experience localized phenomena that violate linearity. For instance, in porous lattice structures, stress concentrations at the nodal junctions of struts can significantly exceed the material's yield strength, leading to localized plasticity or microcracking (L. J. Gibson, 2003). A linear elastic FEA model ignores these energy dissipation mechanisms. Consequently, such models invariably overestimate the structural stiffness and underestimate the total deformation, leading to a non-conservative prediction of the structure's safety factor.

4.2 Advanced Constitutive Models

To bridge the gap between numerical predictions and experimental reality, it is often necessary to move beyond Hooke's Law and employ advanced constitutive formulations that can capture nonlinear phenomena. For metals and many polymers, the onset of permanent deformation must be modeled using plasticity theories, such as the J_2 flow theory (also known as von Mises plasticity) (Lemaitre & Chaboche, 1994). Accurate modeling requires defining not only the yield surface but also the post-yield behavior, known as hardening. Neglecting hardening behavior in FEA can lead to significant discrepancies in predicting the ultimate tensile strength (UTS) and failure displacement (Belytschko et al., 2014a). Moreover, biological tissues and polymeric biomaterials are inherently viscoelastic; their mechanical response depends on the rate of loading and the duration of the load application (Fung, 2013). They exhibit phenomena such as creep (an increase in strain under constant stress) and stress relaxation (a decrease in stress under constant strain). For soft tissues (e.g., cartilage, vessel walls) and elastomers, the relationship between stress and strain is non-linear and valid over large deformations (Holzapfel, 2000). Linear elasticity is wholly inadequate here. Instead, hyperelastic models (e.g., Neo-Hookean, Mooney-Rivlin, Ogden) are employed, which define the material behavior using a Strain Energy Density Function (W). These models are critical for accurately predicting the "locking" behavior (stiffening at high strains) seen in biological materials (Ogden, 1997).

4.3 The Challenge of Parameter Identification

The use of sophisticated material models introduces a secondary challenge: accurately identifying the parameters.

- Data Scarcity: An advanced hyperelastic or damage model may require five or more coefficients to represent the material's properties accurately (Holzapfel, 2000). Determining these requires complex multiaxial testing (e.g., uniaxial tension combined with biaxial tension and shear), which is often unavailable.
- Non-Uniqueness: Attempting to fit complex models to limited datasets (e.g., only uniaxial data) is an ill-posed inverse problem. Multiple sets of parameters might fit the training data equally well but produce vastly different predictions under complex loading states (Tarantola, 2005).

Therefore, the sophistication of the material model must be balanced against the quality of the available experimental data. A poorly calibrated advanced model can introduce more uncertainty than a well-understood simple model.

5. Geometrical Representation and Mesh Effects

While the governing equations and material models provide the mathematical and physical framework for FEA, the geometric representation constitutes the spatial domain upon which these equations are solved. A frequent and often underestimated source of discrepancy between numerical predictions and experimental results lies in the deviation between the virtual geometry used in simulation and the "physical" geometry tested in the laboratory (Oberkampf & Roy, 2010).

Finite element models are typically constructed using idealized CAD geometries (Zienkiewicz & Taylor, 2005). These virtual models consist of perfectly smooth surfaces, nominal dimensions, and mathematically precise connections. For instance, in a lattice scaffold model, a strut is often represented as a perfect cylinder with a uniform diameter. In contrast, experimental specimens invariably contain manufacturing defects, surface roughness, and geometric irregularities that significantly influence mechanical behavior (L. J. Gibson, 2003). These features are often neglected in numerical models due to computational costs or a lack of detailed geometric data (Oberkampf & Trucano, 2002b). Consequently, FEA based on perfect CAD geometry tends to be non-conservative, overpredicting the stiffness and strength of the structure compared to the experimental reality. Finite element investigations on auxetic and lattice metamaterials have shown that even minor geometric variations, such as strut cross-section shape, can lead to pronounced changes in effective stiffness and Poisson's ratio (Kavakli & Ali, 2023).

Mesh discretization plays a crucial role in determining the accuracy of finite element methods. The finite element method divides the continuum into discrete sub-domains. If this division is too coarse, the model cannot capture high gradients of stress or strain, particularly at geometric discontinuities like notches or holes (Bathe, 2006). Coarse meshes act artificially stiff, leading to an underestimation of displacement and peak stresses (Zienkiewicz & Taylor, 2005). Also, the choice of element formulation is as important as mesh density. To ensure that the results are a function of the physics and not the mesh size, mesh convergence studies are rigorously recommended. This involves simulating with successively finer meshes until the key results (e.g., max stress, total strain energy) asymptotically stabilize (Jerier et al., 2010). Although essential for reliability, these studies are frequently omitted in practice due to time constraints or computational limitations, rendering the reported results mathematically uncertain.

6. Boundary Conditions and Loading Definitions

The definition of boundary conditions (BCs) and loading protocols represents the interface between the mathematical model and the external environment. In Finite FEA, boundary conditions mathematically dictate the values of the dependent variables (degrees of freedom) at specific locations. A frequent cause of discrepancy is the erroneous assumption that the idealized mathematical constraints applied in the software perfectly mimic the complex physical interactions occurring in the test rig (Oberkampf & Roy, 2010). In FEA software, it is standard practice to use a Fixed Support condition to a surface, mathematically setting all translational and rotational degrees of freedom to zero (u_x, uy, uz = 0 and θ_x, θy, θ_z = 0) (Bathe, 2006). This assumes that the gripping mechanism is infinitely rigid and that there is absolutely no relative motion between the specimen and the grip. However, in physical experiments, achieving a perfectly rigid constraint is mechanically impossible. While FEA often assumes either frictionless or perfectly bonded contact, reality lies somewhere between these extremes. For instance, in compression testing of a scaffold between steel plates, friction restricts lateral expansion (Poisson effect) at the contact surfaces (Wriggers, 2012). If the FEA model employs a frictionless assumption, it will underpredict the confinement stress, resulting in a discrepancy in the yield behavior.

6.2 Load Application Methods

The distinction between how load is applied is critical, especially for non-linear analyses. Applying a specific force value in experiments, if the material exhibits softening (damage) or buckling, force-controlled tests become unstable and result in catastrophic failure immediately after the ultimate load is reached (De Borst et al., 2012). Instead of prescribing a fixed load, specifying a specific displacement (e.g., 1 mm/min) would yield a better outcome. This approach is preferred in both experiments and nonlinear FEA because it enables the capture of post-yield softening and snap-back behavior without numerical divergence (Belytschko et al., 2014b). Comparing a force-controlled simulation to a displacement-controlled experiment (or vice versa) can lead to fundamentally different failure predictions.

In FEA, loads are often applied directly to nodes or surfaces. According to Saint-Venant's Principle, the difference between the actual load distribution and an idealized equivalent load vanishes at a sufficiently great distance from the point of load application. However, near the loading zone, the stress distribution is highly sensitive to the exact manner in which the load is applied (e.g., a point load vs. a distributed pressure) (Bathe, 2006). Discrepancies often arise when validation data (strain gauges) are collected too close to the gripping area, where local stress concentrations from the fixture dominate the response (Luecke et al., 2010).

7. Numerical Errors and Solver Limitations

While the physical modeling assumptions (material, geometry, boundary conditions) are the most visible sources of discrepancy, the mathematical engine solving the problem is not infallible. Finite Element Analysis relies on numerical algorithms to approximate solutions to systems of equations that are often impossible to solve analytically. These algorithms are subject to limitations inherent to iterative solvers and digital floating-point arithmetic (Van Loan & Golub, 1996). Consequently, a solution that appears visually plausible may still contain significant numerical errors that deviate from the experimental reality (Oberkampf & Roy, 2010).

7.1 Nonlinear Convergence Issues

The Iterative Nature of Nonlinear Solvers

In linear analysis, the system [K][u] = [F] is solved in a single step. However, for nonlinear problems involving plasticity, large deformations, or contact, which are common in biomedical scaffolds, the stiffness matrix [K] changes as the structure deforms (Bathe, 2006). Solvers (such as the Newton-Raphson method) must iteratively update the stiffness matrix to find an equilibrium state where the internal forces balance the external loads (De Borst et al., 2012).

Convergence Tolerances and Drift

The solver stops iterating when the imbalance between external and internal forces falls below a user-defined tolerance (ε).

$$||F_{ext} - F_{int}|| < \varepsilon$$

- Loose Tolerances: To prevent "non-convergence" errors and ensure the simulation completes, users might relax these tolerances. While this aids stability, it allows the model to accept a solution that is not in true equilibrium. This "drift" from the true solution can accumulate over thousands of increments, leading to a final result that artificially deviates from the experimental curve (Babuška & Suri, 1994; Belytschko et al., 2014a).
- Instability: In problems with softening materials (e.g., damage in PLA scaffolds) or buckling struts, the stiffness matrix may become singular or non-positive definite. The solver may struggle to converge, leading to premature termination of the analysis or "chatter" (oscillation) around the solution, creating artificial noise in the stress-strain data (Bathe, 2006; De Borst et al., 2012).

7.2 Accumulation of Numerical Errors

Ill-Conditioned Matrices

The accuracy of the solution depends heavily on the condition number of the stiffness matrix (Trefethen & Bau, 2022). An ill-conditioned matrix occurs when there is a significant disparity in stiffness within the model (for example, modeling a titanium implant (with an elastic modulus of 110 GPa) in direct contact with soft marrow tissue (with an elastic modulus of 1 kPa) (Van Loan & Golub, 1996). This significant difference in elasticity can cause precision issues, as the contributions of the softer material are lost due to the finite precision of computer arithmetic (floating-point operations) (Higham, 2002).

Round-off and Truncation Errors

- **Round-off Error:** Arises from the computer's inability to represent real numbers with infinite precision. While negligible in small models, these errors can accumulate significantly in large-scale simulations with millions of degrees of freedom (DOFs) or in explicit dynamics simulations involving thousands of time steps (Higham, 2002).
- **Truncation Error:** Arises from the finite element method itself, which approximates the solution using polynomial shape functions (truncating the higher-order terms of the true solution) (Hughes, 2003).

Although modern solvers are robust, these strictly numerical sources of error can contribute to the "unexplained" gap between a simulation and a high-precision experiment, particularly in complex multi-material systems.

8. Verification and Validation Frameworks

In the engineering community, the terms verification and validation are often used interchangeably; however, they represent distinct, hierarchical processes in assessing model credibility. To address discrepancies systematically, one must adopt a formal Verification and Validation (V&V) framework, such as those established by the American Society of Mechanical Engineers (ASME V&V 10 and V&V 40). This framework provides the logic to determine whether a mismatch originates from a mathematical error (verification issue) or a physics modeling error (validation issue) (Engineers, 2006).

8.1 Verification: Ensuring Numerical Correctness

Verification is the process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model. In simpler terms, it answers the question: Are we solving the equations correctly? It is strictly a mathematical exercise and does not involve comparison with experimental data (Oberkampf & Roy, 2010). Verification consists of two distinct activities:

Code Verification

This aspect ensures that the software code (the solver algorithms) is free of bugs and mistakes. For commercial software users (e.g., Ansys, Abaqus), this is largely the responsibility of the software vendor. However, researchers writing user-defined subroutines (e.g., UMAT or VUMAT for custom scaffold materials) must perform code verification by comparing their numerical results against known exact analytical solutions for simplified problems (Oberkampf & Roy, 2010).

Calculation Verification

This is the responsibility of the analyst and focuses on estimating the magnitude of numerical errors in a specific simulation. The primary task here is to quantify the Discretization Error.

- **Mesh Convergence Studies:** The most critical step in calculation verification. It involves systematically refining the mesh size and observing the change in a key output variable (e.g., peak displacement or strain energy) (Roache, 1998).
- **Richardson Extrapolation:** A formal method used to estimate the theoretical exact solution by extrapolating from two or more finite grid solutions, thereby providing a quantitative estimate of the discretization error (Roache, 1994).

Unless a model is rigorously verified, any comparison with experiment is meaningless, as the numerical error may mask or artificially cancel out physical modeling errors.

8.2 Validation: Assessing Physical Accuracy

Validation is the process of determining the degree to which a model accurately represents the real world from the perspective of its intended uses (ASME., 2018). It answers the question: Are we solving the right equations?

The Validation Hierarchy

Validation is not a binary state (i.e., a model is neither always valid nor always invalid) (Oberkampf & Roy, 2010). It is a measurement of accuracy within a specific domain. To validate a complex system such as a full bone-implant construct a hierarchical approach is required:

1. Unit Problem: Validating the material model on a simple coupon (e.g., a tensile test of the PLA filament).
2. Benchmark Problem: Validating a structural component (e.g., a single lattice unit cell).

3. Subsystem/System Level: Validating the full complex geometry (e.g., the complete porous scaffold).

Discrepancies must be tracked and minimized at each level of this hierarchy before moving to the next. Validation involves a systematic comparison of computational predictions with experimental data (Sargent, 2010). However, visual comparison (overlaying curves) is subjective. Quantitative metrics, such as the correlation coefficient (R^2), Root Mean Square Error (RMSE), or specific indices like the CORA (Correlation and Analysis) rating, should be used to score the agreement objectively.

Crucially, validation is context-specific. A finite element model validated for static compression of a scaffold is not automatically valid for fatigue loading or dynamic impact (ASME., 2018). The "Domain of Applicability" must be clearly defined; using the model outside this domain introduces unknown errors and reintroduces the discrepancy gap.

9. Strategies for Reducing Discrepancies Between FEA and Experimental Results

Reducing the gap between virtual predictions and physical reality requires a paradigm shift from deterministic, nominal modeling to advanced, data-driven simulation strategies. This section outlines state-of-the-art approaches designed to systematically minimize discrepancies by enhancing the fidelity of both the experimental input and the numerical representation.

9.1 Improved Experimental Characterization

The quality of a finite element validation is inextricably linked to the resolution of the experimental data used to define it. Traditional testing methods often provide global averages that mask critical local behaviors.

Image-Based Geometrical Reconstruction (Micro-CT)

For complex structures, such as additively manufactured porous scaffolds or biological tissues, nominal CAD geometry is often insufficient. The most robust strategy for reducing geometric discrepancy is the use of Micro-Computed Tomography (microtomography) (Evans et al., 2023). By scanning the physical specimen before mechanical testing, researchers can generate a high-resolution 3D reconstruction of the as-manufactured geometry (du Plessis & le Roux, 2018). This captures process-induced deviations, such as strut variance, surface roughness, and microporosity. When this scanned geometry is converted into a finite element mesh (a process known as image-based meshing or voxel-based meshing), the simulation accounts for the exact material distribution present in the physical test, significantly narrowing the error margin (Hollister, 2005b).

Full-Field Strain Measurement (Digital Image Correlation)

Relying solely on global force-displacement curves provides limited validation data. Digital Image Correlation (DIC) transforms experimental validation by providing full-field surface deformation data (Sutton et al., 2009a). DIC tracks the movement of a random speckle pattern on the specimen surface using high-speed cameras. This allows for the visualization of local strain concentrations, crack initiation sites, and complex deformation bands. Validating an FEA model against DIC maps ensures that the model predicts not just the correct amount of deformation, but also the proper location and pattern of deformation, offering a much more rigorous test of model accuracy (Avril et al., 2008).

9.2 Enhanced Numerical Modeling

Reducing discrepancies often requires increasing the complexity of the numerical model to account for physical phenomena previously deemed negligible.

Incorporating Material Stochasticity and Heterogeneity

Real materials are rarely homogeneous. Enhanced modeling strategies involve mapping spatial variations in material properties directly onto the mesh (Keyak et al., 1997). For example, in bone mechanics, gray values from CT scans are often correlated to local density and Young's modulus using empirical power laws ($E = E_0(1 - \varphi/\varphi c)^f$ where E is the effective Young's modulus of the porous material with porosity φ, E_0 is the Young's modulus of the solid material, φc is the porosity at which the effective Young's modulus becomes zero, and f is a parameter that depends on the grain morphology and pore geometry of the porous material (Asef & Farrokhrouz, 2017)). Similarly, for 3D printed parts, stochastic fields can be generated to randomize material properties across the element set, simulating the natural variability of the manufacturing process.

Multi-Scale Modeling

Discrepancies often arise because macroscopic behavior is driven by microscopic events (e.g., micro-crack coalescence) that a continuum model cannot see. Multi-scale modeling (e.g., Computational Homogenization bridges this gap by coupling two scales: a Representative Volume Element (RVE) at the micro-scale calculates the constitutive response, which is then passed to the macro-scale model (Geers et al., 2010). While computationally expensive, this approach captures the structural effects of microstructure without the need for an impossibly fine mesh across the entire part (Fish & Belytschko, 2007).

9.3 Uncertainty Quantification

The most mature strategy for handling discrepancies is to stop treating them as errors to be eliminated and start treating them as uncertainties to be managed.

Probabilistic vs. Deterministic Approaches

A deterministic simulation yields a single output curve. However, given the variability in input parameters (dimensions, material properties, loading), the true output is a probability distribution. Uncertainty Quantification (UQ) frameworks, such as Monte Carlo simulations or Polynomial Chaos Expansions, propagate input uncertainties through the model to generate a probabilistic output (e.g., a 95% confidence interval for stiffness) (Smith, 2024).

Confidence Intervals

Instead of asking "Does the simulation curve match the experimental curve?", UQ allows us to ask "Does the experimental curve fall within the simulation's confidence interval?" If the experimental result falls within the predicted probabilistic bounds, the model is considered valid, even if the mean values do not align perfectly with the expected values (ASME., 2018). This statistical rigor enhances the interpretability and credibility of numerical predictions in safety-critical applications.

10. Case Studies from the Literature

To bridge the gap between theoretical frameworks and practical application, it is instructive to examine specific case studies where the discrepancy between FEA and experimental results has been rigorously documented.

11. Discussion and Future Perspectives

The persistence of discrepancies between numerical and experimental results should not be viewed as a failure of the finite element method (Roache, 1994), but rather as a driving force for the evolution of computational mechanics. As computational power increases and experimental techniques become more sophisticated, the gap between the "virtual" and the "physical" is expected to narrow (Oberkampf & Trucano, 2002b), though perhaps never to close completely.

11.1 The Shift to Probabilistic Design

A critical realization in modern engineering is that the complete elimination of discrepancies is neither realistic nor necessary. The physical world is stochastic; therefore, a single deterministic simulation cannot be expected to perfectly match a single experimental test. Future research and industrial practice are shifting toward Probabilistic Finite Element Analysis (PFEA). In this paradigm, the goal is not to achieve a perfect overlay of two curves, but to ensure that the probability distribution of the numerical output overlaps significantly with the distribution of the experimental data. This transition marks a shift from "accuracy-based" validation to "uncertainty-aware" decision-making.

11.2 Digital Twins and Real-Time Validation

The concept of the "Digital Twin," a virtual replica continuously updated with physical data, represents the next frontier. By integrating sensors into physical prototypes, experimental data can be fed back into the FEA model in real-time to update boundary conditions and material parameters. This dynamic calibration process has the potential to minimize discrepancies continuously throughout the component's lifecycle, rather than just at the design stage.

11.3 Defining "Acceptable" Error

Ultimately, the engineering community must focus on defining what constitutes an acceptable level of discrepancy for different contexts. An error margin of 20% might be acceptable for the preliminary design of a consumer product, whereas a margin of less than 5% might be mandated for a patient-specific cranial implant. Establishing standardized validation metrics (such as ASME V&V 40 for medical devices) is crucial for managing these expectations.

12. Conclusions

The discrepancy between finite element predictions and experimental test results is a multifaceted problem arising from a complex interplay of modeling assumptions, experimental uncertainties, and numerical limitations[10]. Throughout this chapter, we have identified that these mismatches are rarely mono-causal; they stem from the simplification of material behavior (e.g., linear elasticity), the idealization of geometry (e.g., neglecting roughness), the approximation of boundary conditions (e.g., infinite rigidity), and the inherent errors of numerical discretization.

Understanding the roots of these discrepancies is essential for improving the reliability and credibility of computational models[11]. We have explored that the solution lies not in unthinkingly tuning parameters to force a fit (calibration), but in adopting rigorous Verification and Validation (V&V) frameworks. By implementing advanced strategies, such as high-fidelity material modeling, micro-CT-based geometric reconstruction, and Uncertainty Quantification (UQ) engineers can transform FEA from a mere estimation tool into a robust predictive instrument.

In conclusion, while the perfect model may be an unattainable abstraction, the validated model is an achievable and powerful reality. By transparently acknowledging, quantifying, and managing discrepancies, FEA remains the most potent tool available for predicting the behavior of complex engineering systems.

REFERENCES

Ahn, S., Montero, M., Odell, D., Roundy, S., & Wright, P. K. (2002). Anisotropic material properties of fused deposition modeling ABS. *Rapid Prototyping Journal*, *8*(4), 248–257. https://doi.org/10.1108/13552540210441166

Ali, D. (2021). Mimicking Bone Anisotropic Structure with Modified Gyroid Scaffolds; A Finite Element Analysis. *Politeknik Dergisi*, *24*(4), 1637–1646. https://doi.org/10.2339/politeknik.941106

Asef, M. R., & Farrokhrouz, M. (2017). A semi-empirical relation between static and dynamic elastic modulus. *Journal of Petroleum Science and Engineering*, *157*, 359–363. https://doi.org/10.1016/j.petrol.2017.06.055

ASME. (2018). *V&V440-Assessing credibility of computational modeling through verification and validation: Application to medical devices*.

Avril, S., Bonnet, M., Bretelle, A.-S., Grédiac, M., Hild, F., Ienny, P., Latourte, F., Lemosse, D., Pagano, S., & Pagnacco, E. (2008). Overview of identification methods of mechanical parameters based on full-field measurements. *Experimental Mechanics*, *48*(4), 381–402.

Babuška, I., & Suri, M. (1994). The p and h-p Versions of the Finite Element Method, Basic Principles and Properties. *SIAM Review*, *36*(4), 578–632. https://doi.org/10.1137/1036141

Bathe, K.-J. (2006). *Finite Element Procedures*. Klaus-Jurgen Bathe.

Belytschko, T., Liu, W. K., Moran, B., & Elkhodary, K. (2014a). *Nonlinear Finite Elements for Continua and Structures*. John Wiley & Sons.

Belytschko, T., Liu, W. K., Moran, B., & Elkhodary, K. (2014b). *Nonlinear finite elements for continua and structures*. John wiley & sons.

De Borst, R., Crisfield, M. A., Remmers, J. J., & Verhoosel, C. V. (2012). *Nonlinear finite element analysis of solids and structures*. John Wiley & Sons.

DebRoy, T., Wei, H. L., Zuback, J. S., Mukherjee, T., Elmer, J. W., Milewski, J. O., Beese, A. M., Wilson-Heid, A., De, A., & Zhang, W. (2018). Additive manufacturing of metallic components – Process, structure and properties. *Progress in Materials Science*, *92*, 112–224. https://doi.org/10.1016/j.pmatsci.2017.10.001

du Plessis, A., & le Roux, S. G. (2018). Standardized X-ray tomography testing of additively manufactured parts: A round robin test. *Additive Manufacturing*, *24*, 125–136. https://doi.org/10.1016/j.addma.2018.09.014

Ekinci, İ., Karaca, M. M., & Ali, D. (2025). Mechanical performance of Schwarz and lattice diamond regular architecture versus stochastic

Voronoi structure for bone scaffolds: Design, Manufacturing, and FEM analysis. *The International Journal of Advanced Manufacturing Technology*. https://doi.org/10.1007/s00170-025-17190-3

Engineers, A. S. of M. (2006). *Guide for verification and validation in computational solid mechanics*.

Ereiz, S., Duvnjak, I., & Fernando Jiménez-Alonso, J. (2022). Review of finite element model updating methods for structural applications. *Structures*, *41*, 684–723. https://doi.org/10.1016/j.istruc.2022.05.041

Evans, L. M., Sözümert, E., Keenan, B. E., Wood, C. E., & du Plessis, A. (2023). A Review of Image-Based Simulation Applications in High-Value Manufacturing. *Archives of Computational Methods in Engineering*, *30*(3), 1495–1552. https://doi.org/10.1007/s11831-022-09836-2

Fish, J., & Belytschko, T. (2007). *A first course in finite elements* (Vol. 1). Wiley New York.

Fung, Y. C. (2013). *Biomechanics: Mechanical Properties of Living Tissues*. Springer Science & Business Media.

Geers, M. G. D., Kouznetsova, V. G., & Brekelmans, W. A. M. (2010). Computational homogenization. In R. Pippan & P. Gumbsch (Eds), *Multiscale Modelling of Plasticity and Fracture by Means of Dislocation Mechanics* (Vol. 522, pp. 327–394). Springer Vienna. https://doi.org/10.1007/978-3-7091-0283-1_7

Gibson, I., Rosen, D., Stucker, B., & Khorasani, M. (2021). Development of Additive Manufacturing Technology. In I. Gibson, D. Rosen, B. Stucker, & M. Khorasani (Eds), *Additive Manufacturing Technologies* (pp. 23–51). Springer International Publishing. https://doi.org/10.1007/978-3-030-56127-7_2

Gibson, L. J. (2003). Cellular Solids. *MRS Bulletin*, *28*(4), 270–274. https://doi.org/10.1557/mrs2003.79

Higham, N. J. (2002). *Accuracy and Stability of Numerical Algorithms* (Second). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718027

Hill, R. (1963). Elastic properties of reinforced solids: Some theoretical principles. *Journal of the Mechanics and Physics of Solids*, *11*(5), 357–372. https://doi.org/10.1016/0022-5096(63)90036-X

Hollister, S. J. (2005a). Porous scaffold design for tissue engineering. *Nature Materials*, *4*(7), 518–524. https://doi.org/10.1038/nmat1421

Hollister, S. J. (2005b). Porous scaffold design for tissue engineering. *Nature Materials*, *4*(7), 518–524.

Holzapfel, A. G. (2000). *Nonlinear solid mechanics II*. http://elibrary.matf.bg.ac.rs/bitstream/handle/123456789/2762/NonlinearSolidMechHOLZAPFELP2.pdf?sequence=1

Hughes, T. J. (2003). *The finite element method: Linear static and dynamic finite element analysis*. Courier Corporation.

Jerier, J.-F., Richefeu, V., Imbault, D., & Donzé, F.-V. (2010). Packing spherical discrete elements for large scale simulations. *Computer Methods in Applied Mechanics and Engineering*, *199*(25), 1668–1676. https://doi.org/10.1016/j.cma.2010.01.016

Kavakli, H. S., & Ali, D. (2023). Enhancing the Mechanical Properties of Auxetic Metamaterials by Incorporating Nonrectangular Cross Sections into Their Component Rods: A Finite Element Analysis. *Physica Status Solidi (b)*, *260*(3), 2200194. https://doi.org/10.1002/pssb.202200194

Kavakli, H. S., & Ali, D. (2025). Adapting Metamaterial Structures for a New Generation of Bone Substitute Scaffolds with the Desired Stiffness and Permeability: A Finite Element Analysis. *Journal of Materials Engineering and Performance*, *34*(18), 20586–20597. https://doi.org/10.1007/s11665-025-10686-9

Keyak, J. H., Rossi, S. A., Jones, K. A., & Skinner, H. B. (1997). Prediction of femoral fracture load using automated finite element modeling. *Journal of Biomechanics*, *31*(2), 125–133.

Lemaitre, J., & Chaboche, J.-L. (1994). *Mechanics of Solid Materials*. Cambridge University Press.

Luecke, W. E., Ma, L., Graham, S. M., & Adler, M. A. (2010). *Repeatability and Reproducibility of Compression Strength Measurements Conducted According to ASTM E9*. https://ntrs.nasa.gov/citations/20110007223

Malvern, L. E. (1969). *INTRODUCTION TO THE MECHANICS OF A CONTINUOUS MEDIUM* (No. Monograph). https://trid.trb.org/View/199874

Mondal, P., Das, A., Wazeer, A., & Karmakar, A. (2022). Biomedical porous scaffold fabrication using additive manufacturing technique: Porosity, surface roughness and process parameters optimization. *International Journal of Lightweight Materials and Manufacture*, *5*(3), 384–396. https://doi.org/10.1016/j.ijlmm.2022.04.005

Oberkampf, W. L., & Roy, C. J. (2010). *Verification and Validation in Scientific Computing*. Cambridge University Press.

Oberkampf, W. L., & Trucano, T. G. (2002a). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*, *38*(3), 209–272. https://doi.org/10.1016/S0376-0421(02)00005-2

Oberkampf, W. L., & Trucano, T. G. (2002b). Verification and validation in computational fluid dynamics. *Progress in Aerospace Sciences*, *38*(3), 209–272. https://doi.org/10.1016/S0376-0421(02)00005-2

Ogden, R. W. (1997). *Non-linear Elastic Deformations*. Courier Corporation.

Rasheed, S., Lughmani, W. A., Khan, M. M., Brabazon, D., Obeidi, M. A., & Ahad, I. U. (2023). The Porosity Design and Deformation Behavior Analysis of Additively Manufactured Bone Scaffolds through Finite Element Modelling and Mechanical Property Investigations. *Journal of Functional Biomaterials*, *14*(10), 496. https://doi.org/10.3390/jfb14100496

Roache, P. J. (1994). *Perspective: A method for uniform reporting of grid refinement studies*. https://asmedigitalcollection.asme.org/fluidsengineering/article-abstract/116/3/405/411554

Roache, P. J. (1998). *Verification and validation in computational science and engineering* (Vol. 895). Hermosa Albuquerque, NM. https://pdfs.semanticscholar.org/0f3c/728bd0f17e45cce72bda2165707a0eb9e03b.pdf/1000

Sargent, R. G. (2010). Verification and validation of simulation models. *Proceedings of the 2010 Winter Simulation Conference*, 166–183.

Shetty, A., Fathima, A., Anika, B., Shetty, R., Vinyas, Supriya, J. P., & Hegde, A. (2025). Computational optimization of 3D printed bone scaffolds using orthogonal array-driven FEA and neural network modeling. *Scientific Reports*, *15*(1), 30515. https://doi.org/10.1038/s41598-025-15122-5

Smith, R. C. (2024). *Uncertainty Quantification: Theory, Implementation, and Applications, Second Edition*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611977844

Sutton, M. A., Orteu, J. J., & Schreier, H. (2009a). *Image correlation for shape, motion and deformation measurements: Basic concepts, theory and applications*. Springer Science & Business Media.

Sutton, M. A., Orteu, J. J., & Schreier, H. (2009b). *Image Correlation for Shape, Motion and Deformation Measurements: Basic Concepts,Theory and Applications*. Springer Science & Business Media.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898717921

Trefethen, L. N., & Bau, D. (2022). *Numerical Linear Algebra, Twenty-fifth Anniversary Edition*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611977165

Van Loan, C. F., & Golub, G. (1996). Matrix computations (Johns Hopkins studies in mathematical sciences). *Matrix Computations*, *5*, 32.

Viceconti, M., Olsen, S., Nolte, L.-P., & Burton, K. (2005). Extracting clinically relevant data from finite element simulations. *Clinical Biomechanics*, *20*(5), 451–454. https://doi.org/10.1016/j.clinbiomech.2005.01.010

Wriggers, P. (2012). Computational contact mechanics. *Computational Mechanics*, *49*(6). https://go.gale.com/ps/i.do?id=GALE%7CA356580998&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=01787675&p=AONE&sw=w

Zhang, J., Chen, B., Chen, B., Wang, H., Han, Q., Tang, X., & Qin, Y. (2025). Clinical Application of Finite Element Analysis in Meniscus Diseases: A Comprehensive Review. *Archives of Computational Methods in Engineering*, *32*(7), 4163–4195. https://doi.org/10.1007/s11831-025-10265-0

Zienkiewicz, O. C., & Taylor, R. L. (2005). *The Finite Element Method for Solid and Structural Mechanics*. Elsevier.

Chapter 7

Legal Transformation and Changes in Implementation Policies in Building Inspection Legislation

Sıtkı Alper ÖZDEMİR[1], Mehmet UZUN[2]

1. Introduction

Turkey is located on active fault lines in terms of seismicity and has experienced catastrophic earthquakes and disasters throughout history (Dedeoglu et al., 2025). The Marmara Earthquake, which occurred on the North Anatolian Fault (NAF) on August 17, 1999, caused large-scale loss of life and property. A similar and more recent example is the 7.6 and 7.7 magnitude earthquakes centered in Kahramanmaraş on the East Anatolian Fault (EAF) on February 6, 2023, which killed thousands of people and destroyed hundreds of thousands of buildings (Aydan et al., 2024; İnce, 2024; Kazaz, 2025). This situation demonstrates that building safety is a serious issue that must be addressed at the national level (Gurbuz & Cengiz, 2025). In this context, Law No. 4708 on Building Inspection defines the inspection mechanism in the design and construction processes, enabling technical and independent inspection at all stages of the inspection (Engin & Özbil, 2022). In terms of these characteristics, Law No. 4708 is a fundamental regulation aimed at protecting life and property. However, since the law came into force, many problems have been encountered in its implementation (Engin & Özbil, 2022). These problems have revealed inconsistencies between the legislation and its technical application. This situation has necessitated regular revisions to Law No. 4708, which has been in force since 2001, due to the dynamic nature of the sector. The revision of legislation is inevitable, considering both the size of the construction sector and the emergence of new construction technologies and safety standards. For example, according to Law No. 4708, construction supervision services are carried out by supervisory firms and contractors through contracts between them, while structural reforms such as the

---

[1] Iskenderun Technical University/Iskenderun Vocational School, Department of Construction
salper.ozdemir@iste.edu.tr- 0000-0001-7331-8026
[2] Karamanoğlu Mehmetbey University/Faculty of Engineering, Department of Civil Engineering,
mehmetuzun@kmu.edu.tr – 0000-0002-6347-1243

implementation of the Electronic Distribution System (EDS) have been implemented with changes in legislation. With EDS, the link between the construction supervision firm (YDF) and the contractor was severed, aiming to increase the independence of supervision through a centralized and sequential appointment procedure (Engin & Özbil, 2022). Similarly, regulations introduced after 2019 expanded the legal responsibilities of construction supervision firms, enabling them to become more effective in supervision. The lack of alignment between the institutional capacity in the field and the risk level of the YDF has created a significant mismatch between the system's implementation and its regulations. This study examines the revisions made since the YDHK No. 4708 came into force. The effects of the relevant revisions on the building inspection system have been examined in terms of the functionality and effectiveness of the inspection mechanism, and suggestions for a model that can be developed to make the building inspection system more functional have been presented.

2. Methods and Materials

In this study, a multifaceted methodological framework was established by combining qualitative content analysis, legislation review, comparative legal assessment, and literature synthesis methods. First, the laws, regulations, and circulars in force between 2001 and 2025 were classified within a historical flow. The legal justifications, objectives, and practical effects of the revisions were evaluated using documentary analysis. Subsequently, academic studies on the building control system and field data from the construction sector were reviewed to identify prominent problem areas in the literature. This approach enabled a comprehensive comparison of the differences between the normative aspects of the legislation and its actual implementation in the field.

In the impact analysis, the technical and administrative consequences of the revisions on the building inspection system were examined. Taking into account risk factors and the regulatory context, the effects of the revisions on the functioning of the building inspection system and their practical consequences were evaluated. In line with the purpose of the study, an analytical framework was established based on probability-impact assessment and theoretical frameworks.

3. Literature Review

A review of the literature reveals that studies have been conducted on legal liability, institutional capacity, authority limits, and technical competence in the building inspection system.

(Çelik & Ünal, 2017), To identify the structural problems encountered in the implementation of the building inspection system in Turkey, a field survey based on a questionnaire was conducted among 73 technical personnel working for building inspection companies operating in the province of Adana. The study shows that the problems experienced in the context of the institutional structure, operation, and stakeholder relations of building inspection companies largely stem from the mismatch between the level of responsibility stipulated by the legislation and the actual capacity in the field. The findings reveal that issues such as low staff salaries, insufficient experience and expertise among technical personnel, excessive inspection workloads, and the involvement of many inspectors and control personnel in the process solely through their signing authority, without actually being present in the field, directly negatively affect the quality of inspections. Furthermore, the failure of building owners to pay building inspection service fees on time, the commercial pressure exerted by contractors on inspection companies, and bureaucratic delays caused by local authorities have been identified as key factors weakening the system's functioning. A significant portion of the participants acknowledge that Law No. 4708 has generally had a positive impact on the building inspection system. However, they have stated that the current legislation is insufficient and that there are shortcomings, particularly in terms of increasing the powers of inspection organizations, strengthening the central pool system, and guaranteeing the personal rights of engineers. In this respect, the study reveals that revisions to the building inspection legislation will not achieve the expected effectiveness unless they are supported by institutional capacity and practical implementation in the field.

(Boysal & Güvel, 2023), A comprehensive, multi-stakeholder survey was conducted specifically in Osmaniye province to evaluate the implementation of Law No. 4708 on Building Inspection and the relevant implementation regulations. The study stands out in the literature for its comprehensive approach to examining all components of the building inspection system, covering not only technical personnel working in building inspection organizations but also technical personnel working in licensing authorities (municipalities), building contractors, project designers, and site managers. The findings of the study reveal that, despite the theoretical aim of building inspection legislation to improve building quality, it harbors serious structural problems in practice. In particular, the fact that the majority of technical personnel working in building inspection organizations are paid at the minimum wage level, technical training is not provided systematically, inspection service fees are not paid on time, and there is an imbalance between inspection responsibilities and administrative powers have been identified as key factors weakening the effectiveness of the system. Furthermore, the survey results show that while the transition to the post-2011 building inspection system has led to a certain improvement in building quality, the building inspection mechanism alone is not considered sufficient to combat illegal construction, and amnesty applications are viewed negatively in terms of building safety. Insufficient oversight in the licensing and building occupancy permit processes, municipalities' failure to effectively use notification and sanction mechanisms, and the de facto reduction of site management to merely a signature process are among the structural problems limiting the functionality of the inspection system in the field. In this respect, the study empirically demonstrates that there are serious gaps in the implementation of revisions to building inspection legislation and serves as a strong field-based resource emphasizing the need to re-examine the legislation with an approach that not only increases criminal liability but also strengthens institutional capacity and areas of authority.

(Yılmaz et al., 2017), this comprehensive study examines the impact of building inspection officials on occupational health and safety (OHS) inspections, particularly in public construction projects, by addressing OHS legislation and building inspection legislation with an integrated approach. The study evaluates Law No. 4708 on Building Inspection, Laws No. 4734 and 4735 on Public Procurement, and Laws No. 4857 and 6331 on Occupational Health and Safety, which came into force between 2001 and 2012, partly due to the European Union harmonization process. analyzing why these regulations have not sufficiently demonstrated the expected impact in reducing work accidents in the construction sector. Social Security Institution (SGK) data cited in the article reveals that, despite numerous regulatory changes, the construction sector still accounts for approximately one-third of all fatal workplace accidents. The study emphasizes that the fundamental reason for this situation is the fragmented structure of the inspection mechanisms and the lack of clear authority and responsibility boundaries. The authors state that the inspection, warning, and notification duties related to OSH assigned to building inspection organizations under Article 2(f) of Law No. 4708 remained ineffective for a long time and only became more visible after the explanatory amendment made by Law No. 6645 in 2015. The results of a survey conducted with 36 control engineers working in public institutions within the scope of the study show that the vast majority of participants agree that granting construction inspection personnel OSH inspection authority on construction sites would reduce work accidents. However, concerns that legal liability would be attributed to them in the event of possible accidents cause technical personnel to shy away from taking responsibility in this area. The authors argue that building inspection and OSH inspection should be addressed together throughout all stages of the project life cycle, and that the effective implementation of risk analyses and health and safety plans in the design and project planning phase plays a critical role in preventing accidents. In this respect, the study provides a strong contribution to the literature, emphasizing that revisions to building inspection regulations must be addressed within a comprehensive system that establishes a balance of authority and responsibility, covering not only structural quality but also occupational health and safety, and examining the relationship between regulations and implementation with both normative and empirical data.

(Nar, 2023)'s study, he examines in detail the legal liability of building inspection organizations towards the building owner arising from the contract due to damages caused by earthquakes, within the framework of Turkish Law of Obligations and building inspection legislation. The article discusses the nature of the legal relationship established between building inspection organizations and building owners under Law No. 4708 on Building Inspection, revealing that this relationship is a private law contract and that the rights and obligations of the parties are essentially subject to the provisions of the contract and the provisions of the Code of Obligations. The author emphasizes that the liability of building inspection organizations for building damage resulting from extraordinary events such as earthquakes is not automatic and unlimited, and that liability can only arise if the elements of fault, causation, and breach of contract are present simultaneously. The study details the legal liability of building inspection organizations, particularly through the concepts of defective performance, duty of care, and liability for the acts of auxiliary persons. In light of post-earthquake court decisions, it is stated that the decisive criterion is whether the inspection agencies acted in accordance with the project, the applicable legislation, and the rules of science and art. Furthermore, the study discusses whether the liability of building inspection organizations can be assessed within the scope of joint and several liability together with contractors and project designers, and states that the degree of fault must be determined separately in each specific case. The study reveals that the increasingly blurred boundaries between the private law liability of building inspection organizations and public law sanctions, along with the extension of liability periods and the intensification of criminal sanctions in building inspection legislation in recent years, create serious legal risks in practice. In this respect, the article provides an important contribution to the literature by explaining how revisions to building inspection legislation have transformed the post-earthquake legal liability regime for YDFs and by providing a strong legal basis for the discussion of increased liability versus insufficient authority addressed in the present study.

(Özdemir, 2020), the article examines in detail whether administrative fines can be imposed on building inspection organizations under the provisions of the Urban Planning Law No. 3194, within the framework of building inspection regulations and administrative law principles. The article evaluates the legal status of building inspection organizations based on the duties and responsibilities assigned to them by Law No. 4708 on Building Inspection. Although YDFs are not actors directly implementing zoning, it is revealed that they are sometimes subject to fines based on the provisions of the Zoning Law. The author argues that Law No. 4708 provides for a special liability and sanction regime for construction supervision organizations and, therefore, the direct and analogous application of administrative fines regulated in the Zoning Law No. 3194 to YDFs lacks legal basis. The study concludes that, based on an assessment conducted within the framework of the principles of legality, parallelism in authority and procedure, and the relationship between special and general laws, construction supervision organizations can only be subject to sanctions explicitly stipulated in Law No. 4708 and the regulations issued based on this law. Furthermore, the article clearly highlights the differences between the technical liability of building inspection organizations and their administrative liability; it emphasizes that the fault-based legal liability of building inspection organizations and the limits of the administration's power to impose penalties are often confused in practice. In this context, the study reveals that the increasing number of criminal and administrative sanctions in building inspection legislation in recent years, applied without clearly defined normative boundaries, creates legal uncertainty and exposes YDFs to unpredictable risks in terms of the balance of authority and responsibility. In this respect, the article makes an important contribution to the literature, strongly supporting the view that the need for revision in building inspection legislation should be addressed not only from a technical perspective but also in terms of clarifying administrative law-based norms and authority limits.

(Dikmen et al., 2011), by comprehensively addressing Law No. 4708 on Building Inspection and Labor Law No. 4857 from an occupational health and safety (OHS) perspective, it reveals that high accident rates in the construction sector stem not only from implementation deficiencies but also from the fragmented nature of the legislation. Using data from the International Labor Organization (ILO) and the Social Security Institution (SGK), the study shows that the construction sector carries a disproportionately high risk of fatal and permanent disability accidents compared to all other sectors. Statistics for the period 1992–2009, in particular, reveal that this risk persists structurally despite regulatory changes. The authors note that the OSH provisions of Labor Law No. 4857 focus primarily on the implementation phase, whereas Construction Supervision Law No. 4708 has a more comprehensive supervision potential covering the project, design, and implementation processes. However, they argue that this potential is not sufficiently utilized in practice. In particular, a survey conducted with 25 construction supervision organizations operating in Istanbul, Ankara, Izmir, Sakarya, and Eskisehir reveals that the authority granted to construction supervision organizations under Article 2(f) of Law No. 4708 to warn and notify contractors regarding occupational health and safety is either unknown in the sector or is not actually used due to concerns about losing new business. The findings show that the vast majority of construction supervision firms do not have sufficient knowledge of OHS legislation, that almost none of them employ certified OHS experts, and that they refrain from exercising the powers granted to them by the legislation. The study emphasizes that building inspection and OSH inspection have complementary functions throughout all stages of the project life cycle, and that these two areas should be treated as an integrated inspection system rather than separately. In this respect, the article presents a strong contribution to the literature, based on both theoretical and field data, demonstrating that revisions to building inspection regulations should be designed around a comprehensive institutional model with a clearly defined balance of authority and responsibility, covering not only structural safety but also occupational health and safety.

In summary, as seen in the literature review, the fundamental problem in the control system appears to be concentrated in the areas of high accountability, low authority, and weak institutional capacity. The incompatibility between these three elements threatens the long-term sustainability of the system. Table 1 summarizes the studies in the literature.

Table 1. Literature review of building inspection systems

Author Name	Subject	Purpose	Importance	Contribution
(Tantekin Çelik & Ünal, 2017)	Field applications of the building inspection system, corporate and financial issues	Identify the structural and operational problems faced by building inspection organizations	Demonstrating empirically the mismatch between building inspection regulations and actual capacity in the field	By revealing the gap between legislation and practice using field data, it highlights the lack of institutional capacity in the literature.
(Boysal & Güvel, 2023)	Multi-stakeholder assessment of the building inspection system	Analyzing the impact of building inspection regulations on different actors	Examine the building inspection system not only from the YDF perspective, but from the perspective of all stakeholders.	It reveals that building inspection alone does not guarantee building quality and highlights the lack of institutional integration.
(Yılmaz et al., 2017)	The relationship between occupational safety and health and building inspection, public construction projects	Examining the role of building inspection officials in occupational health and safety inspections	Addressing the integration of building inspection and occupational safety and health in the context of public construction projects	It presents a comprehensive approach that highlights the need to strengthen the occupational safety and health dimension of the building inspection system.
(Nar, 2023)	The civil liability of the YDF after an earthquake	Explaining the legal liability of building inspection organizations towards the building owner	Explaining on a legal basis that YDF liability is not automatic and unlimited after an earthquake	Clarifies that the YDF must be assessed in terms of fault, causation, and breach of contract
(Özdemir, 2020)	Administrative sanctions and technical liability	Examine whether penalties can be imposed on YDFs under the Zoning Law	Discussion of legality and limits of authority in administrative fines	It reveals that the confusion surrounding sanctions in building inspection legislation creates legal uncertainty.
(Dikmen et al., 2011)	Integration of occupational health and safety legislation with building inspection legislation	Yapı denetiminin İSG üzerindeki potansiyel etkisini değerlendirmek	Proje yaşam döngüsünde entegre denetim yaklaşımını savunması	Yapı denetiminin yalnızca yapısal değil, iş güvenliği açısından da kritik bir araç olduğunu ortaya koyar

4. Historical Development and Transformation Dynamics of Legislation

The implementation of YDHK No. 4708 throughout Turkey has established a national standard for ensuring structural safety (Ömürbek et al., 2016). An examination of the revisions made to the law since 2001 reveals that the aim is to strengthen the independence of inspections, the transparency of processes, and institutional accountability mechanisms, along with the technical suitability of buildings.

The 2012 regulatory revision expanded the legal liability areas of YDF. The increased level of responsibility has led to a significant increase in the risk level for YDFs. In particular, the enactment of the Occupational Health and Safety (OHS) Law No. 6331 has paved the way for YDFs to be held liable, albeit indirectly, for work accidents that may occur on construction sites.

The introduction of EDS in 2018 was one of the most critical turning points for the building inspection sector. With this application, contractors' right to choose YDFs was removed. This reduced the commercial concerns and difficulties in obtaining work faced by YDFs in the inspection processes. Furthermore, the payment processes for inspection fees have become more secure and disciplined. Thanks to these regulatory revisions, the Building Inspection System has been significantly strengthened, and the principles of independent inspection have become more effective.

Regulatory revisions made in 2019 have increased penalties such as extending the liability period for construction to up to 15 years, revoking guarantees, and suspending operations. This has increased the level of legal liability for YDFs in the building inspection system. However, the corporate volume of YDFs has not increased proportionally, which has also led to risks in the personnel-work balance in the building inspection system.

The most striking change in the 2025 revisions is the removal of the EDS requirement. With this change, contractors are now entitled to enter into contracts with the YDF of their choice for projects with a construction area of 500 m^2 or less (small-scale projects) (Law on Building Inspection, n.d.). This regulation creates commercial pressure, especially for YDFs operating in sparsely populated areas. The fear of not being able to obtain work due to the supervision-payment relationship has created a new risk situation for YDFs. This situation creates an open situation that leads to the legislation imposing increased legal responsibility on YDFs in terms of the historical development of the building inspection system, while at the same time demonstrating a weak institutional stance in the field of application. Table 2 shows the revisions made in the historical process of the legislation governing the building inspection system.

Table 2. Revisions to Regulations Related to the Building Inspection System

Year	Relevant Legislation	Reason/Purpose of Revision	Institution/Stakeholder	Details and Scope
2004	Law No. 5108 and its Implementing Regulation	Extending the law from pilot provinces to the entire country and establishing the legal infrastructure.	Ministry of Environment and Urbanization, YDFs	The law will be gradually made mandatory throughout Turkey.
2011	Decree Law No. 648 and Related Regulations	Prevent contractor pressure (commercial risk) and ensure the financial security of the YDF.	YDFs, Contractors, Building Owners	The requirement to transition to the Electronic Distribution System (EDS) and deposit the audit fees into the Escrow Account.
2012	Occupational Health and Safety (OHS) Law No. 6331	Reducing workplace accidents in the construction sector and expanding the scope of YDF's responsibilities to include occupational health and safety.	YDFs, Ministry of Labor and Social Security	Increased indirect liability of the YDF arising from OSH legislation.
2016-2017	Law and Regulation No. 6745	Improve audit quality by preventing YDFs from taking on excessive workloads.	Ministry of Environment and Urbanization, YDFs	Imposing a limit on the maximum construction area (m²) that YDF can oversee.
2019	Regulation Amendments	Increasing the deterrent effect of criminal penalties and extending the period of legal liability.	Çevre ve Şehircilik Bakanlığı, Yargı, YDF'lar	Seizure of collateral, strengthening of penalties for suspension from activity, and extension of the liability period up to 15 years.
2021-2022	Regulation Amendments	Ensuring the quality of supervisory personnel.	Ministry of Environment and Urbanization, YDFs, Supervisory Staff	Imposing restrictions such as an age limit (e.g., 75 years old) for auditors to take on duties.
2023	Laws and Regulations Issued Following the February 6 Earthquakes	Accelerate post-earthquake emergency action plans and urban transformation processes.	Ministry of Environment, Urbanization, and Climate Change, YDKs	Redefining the role of YDF in urban transformation and expanding its scope of supervision.
2025	Regulatory Changes (Under the Regulations dated June 12, 2025)	Adapting flexibility and control costs to current costs in small-scale structures.	YDFs, Contractors, Ministry of Environment, Urbanization, and Climate Change	For buildings under 500 m², the contractor can choose the YDF themselves (exemption from EDS) and the Building Inspection Service Fee Unit Prices are updated.

5. The Impact of Revisions on the Building Inspection System

Revisions to legislation have had critical effects on the building inspection system. These effects have been analyzed from the perspectives of institutional design theory, risk management models, and administrative sanctions law. In this context, the structural problems of the building inspection system are evaluated along three main axes.

First, the transformation of the principle of independence must be addressed. The introduction of the EDS in 2011 severed the contractor–YDF relationship and made the audit process more transparent and feasible. This practice freed YDFs from commercial pressures. However, the revision made in 2025 has made it possible for contractors to choose the YDF again in small-scale projects (less than 500 m2). This situation has reintroduced the possibility of undermining the principle of independence. From a systemic perspective, asymmetric power relations may re-emerge, which could reduce the quality of supervision.

Secondly, the revisions have expanded the areas of legal liability, but the financial and institutional capacity of the YDF has not been strengthened to the same extent. Extending the liability period to 15 years, increasing penalties for guarantee cancellation, and indirect obligations under OSH have significantly increased the risk assumed by the YDF. This indicates that the system's sustainability has been weakened by a risk-capacity mismatch.

Thirdly, the mismatch between authority and responsibility must be addressed. YDFs do not have the direct authority to halt operations as a result of practices that violate the licensed implementation projects. This situation causes significant time losses in administrative processes and reduces the effectiveness of oversight in the field. YDFs that lack the authority commensurate with their level of responsibility suffer from process weaknesses during the oversight process.

6. Conclusions and Recommendations

When examining revisions to building inspection regulations throughout history, it is evident that the legal and penal dimensions of the system have progressively strengthened. However, its institutional and economic capacity has not developed at the same rate. This situation hinders the achievement of the desired outcomes in terms of building safety. In order for the building inspection system to become more effective and ergonomic, it is necessary to strengthen its financial structure, preserve the principle of independence, ensure the alignment of authority and responsibility, and improve the quality of the inspection personnel.

The building inspection system, which came into effect in 2001, should be analyzed in the context of revisions, applications, and shortcomings made throughout its historical process. Short-, medium-, and long-term goals should be set based on regulatory and systemic experiences. In the short term, inspection fees should be updated, and a financial model should be developed to ensure the financial security of YDFs. In the medium term, YDFs should be granted the authority to directly halt work in cases of non-compliant practices. This will ensure effective supervision and streamlined administrative processes. In the long term, the implementation of artificial intelligence-based supervision systems will eliminate human error. Furthermore, the points where OSH legislation intersects with the building inspection system should be more clearly defined and explicitly stated in the relevant legislation. Implementing all these recommendations will strengthen both the legal and institutional aspects of the building inspection system and create a more sustainable inspection mechanism in the long term.

References

Aydan, Malistani, N., Ulusay, R., & Kumsar, H. (2024). The damage and response of some underground structures by the 2023 February 6 Great Turkish Earthquakes and some considerations. *Engineering Geology*, *336*. https://doi.org/10.1016/j.enggeo.2024.107549

Boysal, V. B., & Güvel, T. (2023). Yapı Denetim Mevzuatında Tespit Edilen Problemlerin Osmaniye İli Özelinde Değerlendirilmesi. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *6*(3), 2296–2315. https://doi.org/10.47495/OKUFBED.1170963

Çelik, G. T., & Ünal, C. (2017). Yapı Denetim Firmalarının Sorunlarının Belirlenmesi ve Adana Örneği. *Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*, *32*(4), 71–78. https://doi.org/10.21605/CUKUROVAUMMFD.371039

Dedeoglu, I. O., Yetkin, M., Tunc, G., & Ozbulut, O. E. (2025). Evaluating earthquake-induced damage in Dogansehir, Malatya after 2023 Kahramanmaras Earthquake sequence: Geotechnical and structural perspectives. *Journal of Building Engineering*, *104*. https://doi.org/10.1016/j.jobe.2025.112266

Dikmen, S. Ü., Akbıyıklı, R., Aytekin, O., & Baradan, S. (2011). İş ve Yapı Denetim Yasalarının İş Sağlığı ve Güvenliği Açısından Bütünleşik İncelenmesi. *Eskişehir Osmangazi Üniversitesi Mühendislik ve Mimarlık Fakültesi Dergisi*, *24*(2), 117–132. https://dergipark.org.tr/tr/pub/ogummf/article/324793?issue_id=30114

Engin, V., & Özbil, B. (2022). Türkiye'de Uygulanan Yapı Denetim Sisteminin Sorunları İle İlgili Yaklaşımlar. *Kent Akademisi*, *15*(4), 1575–1590. https://doi.org/10.35674/KENT.1115470

Gurbuz, T., & Cengiz, A. (2025). Structural damages during the February 06, 2023 Kahramanmaraş Earthquakes in Turkey. *Soil Dynamics and Earthquake Engineering*, *191*, 109214. https://doi.org/10.1016/J.SOILDYN.2025.109214

İnce, O. (2024). Structural damage assessment of reinforced concrete buildings in Adıyaman after Kahramanmaraş (Türkiye) Earthquakes on 6 February 2023. *Engineering Failure Analysis*, *156*. https://doi.org/10.1016/j.engfailanal.2023.107799

Kazaz, İ. (2025). Evaluation of seismic performance, design and practice of frame-wall buildings in Türkiye after the 6 February 2023 Kahramanmaraş earthquakes. *Journal of Building Engineering*, *99*, 111557. https://doi.org/10.1016/J.JOBE.2024.111557

Nar, A. (2023). Yapı Denetim Kuruluşlarının Depremde Meydana Gelen Zararlardan Dolayı Yapı Sahibine Karşı Sözleşmeden Doğan Sorumluluğu. *Erzincan Binali Yıldırım Üniversitesi Hukuk Fakültesi Dergisi*, *27*(1), 77–115. https://dergipark.org.tr/tr/pub/ebyuhfd/article/1311221?issue_id=78249

Ömürbek, N., Karaatlı, M., & Cömert, H. G. (2016). AHP-SAW ve AHP-ELECTRE Yöntemleri ile Yapı Denetim Firmalarının Değerlendirmesi. *Yönetim Bilimleri Dergisi*, *14*(27), 171–199. https://dergipark.org.tr/tr/pub/comuybd/article/533904?issue_id=43601

Özdemir, S. (2020). Yapı Denetim Kuruluşuna İmar Kanunu Kapsamında Para Cezası Verilip Verilemeyeceği Hakkında Bir İnceleme: Yapı Denetim Kuruluşlarının Fennî Mesuliyet Sorumlulukları. *Necmettin Erbakan Üniversitesi Hukuk Fakültesi Dergisi*, *3*(2), 182–207. https://doi.org/10.51120/NEÜHFD.2020.3

Yapı Denetimi Hakkında Kanun, Pub. L. No. 4708. Retrieved December 16, 2025, from https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=4708&MevzuatTur=1&MevzuatTertip=5

Yılmaz, M., Yıldız, S., Bakış, A., & Kanıt, R. (2017). Bir Bütün Olarak İş Sağlığı Güvenliği ve Yapı Denetim Mevzuatı: Yapı Denetim Görevlilerinin Kamu İnşaatlarında İş Sağlığı Güvenliği Denetimine Etkileri. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, *8*(3), 433–442. https://dergipark.org.tr/tr/pub/dumf/article/405195?issue_id=33629

Chapter 8

Large Language Models in Economics: A Chronological Review and Conceptual Assessment

Aytürk KELEŞ[1], **Ali KELEŞ**[2]

1. Introduction

Over the past decade, advancements in natural language processing (NLP) have significantly transformed how economists engage with unstructured text data. Early studies in this field largely treated text as a supplementary data source, converting documents such as news articles, policy statements, and corporate reports into numerical features suitable for integration into standard econometric models.

The introduction of transformer-based architectures markedly enhanced the capacity of machines to understand and process language. These models enabled effective analysis of complex textual structures and long-range contextual dependencies, facilitating the extraction of richer and more meaningful insights from text. Building upon this foundation, large language models (LLMs) have evolved to perform tasks such as generating human-like text, summarizing documents, and classifying content with greater accuracy. These developments have enabled text-based analysis to become more comprehensive and impactful, extending beyond the limitations of traditional econometric approaches and broadening the role of text data in economics.

This study begins with a chronological review of key contributions in the literature between 2015 and 2025. It then examines and organizes the LLM methods used across these studies. In the final section, the paper offers a perspective on the potential of LLM-based approaches in economic research.

---

[1] Prof. Dr., Ağrı İbrahim Çeçen University, Faculty of Education, Department of Computer and Education Technologies Education, ORCID: 0000-0001-9755-295X

[2] Prof. Dr., Ağrı İbrahim Çeçen University, Faculty of Education, Department of Computer and Education Technologies Education, ORCID: 0000-0002-0785-9593

2. Applications of Large Language Models in Economics

The study *"Text as Data"* by Gentzkow, Kelly, and Taddy (2017) provided a foundational and systematic framework for how unstructured text can be used in quantitative economic analysis. Written from an economist's perspective, the paper outlines practical steps for integrating textual data into processes of prediction and causal inference. Methodologically, it begins by defining the target economic concept, followed by labeling and validation procedures applied to the text, and finally transforms the text into numerical signals (e.g., word frequencies, sentiment scores, topic distributions) for analysis.

This framework continues to serve as a core structure for many applications developed during the era of large language models (LLMs). Today, most LLM-based economic pipelines still follow the system proposed by Gentzkow and colleagues: starting with an abstract concept, building supervised data structures, and converting textual content into meaningful, measurable economic indicators.

In later LLM research, this approach has been widely applied in areas such as measuring central bank policy communication, nowcasting using news data, analyzing firm disclosures, and extracting sentiment and topics from text.

In their 2019 article published in the *Journal of Economic Literature*, Gentzkow, Kelly, and Taddy present *"Text as Data"* as a foundational reference that systematically consolidates the main methods and applications of text analysis in economics. This comprehensive work organizes a broad range of natural language processing (NLP) techniques, including supervised and unsupervised learning, dictionary-based approaches, and word embeddings, and illustrates how these tools have been applied across various domains of economic research.

One of the key contributions of this paper to current LLM-based research is its clear articulation of the risks related to identification and measurement, the necessity of validation procedures, and the important distinction between prediction and concept measurement. This distinction is especially relevant in the era of LLMs, which, while capable of generating highly accurate predictions, may not always align precisely with the abstract economic constructs researchers aim to measure.

The most important practical takeaway from the paper is that, even when LLMs offer superior textual representations, researchers must still perform ground truth checks, conduct robustness tests, and carefully assess the construct validity of their models. This ensures that text data is not just used as a raw input to models, but as a meaningful and valid representation of economic concepts.

FinBERT, developed by Araci (2019), is a BERT-based language model specifically adapted to the financial domain, offering a major contribution to improving financial sentiment analysis. The model applies domain adaptation techniques to tailor the original BERT architecture to the unique vocabulary and linguistic patterns of financial texts. As a result, FinBERT is able to perform significantly more accurate and consistent sentiment and information extraction from sources such as news articles, earnings calls, company reports, and other financial documents.

This development holds considerable importance for economics and finance, as sentiment signals derived from such texts are widely used in asset pricing, market reaction analysis, and macro-finance research. By enhancing the quality of textual data processing in these areas, FinBERT has contributed to making the integration of unstructured text into economic models more reliable and methodologically sound.

Methodologically, FinBERT was built by first conducting continued pre-training of the general BERT model on large financial corpora to capture domain-specific language. This was followed by task-specific fine-tuning, particularly on financial sentiment classification tasks using datasets like the Financial PhraseBank and FiQA. As a result, FinBERT not only achieved superior performance in financial text analysis but also served as a foundational example for future domain-specific LLM development.

In his 2023 article published in the *Journal of Economic Literature*, Korinek presents *"Generative AI for Economic Research: Use Cases and Implications"*, a comprehensive and structured roadmap outlining how generative large language models (LLMs) can be integrated across the full lifecycle of economic research. The study maps out applications at various stages of research, such as literature review, coding, drafting, and the automation of micro-tasks, and presents them in a structure aligned with the JEL classification, making it highly relevant for academic economists.

The core message of the paper is clear: while LLMs offer substantial potential for productivity gains, these benefits can only be realized reliably if their use is accompanied by careful evaluation, transparent documentation, and continued human oversight. In other words, although automation is powerful, ensuring the trustworthiness and ethical soundness of AI-assisted research still depends on the judgment and involvement of human researchers.

This paper has quickly become a reference point for methodological norms surrounding the use of generative AI in economics. It provides guidance on how to use AI tools responsibly, not only to enhance research output but also to maintain academic integrity and rigor in an era of rapidly evolving technology.

In his 2023 paper titled *"Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?"*, Horton introduces a groundbreaking approach by proposing the use of large language models (LLMs) as "simulated agents" in economic settings. In this framework, LLMs are assigned characteristics such as endowments, preferences, and information, much like human participants, and their behavior is examined within stylized economic scenarios.

The methodology relies on prompt-based elicitation and experiment-style scenario design. The LLMs are presented with situations inspired by classic economic games or market environments, and their responses are compared to well-documented patterns from behavioral and experimental economics. This allows researchers to assess how closely LLM-generated decisions align with human-like behavior, such as rational choice, preference consistency, or information processing.

The importance of this work lies in its demonstration that LLMs can be used not only as text generators, but also as tools for rapid prototyping in economics research. Researchers can use LLMs to quickly test experimental designs, pre-evaluate survey instruments, or explore behavioral hypotheses in a cost-effective and scalable way. Horton's study positions LLMs as a new class of research tools capable of enhancing flexibility and speed in social science experimentation.

In their 2024 article published in the BIS Quarterly Review, Kwon, Park, Perez-Cruz, and Rungcharoenkitkul present "Large Language Models: A Primer for Economists," which is a comprehensive and practice-oriented guide designed to help economists understand and apply large language models (LLMs) in an end-to-end workflow. The paper outlines the full process in a clear and accessible way, covering data organization, prompt design, signal extraction from text, quantitative analysis, and evaluation.

Importantly, the approach is not just theoretical; the authors demonstrate its application by using a large news corpus to analyze the drivers of stock market dynamics. This real-world example highlights how LLMs can be used to interpret and quantify complex economic phenomena using textual data.

The significance of this work lies in its contribution as a practical reference for economists and policymakers aiming to use LLMs in a transparent, reproducible, and systematic way. In particular, it supports the growing interest in using LLMs within central banking and policy analysis, offering a technically sound and institutionally relevant framework for integrating these tools into economic research and decision-making processes.

In their 2024 article published in the BIS Quarterly Review, Kwon, Park, Perez-Cruz, and Rungcharoenkitkul present “Large Language Models: A Primer for Economists,” which is a practical and economist-friendly guide on how to effectively and systematically use large language models (LLMs) in economic analysis. The paper outlines a complete end-to-end LLM workflow, clearly explaining each step: data organization, prompt design, signal extraction from text, quantitative analysis, and evaluation of results.

Importantly, the study goes beyond theory by including a hands-on demonstration of the proposed approach. The authors apply their LLM-based workflow to a large corpus of news articles in order to analyze the key drivers behind stock market dynamics. This concrete example illustrates how unstructured text data can be transformed into measurable signals and how those signals can enhance our understanding of financial market behavior.

This paper serves as a valuable reference, particularly for central banking and policy analysis contexts, on how to use LLMs in a reproducible and reliable way. With its strong technical foundation, methodological clarity, and accessible guidance, it offers economists and policymakers a structured framework for integrating LLMs into institutional research and analysis.

In their 2024 paper titled *“CB-LMs: Language Models for Central Banking”*, Gambacorta, Kwon, Park, Patelli, and Zhu introduce a class of specialized models known as Central Banking Language Models (CB-LMs), designed to better process and understand the unique language used in central bank communications. These models are encoder-only architectures that have been retrained on domain-specific texts such as central bank speeches, policy documents, and research papers.

The key finding of the study is that these CB-LMs outperform their foundational models on tasks tailored to central bank language, such as masked-word prediction in policy-specific phrasing and terminology. This demonstrates that domain-specific retraining can significantly enhance a model's performance in specialized linguistic contexts.

The broader significance of this work lies in its clear message: general-purpose language models may not be sufficient for understanding nuanced, institution-specific communication. By contrast, custom models trained on institutional language can provide much more accurate and context-aware interpretations. CB-LMs, therefore, offer a compelling case for developing organization-specific language models, not just for central banks but potentially for other domains where precise language understanding is critical.

In their 2024/2025 paper titled “Macroeconomic Forecasting with Large Language Models”, Carriero, Pettenuzzo, and Shekhar conduct a pioneering evaluation of Time Series Language Models (TSLMs), which are large language model-based architectures applied to macroeconomic forecasting. The study systematically compares these models against traditional time-series methods widely used in macroeconomics.

The authors assess the performance of TSLMs across a range of macroeconomic indicators, analyzing under which conditions LLM and transformer architectures excel, and when traditional forecasting approaches remain more effective. In doing so, the paper builds a conceptual and methodological bridge between modern language-model-style forecasting tools and classical macroeconomic benchmarks.

This research is not only theoretical but also offers empirical insights through applied experiments, showing how LLM-based models can handle numerical time-series data and how their predictive power varies across contexts. Importantly, the paper is available as a versioned preprint, with updated versions frequently released that include new experimental results and revised dates, making it a dynamic and evolving contribution to the literature.

Overall, the study brings to the forefront key methodological discussions about the role LLM-based architectures might play in the future of macroeconomic forecasting.

In their 2025 working paper titled *"Large Language Models: An Applied Econometric Framework"*, Ludwig, Mullainathan, and Rambachan propose an econometric framework for the systematic and reliable use of large language models (LLMs) in empirical economic research. The framework explicitly distinguishes between two types of LLM applications: prediction tasks and measurement/estimation tasks, and outlines distinct methodological requirements for each.

One of the paper's key contributions is its clear articulation of the following: in prediction settings, researchers must take care to prevent training-data leakage; in measurement or estimation contexts, they should collect validation data that allows LLM errors to be quantified and formally incorporated into the analysis. This distinction ensures that any uncertainty or bias in LLM outputs can be acknowledged and properly managed within empirical research.

The broader significance of this work lies in its provision of "econometrics-compatible" guardrails, which are guidelines that help economists integrate LLMs into measurement and inference tasks without compromising methodological rigor. While recognizing the power and flexibility of LLMs as tools for empirical analysis, the authors emphasize that this power must be balanced with strict methodological controls and careful model evaluation.

In their 2025 paper titled *"Enhancing GDP Nowcasts with ChatGPT: A Novel Application of PMI News Releases"*, de Bondt and Sun explore how ChatGPT can be used to classify an "activity sentiment score" from Purchasing Managers' Index (PMI) news releases, and how that score performs in nowcasting euro area GDP. In the study, ChatGPT is prompted to automatically analyze PMI texts and generate a sentiment score that reflects the tone and direction of economic activity in a given month.

The main finding is that these text-derived sentiment scores provide incremental information beyond traditional PMI benchmarks, such as diffusion indices, and significantly improve the accuracy of GDP nowcasts. This highlights that text-based interpretations of economic releases can offer valuable predictive signals, complementing standard quantitative indicators.

The significance of this work lies in its demonstration that small but information-rich texts, such as policy statements and economic bulletins, can materially enhance real-time macroeconomic forecasting. It also serves as an early example of how generative AI tools like ChatGPT can be meaningfully integrated into economic analysis and institutional forecasting workflows.

In their 2025 working paper titled *"From Text to Quantified Insights: A Large-Scale LLM Analysis of Central Bank Communication"*, Silva, Moriya, and Veyrune present a fine-tuned large language model (LLM) framework designed to analyze central bank communication at scale. The framework classifies central bank statements at the sentence level across multiple dimensions, including topic, stance, sentiment, and target audience.

The key contribution of this work is its ability to transform qualitative policy language into structured, time-series-like indicators. This allows for the systematic linking of shifts in communication to financial markets, expectations, and macroeconomic outcomes. The LLM-driven analysis makes it possible to measure not only what central banks are saying, but also how their messaging evolves over time and how consistent or influential it is.

From a policy perspective, the framework provides a valuable tool for monitoring the clarity and consistency of central bank communications and for conducting cross-institutional benchmarking of communication strategies. Its ability to analyze the communication of multiple central banks under a unified methodological structure makes it especially relevant for comparative policy analysis and institutional decision-making.

In their 2025 article published in *Economics Letters*, Feyzollahi and Rafizadeh present *"The Adoption of Large Language Models in Economics Research"*, a study that proposes a novel method to estimate how widely large language models (LLMs) have been adopted in academic economics. The authors introduce the concept of a “linguistic footprint,” which refers to a set of textual patterns that indicate LLM-assisted writing, and they apply a difference-in-differences econometric design to analyze changes in writing style across leading economics journals.

The paper focuses particularly on the period following the release of ChatGPT, aiming to detect structural shifts in academic writing that are consistent with the influence of LLM tools. Drawing on a long time horizon of publications from top-tier journals, the study identifies meaningful changes in linguistic features that align with the broader diffusion of AI-assisted writing in economics.

This work is significant because it provides some of the earliest empirical evidence on the diffusion and adoption of LLM tools in academic economic research. It offers a data-driven perspective on how artificial intelligence is beginning to shape scholarly output and research practices in the field of economics.

In the 2025 study titled *“Generating Inflation Expectations with Large Language Models”*, Zarifhonarvar explores how large language models (LLMs) can be used as “agent” proxies to generate inflation expectations in survey-experiment settings. The author uses both proprietary and open-source LLMs, prompting them to act as agents forming expectations about future inflation within carefully designed experimental survey scenarios.

One of the key findings is that the architecture of the model significantly influences both the level and the degree of disagreement in the inflation expectations it generates. Some LLMs exhibit systematic biases relative to actual realized inflation, indicating that not all models are equally accurate or reliable as stand-ins for human expectations.

This research is important because it adds to the growing discussion about the ability of LLMs to simulate expectation formation in economic models and the conditions required for their effective use. It also highlights how sensitive the results are to model selection, offering methodological insights for researchers using LLM agents in behavioral or macroeconomic experiments.

4. Conclusion: Future Outlook for the Use of LLMs in Economics

This review has shown that large language models (LLMs) have rapidly integrated into economics research and have taken on a wide range of roles. However, these developments represent only the beginning. In the coming years, the potential use of LLMs in economics is expected not only to deepen existing methods but also to pave the way for entirely new analytical paradigms.

Several areas are likely to see major advances:

- **Policy Simulation and Decision Support Systems:** LLMs may evolve into “policy simulation agents” capable of analyzing the potential impact of economic policies under different scenarios. This would enable policymakers to test their decisions more rapidly and with data-driven insights.
- **Automated Modeling and Data Narration (Data-to-Text):** LLMs could serve as systems that interpret quantitative data and automatically generate reports, producing natural language summaries for central banks and public institutions.
- **Institution-Specific Models:** As seen in the CB-LM example for central banks, we may see universities, financial institutions, and regulatory bodies developing models tailored to their internal communication styles. This could fundamentally transform how organizations manage knowledge and analyze their own content.
- **Experimental Economics and LLM Agents:** The use of LLMs as simulated “decision-making agents” could go beyond traditional surveys and lab experiments, enabling low-cost and rapid behavioral simulations in experimental economics.
- **New Econometric Approaches:** The integration of LLM outputs into econometric models will require new techniques, including the modeling of measurement error and the incorporation of LLM-derived variables into causal inference frameworks.

In sum, LLMs are not merely auxiliary tools; they are becoming interactive and generative research partners capable of bridging data, theory, and policy in new ways. For the field of economics, this signals the beginning of a new era, both methodologically and epistemologically.

References

Gentzkow, M., Kelly, B., & Taddy, M. (2017). *Text as Data*. NBER Working Paper No. 23276.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). *Text as Data*. *Journal of Economic Literature*, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv:1908.10063.

Korinek, A. (2023). *Generative AI for Economic Research: Use Cases and Implications*. *Journal of Economic Literature*. https://doi.org/10.1257/jel.20231736

Horton, J. J. (2023). *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* arXiv:2301.07543.

Kwon, B., Park, T., Perez-Cruz, F., & Rungcharoenkitkul, P. (2024). *Large language models: a primer for economists*. *BIS Quarterly Review* (Dec 2024).

Kwon, B., Park, T., Perez-Cruz, F., & Rungcharoenkitkul, P. (2024). *Large language models: a primer for economists*. *BIS Quarterly Review* (December 2024).

Gambacorta, L., Kwon, B., Park, T., Patelli, P., & Zhu, S. (2024). *CB-LMs: language models for central banking*. *BIS Working Papers* No. 1215 (01 October 2024).

Carriero, A., Pettenuzzo, D., & Shekhar, S. (2024/2025). *Macroeconomic Forecasting with Large Language Models*. arXiv:2407.00890 (versioned preprint).

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). *Large Language Models: An Applied Econometric Framework*. NBER Working Paper No. 33344.

de Bondt, G. J., & Sun, Y. (2025). *Enhancing GDP Nowcasts with ChatGPT: A Novel Application of PMI News Releases*. *ECB Working Paper Series* No. 3063.

Silva, T. C., Moriya, K., & Veyrune, R. (2025). *From Text to Quantified Insights: A Large-Scale LLM Analysis of Central Bank Communication*. *IMF Working Paper* WP/IEA/2025/109.

Feyzollahi, M., & Rafizadeh, N. (2025). *The Adoption of Large Language Models in Economics Research*. *Economics Letters*, 250, 112265.

Zarifhonarvar, A. (2025). *Generating Inflation Expectations with Large Language Models*. *Journal of Economic Behavior & Organization* (in press/2025).

Chapter 9

Nanofluid-Based Thermal Management Systems: Next-Generation Cooling Approaches for Lithium-Ion Batteries

Mustafa AKKAYA[1]

1.Introduction

The transport sector has undergone significant change as a result of efforts to reduce dependence on fossil fuels and mitigate environmental issues. Electric vehicles (EVs), which offer a zero-emission driving experience, are at the heart of this transformation [1]. The most important component determining the performance, range, and safety of electric vehicles is the energy storage system or battery pack. Lithium-ion (Li-ion) batteries are the most popular technology in electric vehicles due to their advantages, such as high energy density, long cycle life, and low self-discharge rate [2]. However, lithium-ion batteries must be maintained within a specific temperature range to operate safely and effectively. The ideal operating temperature for batteries is typically between 20°C and 40°C (293 K and 313 K) [3]. Exceeding this limit can significantly reduce battery performance. The battery's internal resistance increases at low temperatures, causing a significant drop in power and capacity output. During charging, lithium plating may occur on the anode surface, leading to irreversible capacity loss and safety risks [4]. High temperatures can cause even more harmful consequences. If the battery temperature exceeds 60°C, the chemical reactions inside the battery accelerate and can cause thermal runaway, which is an uncontrollable increase in heat. The battery cell catching fire, exploding, or even the entire battery pack catching fire can lead to catastrophic consequences due to thermal runaway [5]. Therefore, to ensure the safety and longevity of electric vehicles, it is crucial to maintain the battery pack's temperature within the ideal range and ensure an even distribution of heat. The system that performs this task is known as the Battery Thermal Management System

---

[1] Manisa Celal Bayar University, Energy System Engineering, mustafa.akkaya@cbu.edu.tr
0000-0003-4726-062X

(BTMS). An effective BTMS directly affects battery life, safety, and performance. Air cooling, liquid cooling, cooling with phase change materials (PCM), and heat pipes are traditional BTMS approaches [6]. Due to their high heat transfer capacities, liquid cooling systems are one of the most common and effective methods today. However, the thermal conductivity of traditional coolants (e.g., water and ethylene glycol mixtures) is limited. Therefore, smaller and more effective BTMS designs are not possible. At this point, nanofluids, a new technology in heat transfer, come into play. First proposed by Choi and Eastman in 1995, nanofluids are a new generation of fluids created by dispersing nano-sized (1-100 nm) solid particles (metals, metal oxides, carbon nanotubes, etc.) into traditional coolants [7]. The thermal conductivity of the base fluid is significantly increased with the help of these nanoparticles, which improves heat transfer performance. These properties give nanofluids the potential to create more efficient, lighter and smaller thermal management systems for lithium-ion batteries. In recent years, due to developments in semiconductor technology, improving the performance of electronic circuits and batteries has become a highly significant issue. DC circuits and drivers, along with DC-DC power converters, form the core application areas in semiconductor technology-based applications. The efficiency and stability of electronic circuits are system parameters directly affected by the operating temperature of the circuit. Although losses and temperature values remain at acceptable levels in these technologies, under conditions such as high switching frequencies, the loss and temperature values of DC circuit elements can reach abnormal levels. Single-wire structures are generally used when creating electronic circuits. In DC circuits, instead of single-wire conductors, it is possible to use tubular conductors and cool the circuit by circulating a fluid within these tubular conductors [8-10]. Ahammed et al. stated in their work that they achieved thermoelectric cooling of electronic devices using nanofluids they had prepared, with the aid of a heat exchanger. In the study, they used bismuth telluride ($BiTe_3$) as the thermoelectric cooler. They stated that they wanted to reduce the heat of the 20 W power transistor in the electronic device they used in their experiments. In the thermally prepared nanofluid, they used water as the base fluid and Al_2O_3 nanoparticles at volume fractions of 0.1% and 0.2% as the nanoparticles. The nanofluids used in the electronic device with the aid of a heat exchanger were used to dissipate heat. The experiments revealed that thermal conductivity increased with the increase in nanoparticle volume fraction. They reported that the cooling performance coefficient increased by up to 40% when using the nanofluid prepared at a 0.2% volume fraction in the thermoelectric module [11]. Nguyen et al. designed a closed system for cooling microprocessors or other electronic components to increase their operational efficiency. In their study, they reported that they prepared a nanofluid for use in a closed cooling system to improve heat transfer in the comprehensively designed system. The authors stated that they used pure water as the base fluid and Al_2O_3 as the nanoparticle in the nanofluid they prepared. They indicated that they used solid nanoparticles of 36 and 47 nm in size added to the base fluid to prepare the nanofluid. They concluded that as the particle size decreased, the nanofluid provided a higher heat transfer coefficient. As a result of the experiments, they reported that they observed a significant increase in the convective heat transfer coefficient of the cooling block [12]. Ljam and Saidur stated that they used nanofluids for thermal control in high heat dissipation conditions in new generation electronic systems. They used the nanofluid they prepared in a 20 cm long mini-channel heat sink to cool the electronic system. They reported that they used water as the base fluid and SiC and TiO_2 nanoparticles as nanoparticles when preparing the nanofluid. The authors observed that the thermal conductivity of water increased with the dispersion of nanoparticles in the suspension. When using TiO_2 nanoparticles at a volume fraction of 4.0% and water as the base fluid, they reported a 7.63% increase in heat flux when the fluid flow rate in the heat sink was 2 m/s. They stated that when the liquid flow velocity in the heat sink was increased to 6 m/s, the maximum increase in heat flow was calculated to be 12.43% [13]. Colangelo et al. drew attention to the frequent use of new generation technologies. They reported that the biggest problem when using new generation technologies is the cooling of electronic devices. They mentioned that the miniaturisation of electronic devices used today has many advantages, but also highlighted the negative effects of the heat generated due to the reduction in surface area. They reported that nanofluid applications have yielded positive results in many studies on the cooling of electronic devices. They have reported that Al_2O_3 nanoparticles are most frequently used in nanofluid applications due to their thermophysical properties and economic size [14]. Selvaraj and Krishnan have investigated the properties of hybrid nanoliquid materials. Within the nanocomposite material, they investigated the optimum concentration values of the graphene nanoparticles they used to increase dispersion stability. At the end of the experimental process, they stated that the hybrid nanofluids they prepared achieved high dispersion stability (zeta potential: -35.21 mV). Thus, they concluded that nanofluids prepared with graphene-coated alumina nanocomposite materials are highly suitable for thermal management applications in electronic devices and have high potential for heat transfer [15]. Siricharoenpanich et al. reported the necessity of dissipating the heat generated by various components of electronic devices in order to obtain higher efficiency from these devices and increase their reliability. They emphasised the need to remove high-value heat flux from the system in order to operate at lower temperatures in such systems. Previous studies have highlighted that air, water, and fluorochemicals are commonly used coolants to improve the cooling performance of electronic devices. Previous studies have emphasised that air, water, and fluorochemicals are commonly used coolants for improving the cooling performance of electronic devices. However, this study reports the use of nanofluids to increase thermal removal efficiency. The experiments revealed an increase of up to 11.94% in thermal distribution efficiency. They reported that the thermal distribution efficiency from the Ag/Fe_3O_4 nanofluid mixture as a cooling fluid was higher than that from nanofluids prepared with Ag nanoparticles and pure water used as a coolant [16].

The aim of this book chapter is to comprehensively examine the use of nanofluids in the thermal management of lithium-ion batteries. The thermal behaviour of lithium-ion batteries and current BTMS technologies will be addressed in this context. Subsequently, details will be provided on the fundamental concepts, properties, production techniques, and thermophysical behaviour of nanofluids. Finally, issues related to the use of nanofluids and future research approaches will be discussed, presenting the potential of this innovative technology.

2. Thermal behaviour of lithium-ion batteries

During charging and discharging processes, heat is generated due to the electrochemical structure of lithium-ion batteries. The battery's performance, lifespan, and safety are directly affected by this heat generation. Understanding the fundamental behaviour of the battery is the first and most important step in designing an effective thermal management system. Reversible and irreversible heat are the two main sources of heat generation within the battery. Joule heat, which arises from the battery's internal resistance, creates irreversible heat during charging and discharging. This is the most significant component of heat generation. Entropy changes in the electrode materials can cause reversible heat. It can be endothermic (heat absorbing) during charging and exothermic (heat releasing) during discharging [17]. The combination of these two mechanisms determines the total heat production. Temperature significantly affects battery performance. As mentioned earlier, the ideal operating temperature range for batteries is limited. When this limit is exceeded, a number of problems arise:

Low Temperature Effects: Charge transfer reactions in electrode materials slow down when the temperature drops, and the ionic conductivity of the electrolyte decreases. This increases the battery's internal resistance. As a result, both the power output and capacity of the battery decrease significantly. This is the main reason for the reduced range of electric vehicles, especially in cold weather [4].

High Temperature Effects: Batteries age due to high temperatures. Parasitic reactions such as the growth and decomposition of the solid-electrolyte interphase (SEI) layer increase, leading to the loss of active lithium and electrolyte, and thus a continuous decrease in capacity [18]. When the temperature exceeds a certain threshold (60-80 °C), the SEI layer becomes unstable and begins to decompose. This triggers excessively exothermic reactions between the anode material and the electrolyte. These reactions generate more heat and raise the temperature, initiating a chain reaction. This uncontrolled process is known as thermal runaway and can cause the cell to vent, smoke, burn, and even explode [5]. The temperature level and the homogeneity of the temperature distribution within the battery pack are very important. Temperature differences between cells within the pack or between specific areas of the same cell cause the cells to age at different rates. This results in electrical imbalance between the cells, which reduces battery life and overall performance. Generally, the total temperature variation within the battery should not exceed five degrees [19].

3. Battery Thermal Management Systems

Battery Thermal Management Systems (BTMS) are designed to protect the safety, performance and lifespan of lithium-ion battery packs. BTMSs have various architectures and operating principles. Generally, these systems are classified according to the cooling medium they use. Each method has its advantages and disadvantages. This section examines the leading BTMS technologies.

3.1. Air Cooling Systems

Air cooling is the simplest and most cost-effective solution for BTMS. These systems utilise ambient air to dissipate the heat generated by battery cells. Active air cooling systems require airflow via fans or blowers, whereas passive air cooling systems rely on natural convection. Due to its low cost, light weight and simplicity, it is preferred, especially in hybrid vehicles with low power requirements. However, cooling efficiency is low due to the low heat capacity and thermal conductivity of air. Therefore, it is insufficient in hot weather or in high-performance electric vehicles [20].

3.2. Liquid Cooling Systems

The heat capacity and thermal conductivity of liquids are much higher than those of air, which is why they are used in liquid cooling systems. In these systems, a coolant fluid consisting of a mixture of water and ethylene glycol is typically circulated through cooling plates or channels that pass around or through the battery modules. The liquid, which absorbs heat from the battery cells, is discharged into the environment via a radiator. Liquid cooling enables better regulation and more even distribution of battery temperature thanks to its high heat transfer efficiency. Due to these characteristics, it is the standard BTMS solution in most high-performance electric vehicles today [21]. Nanofluids can be used to improve the performance of these systems.

3.3. Phase Change Material (PCM) Based Systems

Phase Change Materials (PCM) can absorb or release large amounts of latent heat during freezing and melting. PCM-based BTMSs coat battery cells with a PCM matrix. When the battery heats up, the PCM begins to melt and absorbs a significant amount of heat. This passive method provides excellent temperature control. However, the biggest problem with PCMs is their low thermal conductivity, which causes them to absorb and dissipate heat slowly. Additionally, once the PCM is completely melted, its heat absorption capacity decreases significantly. To overcome these limitations, hybrid systems are being developed where metal foams or nanoparticles are added to PCMs to enhance their thermal conductivity [22, 23].

3.4. Heat Pipe Systems

Heat pipes, as passive heat transfer devices, possess extremely high thermal conductivity. A heat pipe consists of a sealed container containing a small amount of working fluid. The fluid evaporates at the hot end on the battery side, absorbing heat. The vapour moves towards the cold end, where it condenses and releases the absorbed heat to the surroundings. The wick structure ensures that the liquid returns to the hot end. This closed loop operates continuously without requiring power from any external source. Heat pipes are considered a good technology for BTMS applications due to their efficiency, reliability, and flexible design [24, 25].

3.5. Hybrid Systems

Hybrid systems aim to compensate for the weaknesses of different methods and combine their strengths by integrating two or more of the technologies mentioned above. An active liquid cooling loop can be used to solve the slow heat distribution problem of a PCM system. Similarly, a heat pipe system can effectively transfer heat to a PCM block or a liquid-cooled radiator located away from the batteries. Recently, there has been considerable interest in hybrid approaches combining nanofluids with PCM or other systems [17, 26]. A comparison of battery thermal management systems is provided in Table 1.

Table 1. Comparison of Battery Thermal Management Systems.

BTMS	Advantages	Disadvantages
Air Cooling	Simple, inexpensive, lightweight, maintenance-free.	Low heat transfer efficiency, may be noisy, inadequate in hot climates.
Liquid Cooling	High heat transfer efficiency, precise temperature control, compact design.	Complex structure, risk of leakage, additional pump power requirement, higher cost.
PCM-Based Cooling	Excellent temperature uniformity, passive operation, no moving parts.	Low thermal conductivity, limited heat absorption capacity, may be heavy.
Heat Pipe	Highly effective thermal conductivity, passive operation, reliable, flexible design.	It may be costly, it may be sensitive to gravity (in some designs).
Hybrid Systems	Combines the advantages of different technologies, high performance.	High complexity, high cost, potential weight increase.

4. Nanofluids: Fundamental Concepts and Properties

Over the past twenty years, the field of heat transfer has shown great interest in nanofluids, which have the potential to take the performance of liquid cooling systems to the next level. This section covers the basic definition of nanofluids, their components, and preparation techniques.

4.1. Definition and History of Nanofluids

By its most basic definition, a nanofluid is a two-phase mixture formed by the stable dispersion of solid particles at the nanoscale (nanoparticles) within conventional heat transfer fluids (base fluids). Stephen Choi and Jeffrey Eastman from Argonne National Laboratory first proposed this idea in 1995 [7]. In general, adding very small amounts of solid particles to a liquid demonstrates that solids, such as metals, have much higher thermal conductivity than liquids. Therefore, a new class of fluids with significantly better heat transfer properties than traditional fluids emerges [26].

4.2. Types of Nanoparticles

Nanoparticles are the most important components determining the properties of nanofluids. The thermophysical properties of a nanofluid are directly influenced by the material, size, shape, and concentration of the nanoparticles used. There are three main categories of nanoparticles used in battery thermal management:

Metal Oxide Nanoparticles: Metal oxides such as alumina (Al_2O_3), copper(II) oxide (CuO), titanium dioxide (TiO_2), and silicon dioxide (SiO_2) are the most commonly used types of nanoparticles due to their chemical stability, low cost, and ease of production. Al_2O_3 and CuO are particularly recognised for providing an excellent balance of thermal conductivity and stability [23, 27].

Metal Nanoparticles: Pure metal particles such as copper (Cu), silver (Ag), and aluminium (Al) have much higher thermal conductivity compared to metal oxides. Therefore, they are quite attractive for increasing heat transfer. For example, a study found that silver nanoparticles reduced battery temperature by over 30% [21]. However, the stability of metal nanoparticles can be compromised, and they are more prone to oxidation.

Carbon-Based Nanoparticles: Graphene and carbon nanotubes have very high thermal conductivity values of 3000–6000 W/m·K [6]. These materials can exhibit very high levels of thermal conductivity even at very low volumetric concentrations. Therefore, there is considerable potential for creating high-performance nanofluids [28, 29].

Table 2. Thermal Properties of Commonly Used Nanoparticles and Some Liquids [30, 31].

Nanoparticle	**Thermal Conductivity (W/m·K)**	**Density (kg/m³)**
Silver (Ag)	429	10500
Copper (Cu)	401	8933
Alümina (Al_2O_3)	40	3970
Copper Oxide (CuO)	76.5	6320
Titanium Dioxide (TiO_2)	8.9	4250
Carbon Nanotube (CNT)	~3000-6000	~1350-2100

4.3. Methods for Preparing Nanofluids

The success of nanofluids largely depends on how stably and homogeneously nanoparticles are dispersed in the base fluid. Over time, particle settling or agglomeration can cause blockages in channels. Therefore, the method of producing nanofluids is very important. There are two main preparation methods:

- Single-Step Method (One-Step Method): In this approach, the synthesis of nanoparticles and their dispersion in the base fluid are carried out in a single process. Using methods such as physical vapour deposition or chemical reduction, nanoparticles are produced directly in the liquid phase. The advantage of this method is that it reduces particle agglomeration, creating more stable suspensions. However, it is more complex and expensive [31].
- Two-Step Method: This is the most popular and economical method. The first step involves converting the nanoparticles into a dry powder. This is achieved using physical or chemical methods. In the second step, this nanoparticle powder is incorporated into a base liquid, and different dispersion techniques are used to obtain a stable suspension. These techniques include sonication, which uses high-frequency sound waves to break up particle agglomerates, and the addition of surfactants to prevent particles from sticking together. Surfactants such as cetyltrimethylammonium bromide (CTAB) and sodium dodecyl sulphate (SDS) prevent precipitation by coating the particle surfaces, creating electrostatic or steric repulsive forces [27, 32]. Although the simplicity and low cost of the system are its greatest advantages, the tendency of particles to re-agglomerate is the main drawback.

5. Applications of Nanofluids in Battery Cooling

Due to their superior thermophysical properties, nanofluids serve as an alternative to traditional coolants in various thermal management systems developed for lithium-ion batteries. Researchers have investigated various BTMS configurations to maximise the performance of nanofluids. This section discusses key application areas where nanofluids are used and the resulting performance improvements.

5.1. Performance Comparisons

Research in the literature indicates that various nanofluids and system designs significantly improve the cooling performance of batteries. The table below summarises the quantitative performance metrics of selected studies.

Table 3. Performance Improvements Achieved with Nanofluid Use in Selected Studies.

Reference	Nanoparticle/ Base Liquid	System Type	Performance Improvement
Dhanasekaran (2025) [21]	Ag / Water	Cooling Plate	30.2% decrease at maximum temperature
Wiriyasart et al. (2020) [2]	TiO_2 / Water	Cooling Plate (Corrugated)	A 28.65% decrease at maximum temperature
Bahrami (2025) [20]	Various	Compilation	A 14% increase in heat flow, a 22% decrease in hot spot formation
Liao et al. (2022) [3]	Cu / Water	Cooling Plate (Curved)	A 12.6% decrease in the maximum temperature difference
Venkateswarlu (2024) [19]	Hybrid(Al_2O_3-CuO)	Cylindrical Module	A 5.86% increase in heat transfer rate

These findings demonstrate the potential for lithium-ion batteries to operate at cooler, safer and more homogeneous temperatures when the correct nanofluids, nanoparticles and system design are selected.

6. CONCLUSION

The performance, safety and lifespan of lithium-ion battery packs are crucial for the future of electric vehicle technology. These factors are responsible for the effective management of battery temperature. While traditional liquid cooling systems meet current demands, new-generation batteries with higher power densities and fast-charging technologies require more advanced thermal management solutions. As discussed in this book chapter, nanofluids have the potential to revolutionise this field.

Literature reviews and analyses clearly demonstrate that nanofluids are superior to advanced coolants in heat transfer. The addition of various nanoparticles, such as Al_2O_3, CuO, CNT, and hybrid structures, to base fluids has significantly increased thermal conductivity. This has led to a significant reduction in battery temperatures. Studies have shown that nanofluids extend battery life and enhance safety by improving temperature homogeneity within the battery. Nanofluids have been shown to be beneficial in BTMS designs such as immersion cooling, cooling plates, and microchannels.

However, significant challenges need to be overcome for the technology to transition to real-world applications. Various obstacles stand in the way of the commercial use of nanofluids. These include long-term stability, the need for greater pumping power due to increased viscosity, potential erosion and wear effects, and high cost. Addressing these issues requires interdisciplinary research efforts spanning materials science, chemistry, and engineering disciplines.

In conclusion, nanofluids present a significant opportunity to develop more compact, efficient, and lighter thermal management systems for lithium-ion batteries. Despite the current challenges, this technology is expected to advance and become a standard solution in electric vehicles in the future. Nanofluids not only enhance battery performance but can also accelerate the adoption of electric vehicles, paving the way for a more sustainable transport ecosystem.

Acknowledgments: AI-based tools were employed to support the development and editing of the manuscript. The author(s) remain fully responsible for the content and interpretations presented.

REFERENCES

[1] Olabi, A. G., Maghrabie, H. M., Adhari, O. H. K., Sayed, E. T., Yousef, B. A., Salameh, T., ... & Abdelkareem, M. A. (2022). Battery thermal management systems: Recent progress and challenges. *Renewable and Sustainable Energy Reviews*, 162, 112414.

[2] Wiriyasart, S., Hommalee, C., Sirikasemsuk, S., Prurapark, R., & Naphon, P. (2020). Thermal management system with nanofluids for electric vehicle battery cooling modules. *Case Studies in Thermal Engineering*, 18, 100590.

[3] Liao, G., Wang, W., Zhang, F., Chen, J., & Leng, E. (2022). Thermal performance of lithium-ion battery thermal management system based on nanofluid. *Applied Thermal Engineering*, 216, 118997.

[4] Ma, S., Jiang, M., Tao, P., Song, C., Wu, J., Wang, J., ... & Shang, W. (2018). Temperature effect and thermal impact in lithium-ion batteries: A review. *Progress in Natural Science: Materials International*, *28*(6), 653-666.

[5] Shahid, S., & Dubois, C. (2022). A review of thermal runaway prevention and mitigation strategies for lithium-ion batteries. *Journal of Energy Storage*, 55, 105411.

[6] Zhou, Z., Lv, Y., Qu, J., Sun, Q., & Grachev, D. (2021). Performance evaluation of hybrid oscillating heat pipe with carbon nanotube nanofluids for electric vehicle battery cooling. *Applied Thermal Engineering*, 196, 117277.

[7] Choi, S. U. (1995, November). Enhancing thermal conductivity of fluids with nanoparticles. In *ASME international mechanical engineering congress and exposition. American Society of Mechanical Engineers. 17421, 99-105.*

[8] Balci, S. 2019. “A CFD simulation of the liquid-cooled pipe conductors for the high power and high frequency power electronic circuits”, Measurement, 147, 106885.

[9] Kordetoodeshki, E., & Hassanzadeh, A. 2019. “An ultra-low power, low voltage DC-DC converter circuit for energy harvesting applications”, AEU-International Journal of Electronics and Communications, 98, 8-18.

[10] Ashraf, M., & Masoumi, N. 2014. “High efficiency boost converter with variable output voltage using a self-reference comparator”, AEU-International Journal of Electronics and Communications, 68(11), 1058-1064.

[11] Ahammed, N., Asirvatham, L. G., & Wongwises, S. 2016. “Thermoelectric cooling of electronic devices with nanofluid in a multiport minichannel heat exchanger”, Experimental Thermal and Fluid Science, 74, 81-90.

[12] Nguyen, C. T., Roy, G., Gauthier, C., & Galanis, N. 2007. "Heat transfer enhancement using Al2O3–water nanofluid for an electronic liquid cooling system", Applied Thermal Engineering, 27(8-9), 1501-1506.

[13] Ijam, A., & Saidur, R. 2012. "Nanofluid as a coolant for electronic devices (cooling of electronic devices)", Applied Thermal Engineering, 32, 76-82.

[14] Colangelo, G., Favale, E., Milanese, M., de Risi, A., & Laforgia, D. 2017. "Cooling of electronic devices: Nanofluids contribution", Applied Thermal Engineering, 127, 421-435.

[15] Selvaraj, V., & Krishnan, H. 2020. "Synthesis of graphene encased alumina and its application as nanofluid for cooling of heat-generating electronic devices", Powder Technology, 363, 665-675.

[16] Siricharoenpanich, A., Wiriyasart, S., Srichat, A., & Naphon, P. (2020). "Thermal cooling system with Ag/Fe3O4 nanofluids mixture as coolant for electronic devices cooling", Case Studies in Thermal Engineering, 100641.

[17] Jilte, R., Afzal, A., Ağbulut, Ü., Shaik, S., Khan, S. A., Linul, E., & Asif, M. (2023). Battery thermal management of a novel helical channeled cylindrical Li-ion battery with nanofluid and hybrid nanoparticle-enhanced phase change material. *International Journal of Heat and Mass Transfer*, 213, 124311.

[18] Abdolrasol, M. G., Ansari, S., Sarker, I. A., Tiong, S. K., & Hannan, M. A. (2025). Lithium-ion to sodium-ion batteries transitioning: trends, analysis and innovative technologies prospects in EV application. *Progress in Energy*, *7*(2), 022007.

[19] Venkateswarlu, B., Chavan, S., Joo, S. W., & Kim, S. C. (2024). Impact of hybrid nanofluids on thermal management of cylindrical battery modules: A numerical study. *Journal of Energy Storage*, 82, 110556.

[20] Bahrami, H. R., Allahdadi, O., & Saffari, H. (2025). Nanofluids for advanced thermal management: A multi-criteria review of performance, cost, and environmental impacts. *Results in Materials, 27, 100753.*

[21] Dhanasekaran, A., Subramanian, Y., Dhanasekaran, R., et al. (2025). Silver nanoparticle-based nanofluid vs. conventional cooling approach: a comparative study on thermal management in lithium-ion battery packs. *Journal of Umm Al-Qura University for Engineering and Architecture, 16, 1382–1400.*

[22] Kiani, M., Ansari, M., Arshadi, A. A., Houshfar, E., & Ashjaee, M. (2020). Hybrid thermal management of lithium-ion batteries using nanofluid, metal foam, and phase change material: an integrated numerical–experimental approach. *Journal of Thermal Analysis and Calorimetry*, 140(4), 1867-1880.

[23] Senthilraja, S., Vijayakumar, K., & Gangadevi, R. (2015). A comparative study on thermal conductivity of Al2O3/water, CuO/water and Al_2O_3–CuO/water nanofluids. *Digest Journal of Nanomaterials and Biostructures*, 10(2), 571-579.

[24] Çiftçi, E., Sözen, A., & Karaman, E. (2016). TiO2 İçeren Nanoakışkan Kullanımının Isı Borusu Performansına Etkisinin Deneysel Olarak İncelenmesi, sayfa: 367-376. *Politeknik Dergisi*, *19*(3), 367-376.

[25] Sözen, A., Gürü, M., Khanlari, A., & Çiftçi, E. (2019). Experimental and numerical study on enhancement of heat transfer characteristics of a heat pipe utilizing aqueous clinoptilolite nanofluid. *Applied Thermal Engineering*, *160*, 114001.

[26] Kiani, M., Ansari, M., Arshadi, A. A., Houshfar, E., & Ashjaee, M. (2020). Hybrid thermal management of lithium-ion batteries using nanofluid, metal foam, and phase change material: an integrated numerical–experimental approach. *Journal of Thermal Analysis and Calorimetry*, 140(4), 1867-1880.

[27] Asadi, A., Asadi, M., Siahmargoi, M., Asadi, T., & Andarati, M. G. (2017). The effect of surfactant and sonication time on the stability and thermal conductivity of water-based nanofluid containing Mg (OH) 2 nanoparticles: An experimental investigation. *International Journal of Heat and Mass Transfer*, 108, 191-197.

[28] Shijina, S. S., & Nagulan, B. (2025). Graphene functionalized nano-encapsulated composite phase change material based nanofluid for thermal management of electric vehicle batteries. *Applied Thermal Engineering*, 256, 124561

[29] Liu, M., Lin, M. C., & Wang, C. (2011). Enhancements of thermal conductivities with Cu, CuO, and carbon nanotube nanofluids. *Journal of Nanoparticle Research*, 13(12), 6489-6498.

[30] Tao, Q., Zhong, F., Deng, Y., Wang, Y., & Su, C. (2023). A Review of Nanofluids as Coolants for Thermal Management Systems in Fuel Cell Vehicles. *Nanomaterials*, 13(21), 2861.

[31] Rahman, M. A., Hasnain, S. M., Pandey, S., Tapalova, A., Akylbekov, N., & Zairov, R. (2024). Review on Nanofluids: Preparation, Properties, Stability, and Applications. *ACS omega*.

[32] Ghadimi, A., & Metselaar, I. H. (2013). The influence of surfactant and ultrasonic processing on improvement of stability, thermal conductivity and viscosity of titania nanofluid. *Experimental Thermal and Fluid Science*, 51, 1-9.

Chapter 10

Evolution of Mobile Communication Technologies from 3G to 6G: Architectural Transformation, Performance Analysis, and Simulation-Based Approach

Ali Özhan AKYÜZ[1], Kazım KUMAŞ[2], Mustafa AYAN[3], Durmuş TEMİZ[4]

1. Introduction

Mobile communication technologies have become a critical domain, constituting the backbone of modern society's economic, social, and technological structure. Especially since the first quarter of the 21st century, the dramatic increase in mobile data traffic has necessitated the development of new-generation mobile networks. According to periodic reports by Cisco and the ITU, mobile data traffic grows by an average of 30–40% annually. This rapid increase necessitates higher capacity, lower latency, greater device density, and higher energy efficiency (Cisco, 2020; ITU, 2023).

Mobile communication systems are comprehensive infrastructures that directly impact not only the telecommunications sector but also transportation, health, education, defense, industrial automation, logistics, and urban planning. Therefore, the evolution of mobile technologies is not merely a technical development; it is also a multidimensional process shaping economic growth, social transformation, and digitalization.

The primary objective of this book chapter is to examine the journey of mobile network technologies from the third to the fifth generation within both theoretical and applied frameworks. In this context, the chapter does not address

---

[1] Prof. Dr., Burdur Mehmet Akif Ersoy University, Bucak Emin Gülmez Vocational School, Bucak, Burdur, Türkiye, ORCID ID: 0000-0001-9265-7293

[2] Lecturer, Burdur Mehmet Akif Ersoy University Bucak Emin Gülmez Vocational School, Burdur, Türkiye, ORCID ID: 0000-0002-2348-4664

[3] Lecturer, Burdur Mehmet Akif Ersoy University Bucak Emin Gülmez Vocational School, Burdur, Türkiye, ORCID ID: 0000-0001-6780-0625

[4] Lecturer Dr., Burdur Mehmet Akif Ersoy University Bucak Emin Gülmez Vocational School, Burdur, Türkiye, ORCID ID: 0000-0002-7350-7502

the historical evolution of mobile communication solely chronologically; it also encompasses architectural transformations, innovations in radio access technologies, developments in spectrum management, paradigmatic shifts in core network design, and the diversification of use cases.

Technologies such as 3G HSPA, 4.5G LTE-Advanced, and 5G Network Slicing, in particular, demonstrate how mobile networks have been enhanced at different levels. Innovations such as high-speed packet access starting with 3G, the enhancement of spectral efficiency with 4.5G, and ultra-reliable, ultra-fast, and flexible service delivery with 5G represent the three critical links of the technology chain (Holma & Toskala, 2011; Dahlman et al., 2020; Andrews et al., 2014).

The simulation components in this chapter are designed not only to provide a visual interface but also to support analytical understanding of the working principles of mobile networks. The animated demonstration of packet flows, the visual integration of carrier aggregation, the independent movement of network slices, and map-based coverage analysis are simplified yet conceptually consistent representations of tools used in modern network engineering.

Consequently, this chapter aims to present mobile communication technologies holistically, combining academic content with a strong technical foundation and conceptual explanations, supported by interactive visualization. In this regard, the study specifically targets engineering students and educational environments, aspiring to bridge the gap between abstract theory and practice. By transforming complex network dynamics into observable models, it provides a tangible learning experience that reinforces technical competence in telecommunications (de Jong, Linn, & Zacharia, 2013).

2. Basic Architecture of Mobile Networks

The architecture of mobile communication networks is hierarchical, composed of various functional layers, and is primarily built upon three main components: the Radio Access Network (RAN), the Core Network (CN), and User Equipment (UE). Each of these components relies on different standards, protocols, and architectural design philosophies that have evolved from one generation to the next (Rappaport, 2002).

2.1. Radio Access Network (RAN) and Its Transformation

The Radio Access Network (RAN) is the edge layer that provides the physical and logical connection between mobile devices and the network center. This layer executes critical Radio Resource Management (RRM) functions, including channel allocation, selection of modulation and coding schemes, power control, signal level measurement, and handover operations to support seamless communication for moving users. The RAN architecture has undergone significant structural changes during the transition from 3G to 5G.

In 3G networks, the RAN consists of a two-tier structure comprising the NodeB and the Radio Network Controller (RNC). In this structure, the RNC manages multiple NodeBs, controlling radio connections and coordinating resource allocation. This architecture, in which the majority of control-plane operations are centralized in the RNC, represents a hierarchical, centralized network model. However, with the transition to 4G LTE technology, this hierarchy was simplified, and the RNC was eliminated. In the 4G architecture, the "Evolved NodeB" (eNodeB) consolidated all radio control functions within itself, thereby reducing latency, accelerating handover operations, and establishing a more distributed network structure. 5G NR (New Radio) has taken this evolution a step further by introducing the concept of the gNB (next-generation NodeB). The gNB architecture can separate the control plane and user plane into independently scalable components (C/U-plane split). This separation constitutes the foundation of contemporary Cloud-Based RAN (Cloud-RAN or C-RAN) architectures (Holma & Toskala, 2011; Rappaport, 2002; Ghosh et al., 2012; 3GPP, 2020; Boccardi et al., 2014).

2.2. Core Network (CN): The Brain of the System

Characterized as the brain of mobile systems, the Core Network performs vital functions such as auditing user network access authorizations, authenticating users, routing data packets, and operating quality of service mechanisms.

In 3G technology, the core network is built upon two fundamental elements: the SGSN (Serving GPRS Support Node) and the GGSN (Gateway GPRS Support Node). While the SGSN manages user mobility, the GGSN acts as a gateway connecting the user to external IP networks. In the 4G era, this structure was simplified to a fully packet-switched architecture called the Evolved Packet Core (EPC). Within the EPC structure, low-latency data transmission is achieved using the MME (Mobility Management Entity) for control-plane operations and the SGW (Serving Gateway) and PGW (Packet Data Network Gateway) for data-plane operations.

With 5G, the core network design underwent a radical change, evolving from a traditional monolithic structure to a microservices-based Service-Based Architecture (SBA). In the 5G Core (5GC) structure, functions such as Access and Mobility Management (AMF), Session Management (SMF), User Plane Function (UPF), and Network Repository Function (NRF) are completely decoupled from one another. This modular separation has enabled the realization of advanced capabilities such as network slicing, distributed edge computing, and dynamic service provisioning (Holma & Toskala, 2011; 3GPP, 2020).

2.3. User Equipment (UE) and Hardware Complexity

User Equipment (UE) has gone beyond being a simple communication tool and has become a complex radio processor. A modern UE houses a multi-mode RF transceiver, adaptive antenna structures, precise power control mechanisms, multi-carrier support, and SIM/USIM security subsystems. During the transition from 3G to 5G, device hardware has become more complex; specifically, in 5G-compatible devices, millimeter-wave (mmWave) modules, beamforming units, and advanced antenna arrays have become standard components for high-frequency operation (Boccardi et al., 2014).

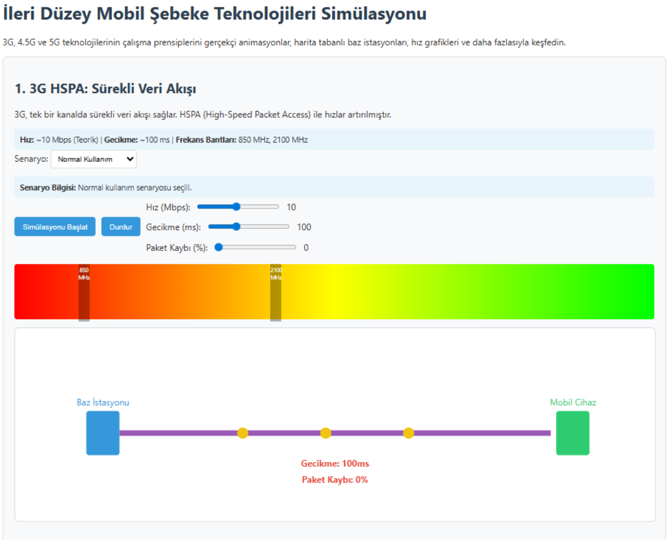
2.4. Fundamentals of Cellular Design

The fundamental working principle of cellular architecture is to divide a specific geographical area into sub-regions, called cells, based on capacity and coverage requirements. Cell sizes vary widely depending on the frequency band used, antenna height, transmission power, and environmental conditions. Macro cells covering wide areas (1–20 km), micro cells providing capacity increases (500–2000 m), pico cells in dense areas (10–200 m), and femtocells for indoor solutions are used in combination to create Heterogeneous Networks (HetNets). This multi-layered structure is critical, particularly in dense city centers, for increasing capacity, managing signal interference, and filling coverage gaps (Ghosh et al., 2012; Ge et al., 2016).

3. Design And Technical Operation of the 3G Hspa Simulation

The technical operation of the developed 3G HSPA simulation is comprehensively evaluated, and the methods by which the HTML, CSS, and JavaScript-based architecture models the unique working dynamics of 3G networks are elucidated in detail. The simulation design aims to transform the single-carrier, continuous-data transmission structure of 3G into a simple yet pedagogically effective animation. When the code's holistic structure is examined, it is observed that both the interface components and the timing mechanisms that manage the data flow are designed to complement one another. The simulation's fundamental visual structure is based on a scene represented in Scalable Vector Graphics (SVG). In this scene, the Base Station (NodeB) is positioned on the left, and the User Equipment (UE) on the right;

data transmission between these endpoints is represented by three data packets moving along a single horizontal channel as illustrated in Figure 1 (Rappaport, 2002).



İleri Düzey Mobil Şebeke Teknolojileri Simülasyonu

3G, 4.5G ve 5G teknolojilerinin çalışma prensiplerini gerçekçi animasyonlar, harita tabanlı baz istasyonları, hız grafikleri ve daha fazlasıyla keşfedin.

1. 3G HSPA: Sürekli Veri Akışı

3G, tek bir kanalda sürekli veri akışı sağlar. HSPA (High-Speed Packet Access) ile hızlar artırılmıştır.

Hız: ~10 Mbps (Teorik) | **Gecikme:** ~100 ms | **Frekans Bantları:** 850 MHz, 2100 MHz

Senaryo: Normal Kullanım

Senaryo Bilgisi: Normal kullanım senaryosu seçili.

Hız (Mbps): 10

Gecikme (ms): 100

Paket Kaybı (%): 0

Gecikme: 100ms

Paket Kaybı: 0%

Figure 1. General view of the 3G HSPA simulation interface. The Base Station (NodeB) is on the left, the User Equipment (UE) is on the right, and data packets are transmitted over a single carrier.

The movement of these packets, defined by circle elements in the code structure, along a single line drawn with the path tag, visually concretizes HSPA’s logic of continuous data flowing over a time-shared but single carrier. At the behavioral center of the simulation lies the simulate3G() function defined in JavaScript. This function utilizes a timer structure to update the positions of data packets moving across the screen at specific intervals. The packet transmission speed depends directly on the data rate (Mbps) selected by the user. In the code, this relationship is modeled by a simple mathematical correlation in which the animation speed is directly multiplied by a coefficient of 0.1. Through this method, an increase in the speed value increases the flow speed of data packets on the screen; thus, the relationship between HSPA’s channel capacity and packet transmission speed becomes visible at an intuitive level.

Critical network parameters, such as latency and packet loss, are directly incorporated into the simulation's dynamic structure. The scenario management function in the code automatically reconfigures speed, latency, and packet loss for different use cases, such as "Normal," "Heavy Traffic," "Rural," "Office," and "Stadium." This approach is highly compatible with the performance characteristics of 3G networks, which are sensitive to environmental conditions and user density. For instance, halving the speed in the "Heavy Traffic" scenario while increasing latency and packet loss rates successfully mimics the scheduling latency and limited radio resources experienced by HSPA under cellular load. One of the most notable components of the simulation, the packet loss mechanism, ensures that data packets become invisible with a certain probability via the applyPacketLoss() function. This process represents block errors, signal fading, or instabilities observed at cell edges in real networks. Finally, the frequency bands used by 3G technology (850 MHz and 2100 MHz) are graphically displayed on the interface to provide frequency awareness to the student. Consequently, this structure allows students to observe the limitations of 3G and establishes a strong foundation for understanding the advantages of the LTE-Advanced technology.

4. Implementation of the 4.5G LTE-A Simulation and Carrier Aggregation Modeling

LTE-Advanced (LTE-A), also known as 4.5G, is a sophisticated radio access architecture designed to overcome the limitations of the 3G HSPA architecture, specifically to dramatically increase data rates. One of the most critical innovations in LTE-A is Carrier Aggregation (CA), which enables the simultaneous use of multiple frequency bands. In the simulation, this structure is represented by a visual model in which three different carriers transmit data concurrently; this arrangement offers a pedagogical abstraction consistent with both the theoretical foundations of LTE-A technology and real-world network behavior (Holma & Toskala, 2011; Ghosh et al., 2012).

On the simulation screen, three separate transmission lines are drawn between the base station and the mobile device. These represent the 800 MHz, 1800 MHz, and 2600 MHz bands, respectively, and are defined by three distinct path objects within the code structure. Each channel possesses its own unique color and positioning structure. As shown in the general interface view in Figure 2, data packets initially move independently through these three channels and, upon reaching the junction point, are consolidated into a single stream on the main carrier. This structure closely resembles the actual physical-layer operation of the CA process in LTE-A. In a real network, each carrier transports data in its own resource blocks; aggregation occurs at the MAC layer. In the simulation, this aggregation process is visualized by a merging line located in the center of the screen.

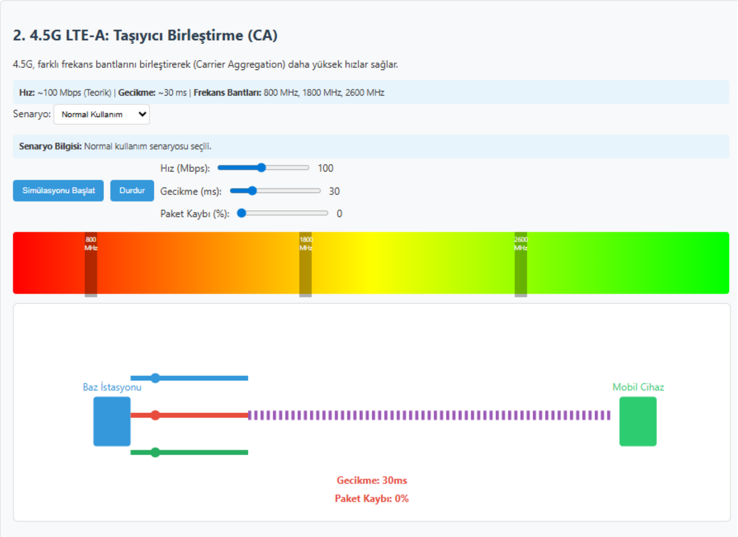
When the JavaScript code is examined, it is seen that the 4.5G simulation is executed by the simulate45G() function, which updates the positions of three independent data packets at regular intervals. The initial positioning of the packets at different vertical coordinates is used to distinguish the three carriers from one another on the scene. In the code, the movement speed of each packet is calculated relative to the user-selected total speed parameter using the formula speed = Mbps × 0.15. In this way, a behavior is observed in which the data flow across all carriers increases faster as the speed increases. While the capacity of each carrier may differ in real LTE-A architecture, the simulation opts for an instructional abstraction, ensuring that all airlines are affected simultaneously by the speed increase.

One of the significant components of the 4.5G simulation is scenario management. Through the updateScenario('45g') function in the code, parameters for speed, latency, and packet loss are readjusted for various conditions such as normal usage, heavy urban traffic, office environments, and stadiums. The variation in these parameter values models the unbalanced capacity behavior of LTE-A under traffic load at a conceptual level. For example, the drop in speed to 30 Mbps and the increase in latency to 80 ms in the stadium scenario correlate with the exhaustion of the base station's carrier aggregation capacity due to excessive user density in the real world. Similarly, the increase in speed and decrease in latency in the office scenario reflect advantageous conditions of the cell, such as low mobility and high signal quality.

The simulation's visual layout also intuitively explains the CA mechanism's logical steps. The fact that packets move independently across three different carriers for a temporary period and then aggregate into a single flow represents the aggregation of radio resources from multiple airlines for presentation to the user. In reality, this aggregation scheduling is performed by the eNodeB; the merging line in the simulation is a simplified metaphor created for this process. In this respect, the simulation explains MAC layer timing and the aggregation of resource blocks—critical building blocks of LTE-A—through an instructive illustration.

In the 4.5G animation, latency and packet loss indicators are also dynamically updated. The gecikme-gosterge45g and paket-kayip-gosterge45g tags within the code are updated with new values when the scenario changes; this allows the student to visually track the performance variations of LTE-A under different traffic conditions. Furthermore, the packet-loss function defined for 4.5G temporarily hides some packets to visualize errors originating from the radio, just as in 3G. This conceptual structure shows that packet loss in CA systems can affect not only the carrier level but also the aggregated channel level.

Consequently, the 4.5G simulation effectively presents the fundamental principles of LTE-A technology by combining representations of carrier aggregation, multi-channel flow, scenario-based parameter management, and packet behavior at the code level. The convergence of three independent carriers into a single data stream, the simultaneous effect of the speed parameter across all airlines, and the direct consequences of different scenarios on network performance exhibit a structure aligned with the capacity-enhancing elements of the LTE-A architecture described in the academic literature. Thus, the simulation concretizes the significant improvements offered by 4.5G over 3G from both theoretical and practical perspectives, laying the conceptual foundation for the transition to 5G network slicing logic.


2. 4.5G LTE-A: Taşıyıcı Birleştirme (CA)

4.5G, farklı frekans bantlarını birleştirerek (Carrier Aggregation) daha yüksek hızlar sağlar.

Hız: ~100 Mbps (Teorik) | **Gecikme:** ~30 ms | **Frekans Bantları:** 800 MHz, 1800 MHz, 2600 MHz

Senaryo: Normal Kullanım

Senaryo Bilgisi: Normal kullanım senaryosu seçili.

Hız (Mbps): 100

Gecikme (ms): 30

Paket Kaybı (%): 0

Gecikme: 30ms

Paket Kaybı: 0%

Figure 2. General View of the 4.5G Simulation Interface

The simulation interface effectively conveys the carrier aggregation process in LTE-Advanced. Independent transmission lines for three different frequency bands are shown on the screen between the base station and the user device; data packets travel in separate channels along these lines. The carriers' positioning demonstrates, at a conceptual level, that each band serves as an individual physical carrier. As shown in Figure 2, when the packets reach the junction point, they are combined into a single data stream on the main channel. This structure symbolically reflects the total capacity increase resulting from aggregating multiple carriers at the MAC layer in the real LTE-A architecture.

This visual representation of the simulation helps understand how carrier aggregation operates, both theoretically and practically, and clearly reveals the multi-layered data transmission advantage offered by 4.5G compared to 3G.

5. Simulation of 5G Network Slicing: Architecture, Visualization, and Performance Modeling

In contemporary telecommunications literature, network slicing is considered a critical feature that enables 5G architecture to serve as a flexible platform, providing performance enhancements and dynamically adapting to heterogeneous service requirements. Therefore, during simulation design, both the technical and functional components of network slicing were considered, and the three fundamental service types—uRLLC, eMBB, and mMTC—were logically segregated at the interface (NGMN Alliance, 2016; Foukas et al., 2017).

The simulation interface is structured around a centrally positioned "5G Core" component, as presented in Figure 3. Three distinct network slices emerge from this core structure, each modeled over an independent transmission line representing its own data flow. This visual arrangement represents a logical abstraction of network slicing in the 5G architecture;. However, the physical infrastructure remains the same, isolation between slices is maintained, and Quality of Service (QoS) parameters are differentiated by service type. This segregation is also clearly evident in the code structure: parameters such as speed, latency, and packet loss are defined as independent variables for each slice and can be dynamically updated via user interaction.

The uRLLC (Ultra-Reliable Low-Latency Communication) slice supports applications that require extremely low latency, such as industrial automation, autonomous vehicles, and critical infrastructure control. In the simulation, data packets belonging to this slice are made visible by keeping the speed multiplier high and the latency at a minimum. In the code, this behavior is achieved by running timer functions with higher refresh rates and shorter delay intervals. Consequently, it is notable that the data packets on the screen move faster and more fluidly than those in other slices. This design approach offers a powerful visual metaphor representing the 1 ms latency norm that 5G offers for uRLLC (3GPP, 2017).

The eMBB (Enhanced Mobile Broadband) slice supports use cases that require high bandwidth, such as video streaming, VR/AR applications, and high-speed data transfer. In this slice, the simulation creates a profile that closely matches real-world network behavior by defining higher speeds than uRLLC but with relatively higher latency. Data packets in the eMBB slice move according to the mid-level speed multiplier and fixed latency parameters specified in the code. Thus, a model is obtained that provides a continuous and robust data flow, albeit not as fast as uRLLC (Andrews et al., 2014).

The mMTC (Massive Machine Type Communication) slice symbolizes the 5G support infrastructure for the widest-scale IoT deployments. Parameters such as low speed, low bandwidth, and massive device density come to the fore in this slice. In the simulation, the slow movement of mMTC packets and their operation with a lower speed multiplier, which represents a small bandwidth, reflect the characteristics of real IoT networks. This approach aligns with design principles that prioritize energy efficiency and connection continuity over low latency for large sensor networks (Ge et al., 2016).

The use of a JavaScript-based animation engine enables simultaneous observation of behavioral differences in the network slicing structure. setInterval-based timing functions update the data packet's position for each slice, and user-defined parameter changes are instantly reflected in the simulation. Thanks to this dynamic structure, when the user modifies values such as latency, speed, and packet loss, the simulation reacts in real time, recalculating the speed or visibility of the data flow in the relevant slice immediately. In this regard, the study presents an interactive experience that models the adaptive control logic of the 5G architecture.

3. 5G: Ağ Dilimleme (Network Slicing)

5G, tek bir altyapıyı farklı ihtiyaçlara göre sanal dilimlere böler (uRLLC, eMBB, mMTC).

Hız: ~1 Gbps (Teorik) | **Gecikme:** ~1 ms (uRLLC) | **Frekans Bantları:** 700 MHz, 3.5 GHz, 26 GHz (mmWave)

Senaryo: Normal Kullanım

Senaryo Bilgisi: Normal kullanım senaryosu seçili.

uRLLC Hızı (Mbps): 500
eMBB Hızı (Mbps): 300
mMTC Hızı (Mbps): 10
uRLLC Gecikmesi (ms): 1
Paket Kaybı (%): 0

Gecikme: 1ms
Gecikme: 10ms
Paket Kaybı: 0%

Figure 3. General View of the 5G Network Slicing Simulation Interface

Another significant component of the interface design shown in Figure 3 is the visual labels belonging to each slice. The fact that the explanatory texts for uRLLC, eMBB, and mMTC are fixed in distinct locations, the use of line drawings supported by color differences, and the indication of data packets with different colors in each slice not only render the information architecture clear but also contribute to the easier cognitive comprehension of 5G's multi-layered architectural structure. This approach, referred to as pedagogical visualization in the literature, is a design strategy that enhances the educational value of the simulation (de Jong, Linn, & Zacharia, 2013).

Consequently, the 5G network slicing simulation transcends mere animation or visual display, offering a comprehensive model that associates 5G's differentiated service-type structure with functional parameters. The modular structure of the code enables each slice to be treated as an independent system component; together with user-interactive control elements, the simulation has transformed the performance diversity offered by the 5G architecture into an understandable experience. In this respect, the study is a strong example of both technical accuracy and instructional clarity.

6. Map-Based Base Station Deployment and Radio Coverage Modeling

Unlike the 3G, 4.5G, and 5G simulations examined in previous sections, this module models not only intra-network packet flows but also variables related to the Physical Layer (PHY), such as geographical location, signal propagation, and coverage area. Consequently, this section focuses on the radio coverage dimension of mobile communication networks, enabling a more holistic evaluation of the simulation.

The simulation interface is built on an interactive map created with the Leaflet.js library. As shown in the interface in Figure 6.1, the user can position a base station on the map and subsequently specify fundamental parameters such as the frequency band, transmitter power, receiver sensitivity, antenna height, and environment type (urban, suburban, or rural). These variables have a decisive effect on the propagation characteristics of radio signals, and the simulation calculates this impact via a mathematical model (Leaflet Team, 2025).

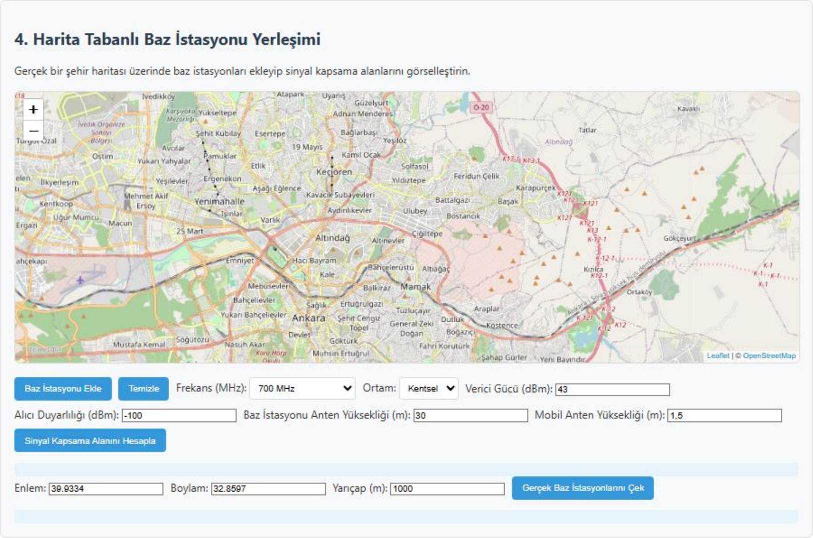
The Okumura-Hata model, one of the most frequently cited empirical models in the literature for macrocell propagation, is used for coverage area calculation. In the code, this model is implemented to include separate coefficients for distance (km), frequency (MHz), transmitter antenna height, mobile antenna height, and environment type. The simulation calculates the maximum allowable path loss based on the selected parameters, then uses a binary search to find the distance at which this value is achieved. Thus, the user obtains not merely a theoretical value but a realistic coverage radius associated with the selected parameters (Hata, 1980; Okumura et al., 1968).

This radius is displayed as a circle layer on the Leaflet map and visually conveys the spatial boundaries of the coverage area, as shown in Figure 4. This visual approach of the simulation intuitively demonstrates to the user the dramatic effect of frequency change on coverage (e.g., the difference between the wide-area coverage of the 700 MHz band and the extremely limited range of 26 GHz mmWave). The fact that the coverage area is wider at low frequencies, while it remains limited at high frequencies, indicates that the physical reality based on radio wave propagation principles has been modeled accurately.

Another notable aspect of the simulation is the capability to retrieve real base station data. This function displays real cells (2G/3G/4G/5G) at specific coordinates on the map via the OpenCellID API. As shown in the sample outputs in Figure 4, base stations retrieved from the API are marked on the map, allowing the user to compare the coverage areas generated by theoretical modeling with the base-station density in the real world. This approach endows the simulation with usability not only for instructional purposes but also for research (OpenCellID, 2025).

Upon examining the code's general architecture, it is observed that the map-based module is highly modular. A separate function performs the addition of base stations; helper functions are defined for clearing coverage areas; and signal calculations have been transformed into a completely independent model function. This structure enhances code readability and provides a flexible foundation for future simulation expansion. For instance, the user can clear all coverage areas with a single button; similarly, new calculations can be performed easily after parameter changes.

Consequently, the map-based base station placement module effectively models the physical variables required for coverage planning in mobile networks. It possesses a dynamic structure that integrates this modeling with user interaction. The simulation's accurate implementation of radio propagation models, its sensitivity to environmental factors, and its support through geographical visualization make this module a strong component from both academic and pedagogical perspectives.



4. Harita Tabanlı Baz İstasyonu Yerleşimi

Gerçek bir şehir haritası üzerinde baz istasyonları ekleyip sinyal kapsama alanlarını görselleştirin.

Frekans (MHz): 700 MHz
Ortam: Kentsel
Verici Gücü (dBm): 43
Alıcı Duyarlılığı (dBm): -100
Baz İstasyonu Anten Yüksekliği (m): 30
Mobil Anten Yüksekliği (m): 1,5
Enlem: 39.9334
Boylam: 32.8597
Yarıçap (m): 1000

Figure 4. Map-Based Coverage Analysis Interface

7. Comparative Visualization of 3G, 4.5G, and 5G Performance Metrics

The speed-latency graph generated in this scope has enabled the presentation of key performance indicators for three generations of mobile technology on a common axis, thereby allowing the user to examine developmental differences between technologies within a comprehensive framework. The graph shown in Figure 5 visualizes the comparative values obtained in the simulation environment and presents the fundamental characteristics of mobile network evolution in a pedagogically understandable manner.

The Chart.js library was used to create the graphs. In this library, the datasets generated by the simulation were processed, and speed and latency values were plotted as two independent series. The simulation used speed values of approximately 10 Mbps for 3G, 100 Mbps for 4.5G, and 1000 Mbps for 5G, consistent with theoretical capacities reported in the literature. Similarly, latency was modeled in the simulation at 100 ms for 3G, 30 ms for 4.5G, and 1 ms for 5G. The graphical presentation of these values ensured that the performance disparities between technologies became distinct both quantitatively and visually (Chart.js Contributors, 2025).

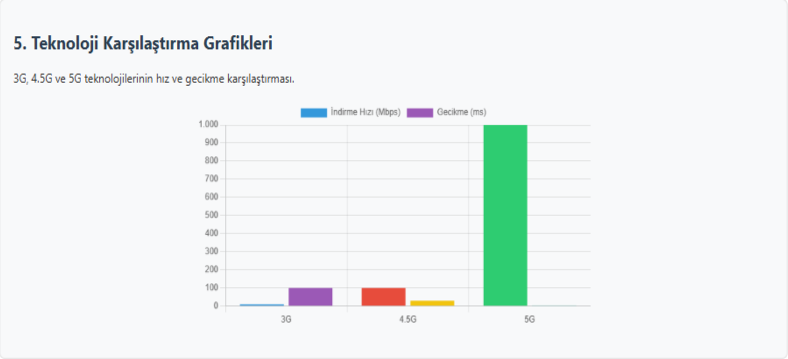
As shown in Figure 5, the simulation results indicate that 5G technology delivers a speed increase of approximately two orders of magnitude compared to 3G and an improvement of 1 order of magnitude compared to 4.5G. Simulation results for latency similarly showed that 5G offers far superior performance compared to both 3G and 4.5G, thanks to its ultra-low latency.

These results indicate that mobile networks have undergone a significant transformation during the evolutionary process, not only in terms of speed but also in end-to-end latency (Andrews et al., 2014).

The graphical modeling performed in the simulation also enabled inferences about the technologies' use cases. For instance, the high latency values of 3G showed that it is unsuitable for real-time applications; conversely, the 1 ms latency performance of 5G obtained in the simulation demonstrated that it offers an ideal infrastructure for applications requiring low latency, such as autonomous systems, industrial automation, and augmented reality. The mid-level performance of 4.5G, on the other hand, indicates that it provides a balanced solution for mobile broadband internet usage.

From a technical perspective, the graphical comparison in the simulation was implemented using Chart.js's categorical bar chart. By displaying two different performance metrics (speed and latency) on the same graph, users could perform a multidimensional performance evaluation. In the simulation, color separation, scaling, and axis definitions were intentionally kept simple to ensure the graphs remain readable across different screen sizes.

Consequently, the speed-latency graphs generated in the simulation offered a powerful tool for analyzing the performance characteristics of mobile network technologies. The visualization in Figure 5 enabled evaluation of the intergenerational development of mobile networks using quantitative data and significantly contributed to the study's holistic simulation structure.



5. Teknoloji Karşılaştırma Grafikleri

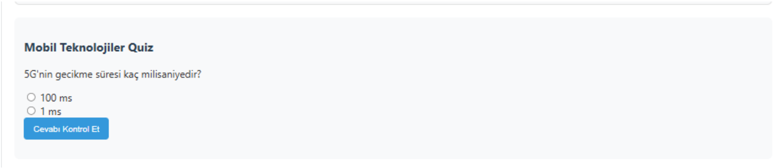
3G, 4.5G ve 5G teknolojilerinin hız ve gecikme karşılaştırması.

Figure 5. Comparative Analysis of Speed and Latency Across Generations

8. Interactive Quiz Module as A Formative Assessment Tool

In this section, the technical operation and pedagogical function of the quiz module included in the study are examined. Whereas previous sections of the simulation elucidated the technical aspects of mobile network technologies through animated visual models and coverage analyses, this module is designed as an interactive assessment tool to verify the extent to which the user has assimilated this technical knowledge. The quiz interface, presented in Figure 6, directs a multiple-choice question to the user and provides instant feedback based on the selected answer.

Although the quiz structure is technically simple, it reflects the formative assessment approach, which holds a significant place in interactive learning literature. The question directed to the user focuses on low latency, a characteristic feature of 5G technology. In the code, each option is defined as an HTML radio button, and when the user selects an option, a JavaScript-based control function (checkAnswer) is triggered. This function compares the specified value to the correct answer and displays "Correct" or "Incorrect" on the screen accordingly. This control mechanism implemented in the simulation provides the user with an instant indication of errors and reinforces the learning process.


Mobil Teknolojiler Quiz

5G'nin gecikme süresi kaç milisaniyedir?

○ 100 ms
○ 1 ms

Figure 6. View of the Interactive Quiz Interface

As seen in Figure 6, the quiz interface possesses a simple yet functional design. The question text is positioned at the top, options are listed vertically, and the entire structure is arranged to facilitate user interaction. This simplicity aligns with design principles for reducing cognitive load in digital learning environments that teach technical concepts. Furthermore, supporting the feedback text with color (green for correct, red for incorrect) is a significant example of sensory reinforcement in the learning sciences (de Jong, Linn, & Zacharia, 2013).

The role of the quiz module within the simulation is not limited merely to measuring knowledge; it also facilitates the active integration of learned information into the user's mental processes. The use of interactive questions transforms the user from a passive observer into an active subject of the learning process. This situation particularly facilitates understanding of the abstract structures of technical concepts and supports reinforcement of topics such as latency, bandwidth, and service classes, which often appear complex to students.

The control function utilized in the simulation represents a method frequently employed in modern web-based assessment systems. This control mechanism, which operates entirely on the client side without constantly sending data to the server, offers advantages in both performance and design simplicity. This same logic can be transformed into an expanded assessment module with a larger dataset; enhancements such as adding multiple questions, increasing difficulty levels, or presenting results via a scoring system can be easily implemented. In this respect, the quiz module can be considered one of the simulation's sections open to future expansion.

Consequently, the quiz module is important as a functional component that enhances the simulation's pedagogical value, supports learning assessment, and is implemented in accordance with interactive digital learning principles. The design presented in Figure 6 demonstrates a successful application example in both technical and educational terms, with its user-friendly structure and real-time feedback mechanism.

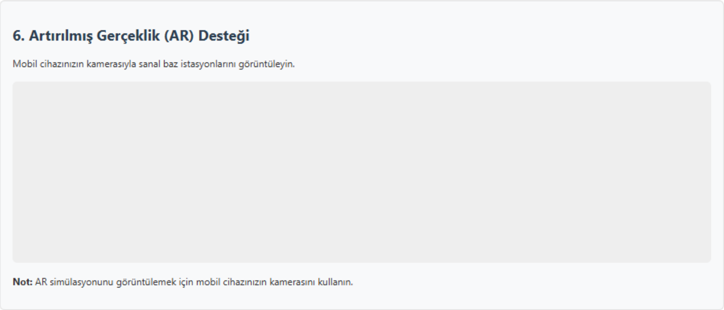
9. Augmented Reality Integration For Visualizing Cellular Infrastructure

In this section, the technical foundations, design approach, and functional value of the Augmented Reality (AR) module included in the simulation are examined in detail within the context of mobile communication technologies. AR technology is an interaction method that enables the placement of digital objects in the physical world, as perceived through a camera. AR applications, which have secured a significant place in the future of mobile networks, have become more functional with 5G, particularly given their need for high bandwidth and low latency. In this context, the AR module added to the simulation allows the user to place a simple base station model in the real environment using the mobile device's camera. An example use case of the AR interface is presented in Figure 7.

From a technical perspective, the module was developed based on the A-Frame and AR.js libraries. A-Frame is a framework that allows defining WebXR-compatible three-dimensional scenes using an HTML-like syntax. AR.js is a lightweight library that enables marker-based augmented reality applications to run within a browser. The AR scene used in the simulation is defined within an <a-scene> tag and is activated via a marker-based trigger. The <a-marker> component in the code activates the 3D objects to be placed in the scene when a specific square pattern is detected in the camera feed. This object is a cube model representing a base station in the simulation (Mozilla Developers, 2025; AR.js Team, 2025).

As shown in Figure 7, when the user directs the camera towards the marker, a red 3D box appears in the scene, with the label "Base Station" displayed above it. Unlike the other 3G/4.5G/5G models used in the simulation, this representation allows the virtual object to be experienced in interaction with the physical environment. Thus, the user can perceive the base station concept not only through two-dimensional graphics and schematic drawings but also in real space. This type of experience, referred to in the literature as contextualized visual learning, is evaluated as a method that reinforces conceptual learning (Billinghurst & Dünser, 2012).

Technically, the AR module is relatively independent of the code structure comprising the entire simulation; however, it serves a complementary function consistent with the application's general purpose. Thanks to AR.js's lightweight architecture, the module runs entirely on the client side and does not require additional server-side processing. This enables an augmented reality experience without performance loss, especially on mobile devices. At the same time, the module serves as a conceptual bridge to understand why the low latency features supported by 5G are critical for AR applications. Real-world AR use cases (e.g., remote maintenance, education, navigation, and industrial AR solutions) cannot operate efficiently without the speed and latency advantages offered by 5G. Therefore, the module is a valuable tool for concretizing the potential application areas of 5G technology.



6. Artırılmış Gerçeklik (AR) Desteği

Mobil cihazınızın kamerasıyla sanal baz istasyonlarını görüntüleyin.

Not: AR simülasyonunu görüntülemek için mobil cihazınızın kamerasını kullanın.

Figure 7. General View of the AR Module Interface

Furthermore, the AR component of the simulation features an architecture highly suitable for expansion. Enhancements such as integrating markerless AR technologies instead of marker-based models, rendering 3D base-station models more realistically, or visualizing base-station coverage areas in real time within the AR environment can be easily implemented. Such expansions will contribute to both the explanation of AR technology and the visualization of physical components in wireless communication infrastructures (Google Developers, 2025).

Consequently, the AR module increases the technical diversity of the simulation and offers a significant contribution to the tangible realization of mobile communication technologies. The interface shown in Figure 7 is a simple yet instructive example of an augmented reality experience. In this respect, the module is evaluated as both an instructional tool conceptualizing the use cases of 5G and a functional complementary element to the holistic simulation structure of the study.

10. Comparative Analysis of Energy Consumption Across Cellular Generations

In this section, the technical and conceptual foundations of the energy consumption comparison module included in the simulation are examined, and the differences in energy efficiency among 3G, 4.5G, and 5G mobile communication technologies are evaluated. Energy consumption is a critical performance indicator for both mobile operators and user devices, and it has emerged as a significant design goal, especially during the development of 5G. In the simulation, these differences are defined by fixed energy consumption values presented to the user; in the interface shown in Figure 8, the energy consumption per unit of data (Wh/GB) for each technology is presented comparatively.

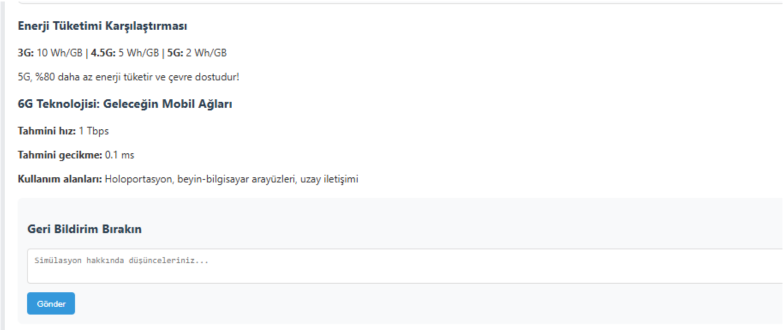
Values of 10 Wh/GB for 3G, 5 Wh/GB for 4.5G, and 2 Wh/GB for 5G were utilized in the simulation. These values are consistent with typical energy-efficiency estimates reported in the literature and reflect improvements in energy performance across generations. Presenting data in terms of fixed-unit consumption allows the user to make a direct, understandable comparison. As seen in Figure 8, 5G technology consumes significantly less energy than previous generations, indicating that 5G has an operational structure approximately 80% more efficient (Lopez et al., 2021).

The fundamental reason for this efficiency disparity is the comprehensive optimizations performed in both the radio access layer and the core network side of the 5G architecture. Structural innovations such as beamforming, massive MIMO, high spectral efficiency, adaptive modulation techniques, and energy-saving scheduling algorithms are among the primary elements reducing 5G's energy consumption. In contrast, while the 3G architecture provides wide coverage, it has higher energy requirements due to its low spectral efficiency and a control structure that requires continuous connection. 4.5G constitutes a balance point between these two technologies; it consumes less energy compared to 3G thanks to carrier aggregation, advanced coding techniques, and more modern radio access protocols (Wu et al., 2015; Chen et al., 2020).

This comparison, performed in the simulation, demonstrates that differences in energy consumption among technologies are not merely theoretical but produce measurable results in daily use. For example, in data-intensive applications such as high-resolution video streaming, a 3G connection rapidly depletes the device battery, whereas device operating times are prolonged due to 5G's lower energy consumption. Similarly, from the perspective of mobile operators, since energy costs are among the highest expenses in infrastructure operations, 5G's more efficient architecture results in significant reductions in operational expenses (OPEX).

When evaluated alongside other performance metrics throughout the simulation, the energy consumption comparison in Figure 8 shows that mobile network evolution progresses not only in terms of capacity and latency but also in energy efficiency. This situation indicates that energy efficiency is a crucial determinant in the design of future sustainable communication networks.

Consequently, the energy consumption comparison in the simulation highlights the significant efficiency advantages of 5G technology over previous generations. It underscores the central role of energy performance in the development of mobile communication infrastructure. The values presented in Figure 8 have contributed to the understanding of this transformation from both conceptual and technical perspectives.



Enerji Tüketimi Karşılaştırması

3G: 10 Wh/GB | **4.5G:** 5 Wh/GB | **5G:** 2 Wh/GB

5G, %80 daha az enerji tüketir ve çevre dostudur!

6G Teknolojisi: Geleceğin Mobil Ağları

Tahmini hız: 1 Tbps

Tahmini gecikme: 0.1 ms

Kullanım alanları: Holoportasyon, beyin-bilgisayar arayüzleri, uzay iletişimi

Geri Bildirim Bırakın

Figure 8. Energy Efficiency Comparison and 6G Vision and Performance Preview

11. Conceptual Outlook On 6G: Future Capabilities and Emerging Communıcation Paradigms

In this section, the conceptual framework of next-generation mobile communication technologies is detailed based on the 6G preview component included in the simulation. Although 6G is currently in the research and standardization phase, it is defined in the literature within a broad vision framework. It is expected to introduce revolutionary innovations in mobile communications. In the simulation, this vision is summarized through high-speed, ultra-low-latency, and next-generation use cases, as seen in the interface shown in Figure 8.

In the 6G component of the simulation, it is indicated that the technology's theoretical speed capacity will be 1 Tbps, and latency will be reduced to the order of 0.1 ms. These values represent an improvement of nearly one order of magnitude over the 1 Gbps speed and 1 ms latency offered by 5G. The simulation results align with predictive estimates from early-stage 6G studies in the literature. For instance, it is foreseen that terahertz (THz) bands will be utilized, photonics-based communication methods will be developed, and AI-supported network optimization will become a fundamental component. In this context, the simulation emphasizes that 6G will offer a significant leap not only in terms of speed and latency but also in architecture (Saad et al., 2020; Letaief et al., 2019).

In the 6G preview interface shown in Figure 8, scenarios beyond current technologies, such as holoportation, Brain-Computer Interfaces (BCI), and space-based communication systems, are included among the technology's potential application areas. Since holoportation requires the real-time transmission of high-resolution 3D images, it is considered one of the most typical use cases of 6G due to its broadband and ultra-low latency requirements. Similarly, BCI applications require ultra-reliable communication infrastructure as they involve complex processes such as the simultaneous processing and wireless transmission of neural signals. Space communication networks integrate LEO (Low Earth Orbit) satellite constellations with 6G networks and are seen as a crucial component of 6G, in line with the goal of global coverage (Rappaport et al., 2019; Giordani et al., 2020).

This conceptual presentation, developed in the simulation, is designed to demonstrate to the user that 6G is not merely an advanced mobile communication standard but a comprehensive technology that will form the infrastructure of the future digital ecosystem. Therefore, the forms of expression used in the simulation support the idea that 6G will create a technological leap of a kind not seen in previous generations. This approach is consistent with conceptualizations frequently mentioned in the literature, such as 6G = AI-native network" and "6G = immersive communication fabric (Zhang et al., 2019).

A significant aspect of the 6G preview included in the simulation is that it does not include a technical animation or calculation, unlike the current-generation simulation modules. Instead, this component lays the groundwork for the future and aims to raise the user's awareness of the next stage of network evolution. From an academic perspective, this design choice is deemed more appropriate for presenting a pedagogical foresight model, since 6G does not yet have concrete implementations.

Nevertheless, this predictive structure used in the simulation is highly open to expansion. For example, coverage models for THz frequency bands can be added, AI-based network optimization can be associated with the simulation, or 6G's satellite-ground integration can be integrated with the map-based module. Such enhancements will enable the creation of a comprehensive 6G simulation platform for both research and education.

Consequently, the 6G preview module offers a conceptual perspective on next-generation mobile networks within the simulation's overall structure and provides the user with a fundamental framework for how mobile communication technologies may evolve. The predictive indicators presented in Figure 8demonstrate that 6G will redefine the boundaries of mobile communication in terms of speed, latency, and use cases. In this respect, the 6G module can be considered a significant component that embodies the study's future-oriented vision (Latva-aho & Leppänen, 2019).

12. Conclusion and Overall Evaluation Of The Simulation Framework

The web-based mobile network simulation developed within the scope of this study aimed to present the fundamental working principles of 3G, 4.5G, and 5G technologies in a visually, interactively, and pedagogically rich structure. When all components of the simulation are evaluated together, it has been demonstrated that modern mobile communication technologies can be understood not only through theoretical concepts but also through interactive models that the user can experience. In this respect, the study qualifies as both an instructional tool for educational environments and a demonstration platform for various engineering analyses.

The simulation modules, designed separately for 3G, 4.5G, and 5G, have been structured to represent the unique architecture and performance characteristics of each generation. The 3G HSPA module successfully modeled the single-carrier and continuous data flow structure with animated data packets; the 4.5G LTE-A module modeled the Carrier Aggregation (CA) mechanism with multiple frequency lines; and the 5G module modeled the Network Slicing approach through three independent service classes. This layered approach clearly shows that the evolution of mobile communication technologies is shaped not only by increases in speed but also by architectural and functional innovations.

The map-based base station placement section of the study integrated coverage modeling, representing the physical layer of mobile networks, with an interactive geographical interface. Coverage areas calculated using the Okumura-Hata model realistically reflected the effects of variables selected by the user—such as frequency, antenna height, transmitter power, and environment type—on radio propagation. Thus, the study contributed to the applied understanding of fundamental concepts related to radio frequency engineering. Additionally, the capability to transfer real base station data to the map via the OpenCellID API added both a practical and research-oriented dimension to the simulation.

The simulation's pedagogical components are also noteworthy. The graphical comparison module clearly showed the performance differences across generations of mobile technologies along the axes of speed and latency; the quiz module allowed the user to test the learned information. The Augmented Reality (AR) component presented an innovative visualization method that combines digital objects with the physical environment, allowing mobile network components to be experienced in a spatial context. When all these components are evaluated, it becomes clear that the simulation has not only technical but also educational value.

The future-oriented dimension of the study was established with the 6G preview module. This module provided foresight into 6G's expected performance targets, use cases, and conceptual features, raising users' awareness of the direction mobile communication technologies may take. In this respect, the study takes a broad perspective that encourages discussion not only of simulating existing technologies but also of next-generation communication paradigms.

In general evaluation, the developed simulation system has succeeded in presenting the technical components of mobile communication technologies in an understandable, interactive, and visually rich manner. Thanks to its modular structure, extensible code architecture, and multi-layered instructional approach, the study has the potential to be a strong resource for both academic research and engineering education. In future studies, adding more detailed integration of quality of service metrics, mmWave coverage analyses, 6G THz channel modeling, or AI-supported network optimization algorithms to the model will significantly deepen and broaden the simulation.

References

Andrews, J. G., et al. (2014). What will 5G be? IEEE Journal on Selected Areas in Communications, 32(6), 1065–1082.

AR.js Team. (2025). Augmented reality for the web. https://ar-js-org.github.io/AR.js-Docs/

Billinghurst, M., & Dünser, A. (2012). Augmented reality in the classroom. Computer, 45(7), 56–63.

Boccardi, F., Heath, R. W., Lozano, A., Marzetta, T. L., & Popovski, P. (2014). Five disruptive technology directions for 5G. IEEE Communications Magazine, 52(2), 74–80.

Chart.js Contributors. (2025). Chart.js: Simple HTML5 charts using the tag. https://www.chartjs.org/

Chen, S., Hu, Y., Hou, Y., Shi, Y., & Ge, X. (2020). Distributed scheduling for energy-efficient massive MIMO systems. IEEE Transactions on Communications, 68(9), 5492–5504.

Cisco. (2020). Cisco annual internet report (2018–2023) white paper. Cisco Systems.

Dahlman, E., Parkvall, S., & Skold, J. (2020). 5G NR: The next generation wireless access (2nd ed.). Academic Press.

de Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and virtual laboratories in science and engineering education. Science, 340(6130), 305–308.

Foukas, X., Patounas, G., Elmokashfi, A., & Marina, M. K. (2017). Network slicing in 5G: Survey and challenges. IEEE Communications Magazine, 55(5), 94–100.

Ge, X., Tu, S., Mao, G., Wang, C.-X., & Han, T. (2016). 5G ultra-dense cellular networks. IEEE Wireless Communications, 23(1), 72–79.

Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., & Zorzi, M. (2020). Toward 6G networks: Use cases and technologies. IEEE Communications Magazine, 58(3), 55–61.

Ghosh, A., Mangalvedhe, N., Ratasuk, R., Mondal, B., & Cudak, M. (2012). Heterogeneous cellular networks: From theory to practice. IEEE Communications Magazine, 50(6), 54–64.

Google Developers. (2025). ARCore: Augmented reality SDK. https://developers.google.com/ar

Hata, M. (1980). Empirical formula for propagation loss in land mobile radio services. IEEE Transactions on Vehicular Technology, 29(3), 317–325.

Holma, H., & Toskala, A. (2011). LTE for UMTS: Evolution to LTE-Advanced (2nd ed.). Wiley.

International Telecommunication Union. (2023). Mobile network traffic evolution and forecasts. ITU-R Reports.

Latva-aho, M., & Leppänen, K. (2019). Key drivers and research challenges for 6G ubiquitous wireless intelligence. 6G Flagship, University of Oulu.

Leaflet Team. (2025). Leaflet: An open-source JavaScript library for mobile-friendly interactive maps. https://leafletjs.com/

Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y. A. (2019). The roadmap to 6G: AI empowered wireless networks. IEEE Communications Magazine, 57(8), 84–90.

Lopez, C. L. G. P. L., et al. (2021). Energy efficiency in 5G networks: A survey. IEEE Access, 9, 13254–13278.

Mozilla Developers. (2025). A-Frame: A web framework for building virtual reality experiences. https://aframe.io/

NGMN Alliance. (2016). Description of network slicing concept (Version 1.0).

Okumura, Y., Ohmori, E., Kawano, T., & Fukuda, K. (1968). Field strength and its variability in VHF and UHF land-mobile radio service. Review of the Electrical Communication Laboratory, 16(9–10), 825–873.

OpenCellID. (2025). The world's largest open database of cell towers. https://www.opencellid.org/

Rappaport, T. S. (2002). Wireless communications: Principles and practice (2nd ed.). Prentice Hall.

Rappaport, T. S., et al. (2019). Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond. IEEE Access, 7, 78729–78757.

Saad, W., Bennis, M., & Chen, M. (2020). A vision of 6G wireless systems: Applications, trends, technologies, and open research problems. IEEE Network, 34(3), 134–142.

3GPP. (2017). Study on scenarios and requirements for next generation access technologies (TR 38.913, Release 14).

3GPP. (2020). System architecture for the 5G system (5GS) (TS 23.501, Release 16).

World Wide Web Consortium. (2020). WebXR device API. https://www.w3.org/TR/webxr/

Wu, J., Zhang, Y., Zukerman, M., & Yung, E. K. N. (2015). Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey. IEEE Communications Surveys & Tutorials, 17(2), 803–826.

Zhang, Z., et al. (2019). 6G wireless networks: Vision, requirements, architecture, and key technologies. IEEE Vehicular Technology Magazine, 14(3), 28–41.

Chapter 11

ĸ-Fold Cross Validation Approach in Image-Based CNN Studies

Olcay PALTA[1], **Seda Yetkin YEŞİL**[2]

INTRODUCTION

Image-based data analysis has undergone a significant transformation in recent years with the rapid advancements in artificial intelligence and deep learning technologies. In particular, convolutional neural networks (CNNs) have enabled models to learn meaningful and hierarchical representations directly from images, eliminating the need for handcrafted feature extraction. As a result, CNN-based approaches have become widely adopted in both academic research and real-world applications.

Despite their remarkable performance, high accuracy values alone are not sufficient to fully demonstrate the effectiveness of a model. Image datasets are often limited in size due to the high cost and complexity of data acquisition processes, which raises concerns about the generalizability of the reported results. Performance metrics obtained from a single train-test split may therefore fail to reflect the true behavior of the model when exposed to unseen data. Similar to dynamic performance variations observed in engineering and energy systems, relying on single average measurements may lead to misleading conclusions, as system behavior can significantly change under different temporal and environmental conditions[1], [2], [3].

To address these limitations, cross-validation techniques provide a more comprehensive and reliable evaluation framework. Among these techniques, K-Fold cross-validation systematically incorporates the entire dataset into the evaluation process by dividing it into multiple subsets. This approach allows CNN-based image analysis studies to achieve more robust, consistent, and trustworthy performance assessments across different data partitions.

---

[1] Bitlis Eren University, Vocational School of Technical Sciences, Bitlis, Turkey, ORCID: (0000-0002-7496-2101), paltaolcay@gmail.com

[2] Bitlis Eren University, Vocational School of Technical Sciences, Bitlis, Turkey, ORCID: (0000-0001-9685-1376), sedayetkinyesil@gmail.com

In machine learning, the model development process is not limited to the training phase alone. Equally important is the evaluation of how well the developed model generalizes to unseen data. In this context, generalization ability is considered a key indicator of model performance. Cross-validation methods provide a statistical evaluation framework that enables systematic analysis of a model’s behavior across different subsets of the dataset.

1. THEORETICAL BACKGROUND OF CROSS-VALIDATION

In image-based machine learning studies, datasets often exhibit non-homogeneous distributions. While some classes may contain visually distinctive features, others may present more ambiguous boundaries. As a result, single and randomly performed train-test splits may lead to insufficient representation of certain classes during the testing phase. Consequently, the reported performance metrics may not accurately reflect the true behavior of the model.

Cross-validation techniques aim to mitigate these limitations by incorporating multiple data partitions into the evaluation process. By repeatedly dividing the dataset into different training and testing subsets, the model’s response to the overall data structure can be examined more comprehensively. Among these approaches, the K-Fold cross-validation technique enables more reliable and robust performance estimation by alternately using each subset for validation. The implementation of the five-fold cross-validation procedure is schematically illustrated in Figure 1.[4], [5], [6], [7], [8]

	Test Set	Train Set			
Original Dataset	1	2	3	4	5
Fold 1	Test	2	3	4	5
Fold 2	1	Test	3	4	5
Fold 3	1	2	Test	4	5
Fold 4	1	2	3	Test	5
Fold 5	1	2	3	4	Test

Figure 1. Implementation of the 5-Fold Cross Validation Technique

2. MECHANISM OF K-FOLD CROSS-VALIDATION

The K-Fold cross-validation mechanism is designed to ensure that the entire dataset contributes systematically to the performance evaluation of a learning model. In this framework, the available data are partitioned into K subsets of approximately equal size, commonly referred to as folds. Rather than relying on a single split, the method iteratively assigns one-fold as the validation set while the remaining folds are used for model training [9], [10].

This iterative procedure continues until each fold has served exactly once as the validation subset. As a result, every sample in the dataset is utilized both for training and for validation at different stages of the process. Performance metrics such as accuracy, precision, recall, or loss are computed independently for each iteration, providing multiple estimates of model behavior under varying data configurations. The final model performance is then obtained by aggregating these estimates, typically through averaging.

One of the key strengths of the K-Fold mechanism lies in its ability to reduce the bias associated with arbitrary data partitioning. In image-based learning tasks, certain samples may unintentionally favor the training or testing phase when a single split is used [11]. By rotating the validation fold across iterations, K-Fold cross-validation alleviates the influence of such sample-specific effects and yields a more balanced evaluation.

In the context of convolutional neural networks, where stochastic initialization and optimization dynamics can lead to variability in training outcomes, the K-Fold mechanism provides an additional layer of robustness. Consistent performance across multiple folds indicates that the learned representations capture intrinsic patterns of the data rather than fold-specific characteristics, thereby supporting stronger conclusions regarding the model's generalization capability [12], [13].

3. RATIONALE FOR USING K-FOLD IN CNN-BASED IMAGE PROJECTS

Convolutional neural networks are complex architectures characterized by a large number of trainable parameters. While this structural richness enables CNNs to learn highly discriminative representations from images, it also increases their susceptibility to overfitting. This issue becomes particularly pronounced when working with image datasets of limited size, where the model may unintentionally memorize training samples instead of learning generalizable patterns. In image-based studies, this behavior is often reflected by very high training accuracy accompanied by seemingly acceptable test performance. However, when the trained model is exposed to data originating from a different distribution or unseen scenarios, a noticeable degradation in performance may occur. For this reason, relying solely on results obtained from a single train-test split is generally considered insufficient for a robust evaluation of CNN-based models. K-Fold cross-validation provides an effective framework for analyzing the generalization capability of convolutional neural networks [11], [14], [15].

By evaluating the model across multiple folds, it becomes possible to observe whether the learned representations are overly dependent on specific subsets of the data. Consistent performance across different folds indicates that the network captures meaningful and transferable features rather than dataset-specific artifacts.

This evaluation strategy is especially valuable in application domains such as medical imaging, microscopic cell classification, and biomedical image analysis, where data acquisition is often expensive, time-consuming, or constrained by ethical considerations. In such contexts, K-Fold cross-validation serves as a fundamental methodological tool to enhance the reliability and credibility of experimental findings. Similar reliability concerns have been reported in complex engineering infrastructures such as power grids, where single-layer or static evaluation strategies were shown to be insufficient for ensuring system stability and robustness, leading to the adoption of multi-stage and feedback-based evaluation frameworks[16], [17]

4. PRINCIPLES OF FOLD CONSTRUCTION IN IMAGE DATASETS

In image-based datasets, the process of fold construction requires a more deliberate strategy compared to classical tabular data. Visual similarities between samples, acquisition conditions, and intra-class variability can lead to biased or unbalanced partitions when folds are created through naive random splitting.Preserving class distribution across folds is one of the most critical considerations in this process. In multi-class image classification tasks, certain classes often contain significantly fewer samples. If these minority classes are underrepresented in specific folds, the resulting performance metrics may become misleading, either overestimating or underestimating the true capability of the model [18].

Another important challenge arises from the risk of data leakage. Images captured under similar conditions or originating from the same source may share latent visual patterns. Distributing such correlated samples across both training and testing folds can artificially inflate model performance. This issue is particularly pronounced in medical and biological imaging studies, where images belonging to the same subject or specimen must not appear in multiple folds simultaneously.

Therefore, fold generation in image-based learning tasks should not rely solely on automated procedures. Instead, it must incorporate domain-specific knowledge alongside algorithmic strategies to ensure a fair, unbiased, and methodologically sound evaluation framework.

5. INTERACTION BETWEEN DATA AUGMENTATION, CNNs, AND K-FOLD

Data augmentation is a widely adopted strategy in image-based CNN[19], [20] studies to enhance model generalization, particularly when dealing with limited datasets. Operations such as rotation, flipping, scaling, and brightness adjustment introduce controlled variability into the training data, enabling the model to learn more robust and invariant feature representations.

When data augmentation is combined with K-Fold cross-validation, however, the process must be carefully designed. If augmented versions of training images inadvertently appear in the validation folds, the model may be evaluated on data that are not truly unseen. This phenomenon leads to information leakage and can result in overly optimistic performance estimates.

To prevent such bias, data augmentation should be strictly confined to the training folds within each iteration. Validation folds must remain untouched by any augmentation procedures. Clearly defining and enforcing this separation is essential to ensure the scientific validity and reliability of K-Fold evaluation results in CNN-based image analysis.

6. COMPUTATIONAL COST AND TRAINING TIME IMPACTS

One of the primary limitations of applying K-Fold cross-validation in CNN-based image analysis lies in the increased computational cost. Convolutional neural networks inherently require substantial processing power due to their deep architectures and large number of trainable parameters. When the training procedure is repeated K times within a cross-validation framework, the total computational burden increases proportionally. This challenge becomes more pronounced in studies involving deep network architectures and high-resolution images, where training a single fold may already take several hours or even days. As a result, researchers must carefully consider hardware resources when designing their experimental setup. In environments with limited computational capacity, compromises are often made by reducing the number of folds or opting for more lightweight model architectures.

Despite these constraints, the methodological advantages offered by K-Fold cross-validation often outweigh its computational drawbacks, particularly in academic research. In doctoral studies and comprehensive experimental investigations, the reliability and reproducibility of results are generally prioritized over training time, making K-Fold cross-validation a preferred evaluation strategy [4], [5].

7. HARDWARE INFRASTRUCTURE AND PARALLEL COMPUTING

The practical applicability of K-Fold cross-validation in CNN-based image analysis is strongly influenced by the available hardware infrastructure. In particular, graphics processing units (GPUs) play a crucial role in accelerating convolutional neural network training due to their inherent parallel computation capabilities. The widespread adoption of GPU-accelerated systems has significantly reduced training times, thereby making computationally demanding evaluation strategies such as K-Fold cross-validation more feasible in practice.

Modern deep learning frameworks provide optimized implementations that efficiently utilize GPU resources, enabling faster forward and backward propagation across multiple network layers. These optimizations are especially beneficial in fold-based training scenarios, where the same model architecture must be trained repeatedly on different data partitions. However, since each fold represents an independent training process, careful management of GPU memory and computational resources is required. Issues such as memory fragmentation, batch size limitations, and concurrent process handling must be addressed to ensure stable and efficient training.

In advanced experimental setups, parallel K-Fold strategies have been proposed, where different folds are trained simultaneously on separate GPUs or computing nodes. This approach substantially reduces the overall experimentation time and allows large-scale cross-validation studies to be conducted within reasonable timeframes. Nevertheless, such parallelization increases the demand for high-performance hardware and sophisticated resource scheduling mechanisms. As a result, the choice between sequential and parallel K-Fold implementations often depends on the balance between available computational resources, time constraints, and the desired level of experimental rigor.

Overall, while K-Fold cross-validation introduces additional computational overhead, the combination of GPU acceleration and optimized deep learning frameworks has made it a practical and powerful evaluation tool for CNN-based image processing studies, particularly in academic research where robustness and reproducibility are of paramount importance.

8. FOLD-WISE ANALYSIS OF PERFORMANCE METRICS

K-Fold cross-validation enables the evaluation of model performance not only through a single average metric but also by considering the distributional characteristics of performance across different data subsets. Performance metrics computed separately for each fold provide a detailed view of how the model behaves under varying data partitions. This aspect is particularly valuable in image-based studies, where datasets often exhibit complex structures and inherent variability.

Although accuracy is commonly the first metric examined in image classification tasks, it may offer a misleading assessment when class imbalance is present. A model may achieve high overall accuracy by favoring dominant classes while failing to adequately learn minority classes. For this reason, complementary metrics such as precision, recall, F1-score, and class-wise accuracy should be analyzed on a fold-by-fold basis to obtain a more comprehensive and reliable evaluation of model performance.

Low variability among fold-level results indicates a stable learning process and suggests that the model generalizes well across different subsets of the data. In contrast, high variance between folds may reveal excessive sensitivity to specific data samples or structural inconsistencies within the dataset. Such analyses not only contribute to a deeper understanding of model robustness but also provide insights into dataset characteristics, including class separability and sample diversity. Consequently, K-Fold cross-validation transforms performance evaluation from a single-point estimate into a multidimensional analytical framework.

9. STATISTICAL INTERPRETATION OF K-FOLD RESULTS

The evaluation of K-Fold cross-validation results should not be limited solely to average performance values. In order to obtain a more comprehensive understanding of the model's behavior across different data subsets, statistical measures must be considered jointly. In this context, reporting the mean performance together with the standard deviation provides valuable insight into the reliability and stability of the obtained results. The standard deviation reflects the variability of the model's performance across different folds and serves as an indicator of learning consistency. A low standard deviation suggests that the model produces similar outcomes across multiple data partitions, indicating strong generalization capability. Conversely, a high standard deviation implies that the model is overly sensitive to specific data subsets, raising concerns regarding its robustness and stability.

This analysis becomes particularly important in studies that compare different model architectures, preprocessing strategies, or feature extraction techniques. In such cases, fold-wise performance results should be examined from a statistical significance perspective rather than relying solely on numerical differences. Statistical evaluation helps determine whether observed performance gaps are meaningful or merely the result of random variation.

By incorporating statistical interpretation into K-Fold cross-validation analysis, researchers can achieve more objective and convincing conclusions. Presenting results in the form of mean ± standard deviation has become a widely accepted practice in academic studies, as it enhances transparency and enables more reliable comparisons across different experimental setups [21], [22], [23].

10. K-FOLD AND MODEL SELECTION IN CNN-BASED IMAGE STUDIES

K-Fold cross-validation serves not only as a tool for evaluating the performance of a given model, but also as an effective framework for selecting the most appropriate model among multiple alternatives. This aspect is particularly important for convolutional neural networks, which involve complex architectural designs and a large number of tunable hyperparameters.

Within this framework, different CNN architectures, network depths, learning rates, or regularization strategies can be systematically compared under identical evaluation conditions. By analyzing the fold-wise performance of each model configuration, more balanced and reliable conclusions can be drawn without relying on a single train-test split. This process helps identify models that consistently perform well across different data subsets rather than those that benefit from favorable data partitioning.

For this reason, K-Fold based model selection has become a common practice in academic studies involving architectural comparisons. Decisions based on multiple-fold evaluations are considered more robust than those derived from a single data split, as they reduce the influence of random variation in data partitioning. Consequently, models selected through K-Fold validation are more likely to demonstrate stable and generalizable performance when applied to unseen data or related problem domains [24].

11. GENERAL DISCUSSION

The K-Fold cross-validation approach discussed in this section adds substantial depth to the model evaluation process in CNN-based image analysis studies. Validation strategies that rely on a single train-test split are often insufficient, particularly when working with limited or imbalanced datasets, as they may yield overly optimistic or misleading performance estimates. In such scenarios, the reported results may fail to accurately represent the true generalization capability of the model.

In contrast, K-Fold cross-validation systematically incorporates the entire dataset into the evaluation pipeline by iteratively assigning different subsets as training and validation data. This structured reuse of data significantly reduces the dependency on a particular data partition and enables a more reliable assessment of model behavior. Especially in image-based deep learning applications, where data collection is both costly and time-consuming, maximizing the utility of available samples is of critical importance. K-Fold cross-validation ensures that each image contributes to both training and testing phases across different iterations, thereby supporting a more comprehensive evaluation of the learned representations.

Beyond average performance values, K-Fold cross-validation also provides insights into the stability and robustness of CNN models. Performance variations observed across folds reveal how sensitive the model is to changes in data composition. Low variance across folds generally indicates consistent learning behavior, whereas high variance may signal overfitting or excessive dependence on specific data subsets. These characteristics make K-Fold particularly valuable in scientific studies where methodological rigor and reproducibility are essential.

Nevertheless, the benefits of K-Fold cross-validation come with certain practical limitations. As summarized in Table 1, the most notable disadvantage of the method is the increased computational cost. Since model training must be repeated K times, the total training time grows proportionally with the number of folds. This issue becomes especially pronounced when deep CNN architectures and high-resolution images are involved. Consequently, the selection of an appropriate K value often represents a trade-off between evaluation reliability and computational feasibility.

Additionally, improper implementation of K-Fold cross-validation may introduce methodological risks, such as data leakage or biased fold construction. Data augmentation strategies, for example, must be applied exclusively within training folds to prevent artificially inflated performance. Similarly, in medical or biological imaging studies, care must be taken to ensure that images originating from the same subject are not distributed across both training and validation folds. Failure to address these concerns can undermine the scientific validity of the results.

Overall, the advantages and disadvantages of K-Fold cross-validation are complementary rather than contradictory, as outlined in Table 1. When implemented with careful consideration of dataset characteristics, hardware resources, and experimental objectives, the advantages clearly outweigh the limitations. For this reason, K-Fold cross-validation should be regarded not merely as a performance evaluation tool, but as a fundamental methodological component that enhances the credibility of CNN-based image analysis studies, particularly in academic research, doctoral dissertations, and investigations involving limited datasets.

Table 1. Advantages and disadvantages of k-fold cross-validation

ADVANTAGES	**DISADVANTAGES**
Utilizes the entire dataset for both training and testing, leading to more reliable performance evaluation.	Requires training the model K times, significantly increasing computational cost and training time.
Reduces the influence of random data splits by avoiding reliance on a single train-test partition.	May be impractical for large datasets or very deep CNN architectures.
Enhances generalization analysis, especially for image datasets with limited samples.	Demands higher hardware resources such as GPUs and memory.
Enables fold-wise performance comparison, facilitating model stability analysis.	Incorrect implementation (e.g., data leakage, weight reuse) can lead to misleading results.
Allows fair comparison of different CNN architectures and hyperparameter settings.	Experimental design and implementation are more complex than simple validation methods.
Improves methodological reliability in medical and biomedical imaging studies.	Not suitable for real-time systems or rapid prototyping scenarios.
Supports statistical interpretation through reporting mean and standard deviation of metrics.	Poor choice of K value may result in unbalanced evaluation.

CONCLUSION

K-Fold cross-validation has been comprehensively examined in the context of CNN-based image analysis studies, with particular emphasis on its methodological contributions to reliable model evaluation. The underlying mechanism of the approach, its practical implementation, as well as its advantages and inherent limitations have been systematically discussed throughout this study. When compared to single train-test validation strategies, K-Fold cross-validation provides a more robust framework for assessing model generalization, especially in scenarios involving limited or imbalanced image datasets.

One of the key strengths of the K-Fold approach lies in its ability to ensure that all available data samples contribute to both training and testing phases across different iterations. As summarized in Table 2, this systematic reuse of data enhances the stability of performance estimates and reduces the risk of biased evaluation caused by unfavorable data splits. In CNN-based image classification tasks, where performance is often sensitive to dataset composition, this property significantly improves the reliability of reported results.

Table 2. Advantages and Limitations of K-Fold Cross-Validation

Advantages	**Limitations**
Ensures that the entire dataset is included in the evaluation process	Increases training time by a factor of K
Eliminates dependence on a single train-test split	May require high computational resources
Provides a more reliable estimation of model generalization performance	Incorrect fold construction may lead to data leakage
Enables analysis of fold-wise performance variability	Becomes computationally demanding for large-scale datasets
Enhances methodological reliability in academic studies	Inappropriate data augmentation strategies may bias results

In addition to performance robustness, K-Fold cross-validation also supports informed model selection and hyperparameter optimization. As outlined in Table 3, fold-based evaluation enables fair comparisons between different CNN architectures, optimization strategies, and preprocessing pipelines. This aspect is particularly valuable in academic studies where multiple model variants are assessed, and conclusions must be drawn based on consistent and reproducible criteria rather than isolated experimental outcomes.

Table 3. Comparison of Single Train-Test Split and K-Fold Cross-Validation

Criterion	Single Train-Test Split	K-Fold Cross-Validation
Data usage	Only a portion of the dataset is used for testing	Entire dataset is used for both training and testing across folds
Sensitivity to data partition	Highly sensitive to random split selection	Less sensitive due to multiple evaluations
Generalization reliability	Limited and potentially misleading	More reliable and robust
Performance stability	Depends on a single split	Averaged over multiple folds
Suitability for small datasets	Generally insufficient	Highly suitable
Computational cost	Low	Higher (training repeated K times)

Despite its methodological advantages, K-Fold cross-validation introduces notable computational challenges, particularly when applied to deep CNN architectures and high-resolution image data.

The trade-off between evaluation reliability and computational cost is highlighted in Table 4, which summarizes practical considerations related to training time, hardware requirements, and resource utilization. These factors underline the importance of selecting an appropriate K value that balances methodological rigor with feasible computational effort.

Table 4. Performance Analysis Metrics in K-Fold Cross-Validation

Metric	Description	Importance in K-Fold Evaluation
Accuracy	Overall proportion of correctly classified samples	Provides general performance overview
Precision	Ratio of correctly predicted positives to total predicted positives	Important for reducing false positives

Recall (Sensitivity)	Ratio of correctly predicted positives to actual positives	Critical for detecting minority classes
F1-Score	Harmonic mean of precision and recall	Balances precision-recall trade-off
Standard Deviation	Variability of metric across folds	Indicates model stability
Class-wise Accuracy	Accuracy computed separately for each class	Essential for imbalanced datasets

Finally, correct implementation plays a decisive role in the scientific validity of K-Fold results. Issues such as data leakage, improper fold construction, and incorrect handling of data augmentation can severely compromise evaluation outcomes. As emphasized in Table 5, adherence to best practices-including class-balanced fold generation, strict separation of training and testing data, and controlled augmentation strategies is essential to fully benefit from the K-Fold methodology.

Table 5. Practical Considerations for Applying K-Fold Cross-Validation in CNN-Based Image Analysis

Aspect	**Consideration**	**Impact on Results**
Choice of K value	Typically 5 or 10	Affects bias-variance trade-off
Hardware resources	GPU availability and memory	Determines feasibility and training time
Data augmentation	Must be applied only to training folds	Prevents data leakage
Fold construction	Stratified or subject-wise splitting	Preserves class distribution
Training initialization	Random weight initialization	Influences fold-level variability
Parallel execution	Multi-GPU support	Reduces total experiment time

In conclusion, K-Fold cross-validation should not be regarded merely as an auxiliary evaluation technique, but rather as a fundamental methodological component in CNN-based image analysis research. When thoughtfully designed and carefully implemented, it provides robust, interpretable, and convincing performance assessments. This makes K-Fold cross-validation particularly well suited for doctoral dissertations, peer-reviewed journal articles, and image-based studies where data availability is limited and methodological reliability is of paramount importance.

REFERENCES

[1] M. Kaynaklı, et al., “Solar Radiation and Temperature Effects on Agricultural Irrigation Systems,” *Bitlis Eren University Journal of Science and Technology*, vol. 6, no. 1, pp. 53–58, Jun. 2016, doi: 10.17678/BEUJST.79941.

[2] S. RÜSTEMLİ et al. “Active filter solutions in energy systems,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 23, no. 6, pp. 1587–1607, Jan. 2015, doi: 10.3906/elk-1402-212.

[3] D. Passos, “Deep tutti-frutti II: Explainability of CNN architectures for fruit dry matter predictions,” *Spectrochim Acta A Mol Biomol Spectrosc*, vol. 337, Sep. 2025, doi: 10.1016/j.saa.2025.126068.

[4] S. Wang and S. Cheng, “Parameter Optimization for Model-Free Predictive Current Control of PMSMs Using ANN and K-Fold Cross-Validation,” *Chinese Control Conference, CCC*, pp. 3048–3053, 2025, doi: 10.23919/CCC64809.2025.11178539.

[5] K. Jung, D. H. Bae, M. J. Um, S. Kim, S. Jeon, and D. Park, “Evaluation of Nitrate Load Estimations Using Neural Networks and Canonical Correlation Analysis with K-Fold Cross-Validation,” *Sustainability 2020, Vol. 12,* vol. 12, no. 1, Jan. 2020, doi: 10.3390/SU12010400.

[6] O. Saidani *et al.*, “White blood cells classification using multi-fold pre-processing and optimized CNN model,” *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–14, Feb. 2024, doi: 10.1038/s41598-024-52880-0.

[7] M. K. Lubis and H. SyahPutra, “Classification of Orange Peel Conditions Using Transfer Learning MobileNetV2 with K-Fold Cross Validation,” *Journal of Artificial Intelligence and Engineering Applications (JAIEA)*, vol. 5, no. 1, pp. 2027–2031, Oct. 2025, doi: 10.59934/JAIEA.V5I1.1778.

[8] O. Palta, M. Çıbuk, and H. Güldemir, “Deep Learning-Based Classification Of White Blood Cells: An Ensemble Model Approach,” *7. INTERNATIONAL MEDITERRANEAN CONGRESS*, pp. 262–270, 2025.

[9] H. Ziang et al. “Framework for lung CT image segmentation based on UNet++,” Jan. 2025, Accessed: Dec. 27, 2025. [Online]. Available: https://arxiv.org/pdf/2501.02428

[10] D. Ayman, A. Taleb, M. M. Morsey, M. Omar, and E.-S. M. El-Horbaty, “Predicting Early-stage ASD using symptomatic dataset for Arab children: A comparison of K-fold cross validation and train-test approaches,” *International Journal of Intelligent Computing and Information Sciences*, vol. 25, no. 2, pp. 32–41, Jun. 2025, doi: 10.21608/IJICIS.2025.378373.1390.

[11] "Ensemble based brain tumor classification technique from MRI based on K fold validation approach | Journal of Integrated Science and Technology." Accessed: Dec. 10, 2025. [Online]. Available: https://pubs.thesciencein.org/journal/index.php/jist/article/view/a1114

[12] R. Göçmen, M. Çıbuk, and E. Akin, "Comparative Analysis of Deep Learning Algorithms in Fire Detection," *Balkan Journal of Electrical and Computer Engineering*, vol. 12, no. 3, pp. 255–261, Sep. 2024, doi: 10.17694/BAJECE.1533966.

[13] Ü. Budak, M. Çıbuk, Z. Cömert, and A. Şengür, "Efficient COVID-19 Segmentation from CT Slices Exploiting Semantic Segmentation with Integrated Attention Mechanism," *Journal of Digital Imaging 2021 34:2*, vol. 34, no. 2, pp. 263–272, Mar. 2021, doi: 10.1007/S10278-021-00434-5.

[14] İ. Söylemez, M. E. Nalici, and R. Ünlü, "High-Accuracy Identification of Durian Leaf Diseases: A Convolutional Neural Network Approach Validated with K-Fold Cross-Validation and Bayesian Optimization," *Applied Fruit Science 2025 67:6*, vol. 67, no. 6, pp. 464-, Nov. 2025, doi: 10.1007/S10341-025-01698-9.

[15] Ç. Cengiz, "Nonparametric estimation of a renewal function in the case of censored sample," *Bitlis Eren University Journal of Science and Technology*, vol. 9, no. 2, pp. 54–57, Dec. 2019, doi: 10.17678/BEUSCITECH.556451.

[16] M. Kaynaklı, et al. "Cooperation of Conventional Electric Power Grids and Smart Power Grids," vol. 11, no. 6, pp. 23–27, 2016, doi: 10.9790/1676-1106042327.

[17] S. Rüstemli, et al., Elektronik Mühendisliği Bölümü, and T. Bilimler Meslek Yüksekokulu, "Fotovoltaik Paneller: Güneş Takip Sistemleri ve İklimlendirme Sistemleri," *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, vol. 2, no. 2, pp. 141–147, Dec. 2013, Accessed: Dec. 27, 2025. [Online]. Available: https://dergipark.org.tr/en/pub/bitlisfen/article/48984

[18] O. Palta, M. Çıbuk, and H. Güldemir, "Image Enhancement In Whıte Blood Cells Usıng Hybrıd (Chale+Gauss) Fılter And Flask-Based Web Servıce Integratıon," *7. INTERNATIONAL MEDITERRANEAN CONGRESS*, pp. 897–906, 2025.

[19] Z. Cömert, et al. "Convolutional Block Attention Module and Parallel Branch Architectures for Cervical Cell Classification," *Int J Imaging Syst Technol*, vol. 35, no. 2, Mar. 2025, doi: 10.1002/IMA.70048.

[20] A. Sengur, Y. Akbulut, U. Budak, and Z. Comert, "White Blood Cell Classification Based on Shape and Deep Features," *2019 International*

Conference on Artificial Intelligence and Data Processing Symposium, IDAP 2019, Sep. 2019, doi: 10.1109/IDAP.2019.8875945.

[21] G. Ozmen Koca Associate Professor and S. Yetkin, “A Hybrid Routing Approach Using Two Searching Layers,” *Electrical, Control and Communication Engineering*, vol. 17, no. 1, pp. 59–66, Jun. 2021, doi: 10.2478/ECCE-2021-0007.

[22] G. O. Koca *et al.*, “Three-Dimensional Modeling of a Robotic Fish Based on Real Carp Locomotion,” *Applied Sciences 2018, Vol. 8,* vol. 8, no. 2, Jan. 2018, doi: 10.3390/APP8020180.

[23] B. Erdem et al., “Cheating Detection in Online Exams Using Deep Learning and Machine Learning,” *Applied Sciences (Switzerland)*, vol. 15, no. 1, p. 400, Jan. 2025, doi: 10.3390/APP15010400.

[24] Ç. Cengiz and H. Aydoğdu, “GAMMA Renewal Function in Censored Data,” *Bitlis Eren University Journal of Science and Technology*, vol. 5, no. 2, pp. 97–101, Dec. 2015, doi: 10.17678/BEUJST.09962.

Chapter 12

A Comprehensive Overview to Power Quality Issues in Smart Grids

Harun ÖZBAY[1], Serhat Berat EFE[2]

1. Introduction

The transition from conventional power systems to smart grids has introduced a paradigm shift in the way electrical energy is generated, distributed, and consumed. This transformation is driven by the integration of distributed energy resources (DER), bidirectional power flows, advanced metering infra-structure (AMI), and intelligent control mechanisms. While these innovations offer enhanced flexibility, efficiency, and sustainability, they also pose significant challenges to maintaining power quality (PQ) across increasingly complex and dynamic networks (Baidya et al. 2021). Power quality, encompassing parameters such as voltage stability, harmonic distortion, flicker, and transient phenomena, remains a critical determinant of system reliability and end-user satisfaction. In smart grid environments, the proliferation of nonlinear loads, inverter-based generation, and real-time switching operations exacerbates PQ disturbances, necessitating robust monitoring and mitigation strategies.

The increasing penetration of renewable energy sources, particularly wind and solar photovoltaic systems, introduces inherent variability and uncertainty into power generation profiles. Unlike conventional synchronous generators that provide natural inertia and voltage support, inverter-based resources exhibit fundamentally different dynamic characteristics that can com-promise grid stability during disturbances. The intermittent nature of solar irradiance and wind speed creates rapid voltage fluctuations at distribution feeders, while the power electronic interfaces of these systems inject harmonics and interharmonics across a wide frequency spectrum. Furthermore, the

---

[1] Assoc. Prof. Dr., Bandırma Onyedi Eylül University, Dept. Of Electrical Engineering, 10200, Bandırma. hozbay@bandirma.edu.tr, ORCID: 0000-0003-1068-244X

[2] Assoc. Prof. Dr., Bandırma Onyedi Eylül University, Dept. Of Electrical Engineering, 10200, Bandırma. sefe@bandirma.edu.tr, ORCID: 0000-0001-6076-4166

bidirectional power flow characteristic of modern distribution networks—where consumers can simultaneously act as prosumers—challenges traditional voltage regulation schemes and protective coordination strategies that were designed under the assumption of unidirectional power flow from transmission systems to end users (YILDIRAN and TACER 2016).

The proliferation of electric vehicles (EVs) and their associated charging infrastructure adds another layer of complexity to power quality management in smart grids. EV chargers, particularly fast-charging stations utilizing high-power rectifiers and DC-DC converters, represent significant nonlinear loads that contribute to harmonic pollution and can cause localized voltage deviations during charging events. The stochastic nature of EV charging behavior, coupled with the potential for simultaneous charging of multiple vehicles in residential areas during peak evening hours, can lead to transformer overloading, phase unbalance, and voltage sags that propagate through the distribution network. Additionally, the integration of vehicle-to-grid (V2G) technology, while offering potential benefits for grid support and energy storage, introduces bidirectional power electronics that further complicate harmonic management and reactive power control.

Advanced metering infrastructure and the deployment of numerous intelligent electronic devices throughout smart grids have created an environment where high-frequency disturbances and electromagnetic interference are increasingly prevalent. The dense communication networks required for real-time monitoring and control—including power line carrier systems, wireless sensor networks, and fiber optic communication—can be susceptible to power quality disturbances, potentially com-promising the very intelligence that defines smart grids. More-over, the proliferation of consumer electronics, LED lighting, variable frequency drives in industrial applications, and other switching-mode power supplies has fundamentally altered the load composition of modern power systems, creating aggregate harmonic currents and demanding more stringent power quality standards to ensure equipment compatibility and longevity. These interconnected challenges underscore the critical need for comprehensive power quality assessment frameworks, advanced monitoring technologies, and adaptive mitigation strategies that can operate effectively in the dynamic, distributed, and data-intensive environment of contemporary smart grids (Luo et al. 2016).

This chapter provides a comprehensive overview of power quality management within smart grids, focusing on international standards, emerging technologies, and future directions. It aims to bridge the gap between theoretical frameworks and practical implementations, offering insights into how intelligent systems can ensure compliance with PQ norms while adapting to evolving grid conditions.

2. Methodology

The management of PQ in smart grids is governed by a set of international standards that define acceptable limits for electrical parameters and provide guidelines for measurement, evaluation, and mitigation. These standards are essential for ensuring interoperability, reliability, and safety across diverse grid infrastructures and technologies.

Figure 1 shows the smart grid architecture with PQ monitoring layers that visualizes how PQ data flows from field devices to centralized decision systems. The block diagram showing DERs (solar, wind, storage), loads (residential, industrial), PQ sensors and smart meters, communication layers (IoT, edge computing, cloud) and control center with AI-based PQ analytics.

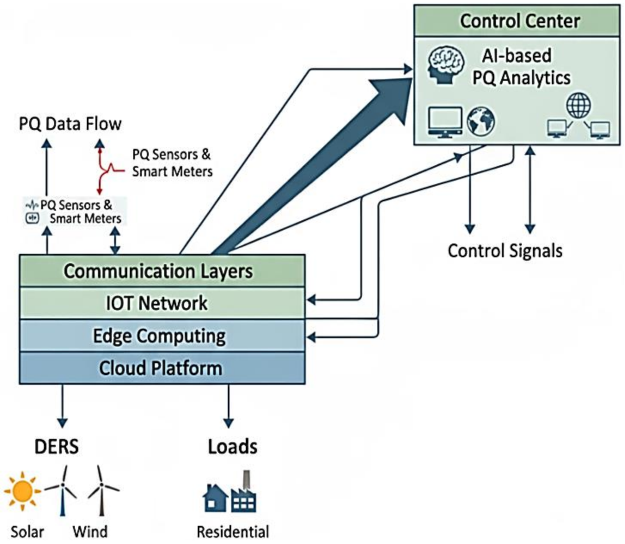


Figure 1. Smart Grid Architecture with PQ Monitoring Layers

The IEEE 519 standard, titled IEEE Recommended Practices and Requirements for Harmonic Control in Electrical Power Systems, is one of the most widely adopted frameworks for managing harmonic distortion. It establishes limits for voltage and current harmonics at the point of common coupling (PCC), tailored to system voltage levels and load characteristics.

IEEE 519 defines Total Harmonic Distortion (THD) thresholds for voltage and current, differentiating between individual harmonic orders and aggregate distortion, and provides guidance for utilities and end-users to maintain harmonic compliance. In smart grids, where inverter-based Distributed Energy Resources (DERs) and nonlinear loads are prevalent, IEEE 519 serves as a foundational reference for designing filtering systems and validating harmonic performance. Similarly, the European standard EN 50160 specifies the main voltage quality parameters in public low-voltage and medium-voltage distribution networks, outlining the expected ranges for frequency, voltage magnitude, flicker, harmonics, and voltage dips. Key parameters include supply voltage variations within ±10% of nominal, frequency stability of typically 50 Hz ±1%, short-duration voltage dips and interruptions, and long-term flicker severity ($P_{st} \leq 1.0$). EN 50160 is particularly relevant for smart grid applications in Europe, where compliance with voltage quality norms is essential for ensuring the reliable operation of grid-connected devices and maintaining consumer satisfaction. Additionally, the IEC 61000 series, developed by the International Electrotechnical Commission, covers a broad spectrum of power quality and electromagnetic compatibility (EMC) topics, including measurement techniques and immunity requirements. Notable subsections include IEC 61000-4-30 for power quality measurement methods, IEC 61000-2-2 for compatibility levels in low-voltage systems, and IEC 61000-3-2/3-12, which set limits for harmonic emissions from equipment.

These standards are critical for ensuring that smart grid components such as smart meters, controllers, and communication devices operate reliably under varying electromagnetic conditions. While IEEE 519, EN 50160, and the IEC 61000 series share common objectives related to maintaining power quality and electromagnetic compatibility, their scope and application differ depending on regional regulations and grid architectures. Consequently, smart grid deployments often adopt hybrid compliance strategies that integrate multiple standards to meet diverse regulatory and technological requirements. Table 1 shows the summary of PQ standards and key parameters (Hossein Pourarab, Nakhodchi, and Monfared 2015; International Electrotechnical Commission (IEC). IEC 61000-3-2:2018. n.d.).

Table 1. Summary of PQ Standards and Key Parameters

Standard	**Focus Area**	**Applicable Region**	**Key PQ Parameters**
IEEE 519	Harmonic distortion	Global (incl. US)	THD, individual harmonics
EN 50160	Voltage characteristics	Europe	Voltage variation, flicker, dips
IEC 61000	EMC and PQ measurement	Global	Measurement protocols, compatibility

3. Results and Discussion

The implementation of advanced technological solutions for power quality monitoring and control in smart grids has demonstrated significant progress in addressing the complex challenges posed by distributed energy resources and dynamic load conditions. The integration of Internet of Things-based monitoring systems has revolutionized the traditional approach to power quality management by enabling real-time data acquisition from multiple grid nodes, including voltage, current, harmonics, and flicker measurements. These systems leverage wireless communication protocols such as Zigbee, LoRa-WAN, and NB-IoT to establish interconnected networks of sensors and smart meters, providing utilities with granular visibility across the grid infrastructure. The cloud-based storage and analytics capabilities of these systems facilitate comprehensive historical trend analysis, enabling early detection of power quality anomalies and localized disturbances. This scalable architecture has proven particularly effective in both urban and rural deployments, significantly enhancing situational awareness for grid operators and enabling proactive management strategies rather than reactive responses (Eslami et al. 2022).

The application of artificial intelligence and machine learning techniques has emerged as a transformative approach to power quality diagnostics, addressing the challenge of analyzing large volumes of data generated by modern monitoring systems (Akdeniz and Efe 2025). Unsupervised learning algorithms, including autoencoders and clustering methods, have demonstrated remarkable capability in identifying deviations from normal power quality patterns without requiring extensive labeled datasets. Supervised models such as support vector machines (Hu, Zhu, and Ren 2008) and random forests have shown high accuracy in classifying specific power quality events, including voltage sags, swells, and transients, while time-series forecasting models like LSTM networks and ARIMA have enabled predictive maintenance by anticipating future disturbances based on historical trends (Wang et al. 2018; Žnidarec et al. 2019).

These AI-driven approaches have significantly reduced false alarm rates and improved diagnostic accuracy, while their adaptive learning capabilities allow them to evolve with changing grid conditions. The integration of these machine-learning models with decision support systems has enabled automated response mechanisms, reducing the burden on human operators and accelerating mitigation actions (Naik et al. 2016; Özer, Efe, and Özbay 2021; Xue et al. 2020). Table 2 shows a summary of AI techniques used for PQ analysis.

Table 2. AI Techniques for PQ Analysis

Technique	**Application Area**	**Strengths**
SVM	PQ event classification	High accuracy, interpretable
Autoencoders	Anomaly detection	Unsupervised, adaptive
LSTM	PQ forecasting	Temporal modeling, long-term memory
Random Forest	Feature importance	Robust to noise, fast inference

The development and deployment of hybrid filtering and compensation techniques have proven essential in mitigating power quality disturbances, particularly harmonics and voltage fluctuations that are prevalent in smart grid environments with high penetrations of power electronic converters (Efe 2015; Rüstemli and Cengiz 2015). Hybrid filters, which combine the simplicity and cost-effectiveness of passive LC circuits with the dynamic performance of active power filters, have demonstrated superior performance in managing harmonic content across a broad frequency spectrum. Advanced control strategies, including instantaneous reactive power theory, synchronous reference frame control, and model predictive control, have enabled these systems to provide dynamic compensation that adapts to rapidly changing load profiles. Field implementations in industrial and commercial settings with high harmonic content have validated the effectiveness of these systems, showing significant improvements in total harmonic distortion levels and voltage stability. The application of edge computing and distributed intelligence has further enhanced power quality management by enabling localized data processing and decision-making at measurement points, substantially reducing latency and bandwidth requirements while improving system resilience through decentralized control architectures (Gong et al. 2021).

Looking toward future perspectives in power quality management, the evolution from static to dynamic compliance frameworks represents a fundamental shift in how power quality standards are conceived and implemented. Traditional fixed-threshold standards are increasingly recognized as inadequate for smart grids characterized by time-varying conditions, diverse distributed energy resource behaviors, and complex load profiles. The development of context-aware compliance frameworks that dynamically adjust thresholds based on real-time grid topology, time-of-day patterns, and distributed energy resource penetration levels promises to provide more realistic and effective power quality governance. This evolution toward probabilistic limits, event-based tolerances, and AI-assisted compliance scoring reflects a broader recognition that power quality management must be fundamentally adaptive to accommodate the inherent variability and uncertainty of modern grid operations.

The integration of power quality considerations into grid planning and design phases represents a paradigm shift from treating power quality issues as post-deployment problems to addressing them proactively during system development. Power quality simulation tools integrated into distribution system planning processes enable utilities to predict and mitigate potential disturbances before they materialize, optimizing the placement of filters, compensators, and sensors based on comprehensive harmonic load flow and flicker impact analyses. Digital twin technology has emerged as a particularly powerful platform for power quality simulation and optimization, providing real-time virtual replicas of physical grid components that can mirror actual conditions and enable scenario testing for various operational conditions, including distributed energy resource integration, fault events, and load changes. These digital representations facilitate predictive power quality optimization through the application of artificial intelligence and advanced control algorithms, allowing utilities to explore mitigation strategies in a risk-free virtual environment before implementing them in the physical grid.

Figure 2 illustrates the integration of passive and active filtering components within a hybrid power quality mitigation framework. Passive filters are employed to target specific harmonic orders through tuned LC circuits, while active power filters dynamically inject compensating currents based on real-time control signals. The control loop governs the coordination between these components, ensuring adaptive response to varying load conditions and harmonic profiles. This schematic serves to demonstrate the practical implementation of hybrid PQ mitigation strategies, highlighting the synergy between conventional and power-electronic-based solutions in enhancing grid stability and compliance with harmonic standards.

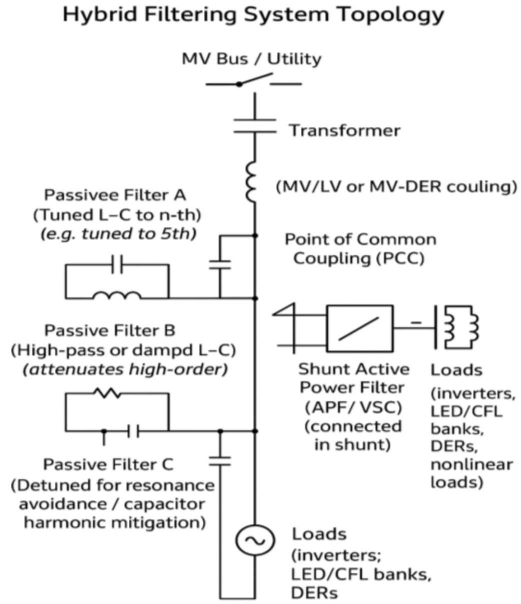


Figure 2. Hybrid Filtering System Topology

The potential application of blockchain technology to power quality management introduces novel possibilities for enhancing data integrity, transparency, and stakeholder engagement in multi-actor smart grid ecosystems. Immutable logging of power quality events and compliance records on distributed ledgers can provide unprecedented transparency and accountability, while smart contracts offer mechanisms for incentivizing power quality compliance by prosumers and other grid participants.

Decentralized verification of power quality metrics through blockchain-based systems could streamline regulatory audits and foster trust-based governance frameworks that are essential for the collaborative operation of future smart grids. Furthermore, the emerging focus on human-centric power quality metrics represents an important evolution beyond traditional technical parameters to incorporate factors directly relevant to end-user experience, comfort, and health. Dynamic lighting quality linked to circadian rhythms, voltage stability thresholds tuned to sensitive medical equipment requirements, and user feedback loops for adaptive control interfaces exemplify this shift toward aligning power quality management with broader societal goals of sustainability, inclusivity, and digital empowerment, ultimately recognizing that power quality is not merely a technical specification but a critical determinant of quality of life in increasingly electrified societies.

4. Conclusion

Power quality management in smart grids is entering a transformative era characterized by intelligence, adaptability, and the recognition of multidimensional value that extends far beyond traditional technical specifications. The convergence of advanced technologies with evolving regulatory frameworks is fundamentally reshaping how power quality is conceptualized, monitored, and controlled within modern electrical systems. This paradigm shift reflects a deeper understanding that power quality in the context of smart grids cannot be adequately ad-dressed through conventional static standards and reactive mitigation strategies, but rather requires dynamic, intelligent approaches that can adapt to the complex and continuously evolving nature of distributed energy systems.

The alignment of international standards with emerging technologies such as the Internet of Things, artificial intelligence, digital twins, and blockchain represents a critical pathway toward achieving comprehensive and effective power quality governance in future energy systems. IoT-enabled monitoring infrastructure provides the foundational layer of real-time visibility and data acquisition necessary for informed decision-making, while artificial intelligence and machine learning techniques transform this data into actionable insights through sophisticated pattern recognition, anomaly detection, and predictive analytics. Digital twin technology bridges the gap be-tween physical and virtual domains, enabling risk-free experimentation and optimization of power quality mitigation strategies before their implementation in actual grid operations. Blockchain technology introduces unprecedented levels of transparency, data integrity, and stakeholder accountability, facilitating trust-based collaborative frameworks essential for multi-actor smart grid ecosystems. The integration of these technologies into coherent power quality management frame-works promises not merely incremental improvements but fundamental transformations in how utilities, regulators, and end-users interact with and govern power quality.

The evolution of power quality frameworks toward encompassing technical compliance alongside broader objectives of resilience, transparency, and user-centric innovation signals a maturation of the field beyond its traditional engineering boundaries. Resilience considerations recognize that power quality management must contribute to the overall robustness and adaptive capacity of energy systems in the face of increasing climate-related disruptions, cybersecurity threats, and operational uncertainties. Transparency imperatives acknowledge that modern energy consumers and prosumers demand clear, accessible information about power quality performance and its implications for their equipment, comfort, and economic interests. User-centric innovation reflects the growing recognition that power quality standards and control strategies must be de-signed with explicit consideration of human factors, incorporating metrics that relate directly to perceived quality of service, health outcomes, and productivity rather than solely technical parameters. This holistic approach positions power quality management as an integral component of sustainable, inclusive, and equitable energy transitions.

As smart grids continue to mature and proliferate globally, the strategic repositioning of power quality from a technical constraint to a valuable asset represents perhaps the most significant conceptual shift in the field. This transformation acknowledges that superior power quality can serve as a differentiator in competitive energy markets, a prerequisite for hosting sensitive digital infrastructure, and an enabler of advanced grid services and flexibility mechanisms. Power quality excellence can facilitate higher penetrations of distributed energy resources by mitigating their potentially disruptive effects, support the electrification of transportation and heating sectors that demand high-quality electrical supply, and enable new business models based on premium power quality services tailored to specific customer segments. The strategic management of power quality assets through predictive analytics, optimal resource allocation, and dynamic pricing mechanisms can unlock significant economic value while simultaneously enhancing system reliability and sustainability.

Ultimately, the future trajectory of power quality management in smart grids will be determined by the collective ability of utilities, technology providers, regulators, and researchers to embrace innovation while maintaining focus on fundamental objectives of reliability, safety, and accessibility. The techno-logical capabilities now emerging provide unprecedented opportunities to reimagine power quality governance, but their successful deployment requires careful attention to standardization, interoperability, cybersecurity, and equitable access.

As these challenges are progressively addressed through collaborative international efforts, power quality management will increasingly serve as a cornerstone of resilient, intelligent, and sustainable energy systems that meet the diverse needs of twenty-first-century societies. The integration of advanced monitoring, artificial intelligence-driven diagnostics, adaptive control systems, and human-centric metrics positions power quality management at the forefront of the smart grid revolution, ultimately shaping the future of energy systems worldwide and contributing to the broader transition toward decarbonized, digitalized, and democratized electricity infrastructures.

References

Akdeniz, Metin, and Serhat Berat Efe. 2025. "Forecasting of Harmonic Distortions in MV Distribution Network Caused by Illumination Devices." *Light and Engineering* 33(2):64–69. doi:10.33383/2024-088.

Baidya, Sanghita, Vidyasagar Potdar, Partha Pratim Ray, and Champa Nandi. 2021. "Reviewing the Opportunities, Challenges, and Future Directions for the Digitalization of Energy." *Energy Research and Social Science* 81.

Efe, Serhat Berat. 2015. "Harmonic Filter Application for an Industrial Installation." Pp. 31–34 in *2015 13th International Conference on Engineering of Modern Electric Systems, EMES 2015*.

Eslami, Ahmadreza, Michael Negnevitsky, Evan Franklin, and Sarah Lyden. 2022. "Review of AI Applications in Harmonic Analysis in Power Systems." *Renewable and Sustainable Energy Reviews* 154:1–26.

Gong, Jie, Dayi Li, Tingkang Wang, Wenhao Pan, and Xinzhi Ding. 2021. "A Comprehensive Review of Improving Power Quality Using Active Power Filters." *Electric Power Systems Research* 199:1–15.

Hossein Pourarab, Morteza, Naser Nakhodchi, and Mohammad Monfared. 2015. "Harmonic Analysis of Led Street Lighting According to IEC61000-3-2 ; a Case Study." Pp. 1–4 in *23rd International Conference on Electricity Distribution*.

Hu, Guo Sheng, Feng Feng Zhu, and Zhen Ren. 2008. "Power Quality Disturbance Identification Using Wavelet Packet Energy Entropy and Weighted Support Vector Machines." *Expert Systems with Applications* 35(1–2):143–49. doi:10.1016/j.eswa.2007.06.005.

International Electrotechnical Commission (IEC). IEC 61000-3-2:2018. n.d. "Electromagnetic Compatibility (EMC)—Part 3-2: Limits—Limits for Harmonic Current Emissions (Equipment Input Current_16 A per Phase)." https://webstore.iec.ch/publication/62553.

Luo, An, Qianming Xu, Fujun Ma, and Yandong Chen. 2016. "Overview of Power Quality Analysis and Control Technology for the Smart Grid." *Journal of Modern Power Systems and Clean Energy* 4(1):1–9.

Naik, Chirag, Faizal Hafiz, Akshya Swain, and A. K. Kar. 2016. "Classification of Power Quality Events Using Wavelet Packet Transform and Extreme Learning Machine." Pp. 1–6 in *2016 IEEE 2nd Annual Southern Power Electronics Conference, SPEC 2016*.

Özer, İlyas, Serhat Berat Efe, and Harun Özbay. 2021. "CNN / Bi-LSTM-Based Deep Learning Algorithm for Classification of Power Quality Disturbances by Using Spectrogram Images." *International Transactions on Electrical Energy Systems* 31(12):1–16. doi:10.1002/2050-7038.13204.

Rüstemli, Sabir, and Mehmet Sait Cengiz. 2015. “Active Filter Solutions in Energy Systems.” *Turkish Journal of Electrical Engineering & Computer Sciences* 23:1587–1607. doi:10.3906/elk-1402-212.

Wang, Fei, Yili Yu, Zhanyao Zhang, Jie Li, Zhao Zhen, and Kangping Li. 2018. “Wavelet Decomposition and Convolutional LSTM Networks Based Improved Deep Learning Model for Solar Irradiance Forecasting.” *Applied Sciences (Switzerland)* 8(8):1–30. doi:10.3390/app8081286.

Xue, Haihua, Alian Chen, Deqiang Zhang, and Chenghui Zhang. 2020. “A Novel Deep Convolution Neural Network and Spectrogram Based Microgrid Power Quality Disturbances Classification Method.” *Conference Proceedings - IEEE Applied Power Electronics Conference and Exposition - APEC* 2020-March:2303–7. doi:10.1109/APEC39645.2020.9124252.

YILDIRAN, Nezihe, and M. Emin TACER. 2016. “The Future’s Power System: Smart Grid.” *International Journal of Electronics, Mechanical and Mechatronics Engineering* 6(1):1047–55.

Žnidarec, Matej, Zvonimir Klaić, Damir Šljivac, and Boris Dumnić. 2019. “Harmonic Distortion Prediction Model of a Grid-Tie Photovoltaic Inverter Using an Artificial Neural Network.” *Energies* 12(5)..

Chapter 13

Artificial Intelligence Prompt Writing: Theory and Practice

Rabia KORKMAZ TAN[1]

I. INTRODUCTION

Artificial intelligence (AI) and large language models (LLMs) have fundamentally transformed the landscape of natural language processing by enabling machines to comprehend, interpret, and generate human language with unprecedented sophistication. These advancements have paved the way for numerous applications, ranging from automated content creation and language translation to intelligent conversational agents and personalized virtual assistants. At the heart of these innovations lies the concept of prompt writing, which involves crafting clear, precise, and contextually relevant inputs that effectively steer AI systems toward producing accurate and meaningful outputs(D Kulkarni, 2024; D. Park et al., 2024).

Prompt writing is a critical skill because the nature and quality of the prompt directly determine the AI's response quality. A well designed prompt reduces ambiguity and guides the AI to focus on the intended subject matter, style, or format, thereby enhancing the coherence, relevance, and overall usefulness of the generated content. This process requires an understanding of both the AI's capabilities and limitations, as well as the specific goals of the user, to ensure that the interaction yields productive and reliable results. As AI technologies continue to evolve and become more integrated into various academic, professional, and creative workflows, the ability to master prompt writing becomes essential for maximizing the benefits of these tools(J. Park & Choo, 2024).

Moreover, effective prompt writing fosters efficient communication between humans and AI, enabling users to harness the full potential of language models without extensive trial and error. By carefully structuring prompts, users can elicit responses that are not only contextually appropriate but also aligned with nuanced requirements such as tone, complexity, and specificity. This precision is particularly vital in domains where accuracy and clarity are paramount, such

---

[1] Tekirdağ Namık Kemal University, 0000 0002 3777 2536

as academic research, legal documentation, and technical writing(J. Park & Choo, 2024).

In addition, prompt writing plays a vital role in mitigating common challenges associated with AI generated content, such as irrelevant or off topic responses, factual inaccuracies, and stylistic inconsistencies. By anticipating potential ambiguities and explicitly defining the scope and constraints within the prompt, users can significantly improve the reliability and applicability of AI outputs. This proactive approach not only saves time and resources but also enhances user confidence in deploying AI tools across sensitive or high stakes environments (Dobbe et al., 2021; Swathi & Veeramanju, 2024).

Furthermore, as AI models continue to advance, incorporating more complex reasoning abilities and broader contextual understanding, the sophistication of prompt writing must evolve in tandem. Users are encouraged to develop iterative prompt refinement strategies, leveraging feedback loops to progressively enhance output quality. This dynamic interaction transforms prompt writing from a static instruction into an adaptive dialogue, where continuous learning and adjustment enable more nuanced and tailored AI responses. Consequently, mastering prompt writing is not only about initial input design but also about cultivating an ongoing process of optimization that aligns AI generated content with evolving user needs and expectations.

The significance of prompt writing extends beyond just improving output quality; it also influences the ethical and responsible use of AI technologies. By carefully framing prompts, users can reduce the risk of generating biased, misleading, or inappropriate content, thereby promoting more trustworthy AI interactions. This ethical dimension underscores the responsibility of users to engage thoughtfully with AI systems, ensuring that the outputs contribute positively to their intended contexts(Flores Romero et al., 2025; Lo, 2023).

Ultimately, the art and science of prompt writing empower users to unlock more sophisticated, tailored, and impactful AI generated content, making it a foundational component in the ongoing advancement and practical application of artificial intelligence. As AI becomes increasingly embedded in diverse sectors, proficiency in prompt writing will distinguish successful users who can fully leverage these technologies to enhance creativity, productivity, and decision making across a wide spectrum of disciplines. This evolving expertise will be crucial in shaping the future of human AI collaboration, where clear communication and precise guidance enable AI to serve as an effective partner in knowledge generation and problem solving(Molla, 2024; Torricelli et al., 2024).

II. THEORETICAL FOUNDATIONS OF PROMPT WRITING

a. The principles of human computer interaction (HCI)

Human computer interaction (HCI) constitute an essential framework for prompt design, ensuring that communication between users and AI systems is not only intuitive but also highly efficient. Effective prompt design demands a comprehensive understanding of how humans perceive, process, and respond to information presented through digital interfaces. By integrating HCI principles, designers can develop prompts that are clear, straightforward, and closely aligned with the natural cognitive patterns of users. This alignment significantly reduces the likelihood of user confusion or frustration, fostering smoother, more productive interactions with AI tools. Key elements such as linguistic clarity, structural simplicity, and the natural progression of conversational flow are emphasized to minimize cognitive load, which in turn enhances overall usability and user satisfaction. This thoughtful approach makes the interaction feel seamless and user friendly, encouraging users to engage confidently with AI powered systems(Mitra, R. et al, 2023; Shafik, 2024).

Natural language understanding (NLU) plays a pivotal role in interpreting prompts with a high degree of accuracy, moving beyond simple keyword recognition to involve sophisticated analyses of syntax, semantics, and contextual subtleties. This advanced capability enables AI systems to grasp the deeper intent behind user inputs, which is critical for generating responses that are relevant, coherent, and contextually appropriate. By dissecting linguistic elements in detail, NLU effectively bridges the gap between human communication styles and machine processing, facilitating interactions that are more fluid, meaningful, and human like. Furthermore, cognitive load considerations are integral to prompt formulation. Designing prompts that avoid overly complex, ambiguous, or jargon heavy language prevents overwhelming users, thereby facilitating easier comprehension and more efficient exchanges. Striking an optimal balance between clarity, cognitive simplicity, and conversational naturalness ensures that prompts effectively guide AI behavior while maintaining high levels of user engagement, trust, and satisfaction throughout the interaction. This comprehensive approach not only optimizes the AI's responsiveness but also enhances the overall user experience, making AI powered tools more accessible and effective across diverse applications(Romeo & Conti, 2025; Talreja Wassan & Ghuriani, 2024).

Moreover, the application of HCI and NLU principles extends beyond mere prompt construction to influence the broader interaction design of AI systems. For instance, adaptive prompt strategies that respond dynamically to user behavior and feedback can further reduce cognitive effort and improve task completion rates. By continuously analyzing user input patterns and adjusting prompt complexity or style accordingly, AI systems can create personalized experiences that accommodate varying levels of user expertise and preferences. This dynamic adaptability ensures that novice users receive more guided, simplified prompts, while expert users can benefit from more concise and technical interactions, thereby optimizing usability across a diverse user base(Kishore et al., 2024).

Additionally, multimodal interaction considerations—such as integrating visual, auditory, or tactile cues alongside textual prompts—can complement linguistic clarity and foster more naturalistic communication channels. These enhancements leverage the full spectrum of human sensory and cognitive capabilities, further bridging the divide between human users and AI systems. For example, visual aids like icons or progress indicators can reinforce textual instructions, auditory feedback can confirm receipt of commands, and haptic responses can provide subtle cues during interaction. Such multimodal designs contribute to reducing cognitive strain and improving accessibility, particularly for users with diverse needs and preferences (Lee, 2024; Shafik, 2024).

In sum, the synergy of HCI and NLU principles in prompt design is critical for developing AI systems that are not only technically proficient but also user centric. By prioritizing clarity, simplicity, contextual understanding, and cognitive load management, prompt designers enable AI tools to interact with users in ways that feel intuitive, engaging, and efficient. This holistic approach ultimately drives greater adoption, satisfaction, and trust in AI technologies, paving the way for more sophisticated and human aligned artificial intelligence applications in the future. As AI continues to evolve, embedding these principles into design processes will be essential for creating systems that not only respond accurately but also resonate with users on a cognitive and experiential level, fostering long term engagement and effective human AI collaboration(Benlalia et al., 2025; Hasyim & Bakri, 2024).

b. Information Retrieval and Query Formulation

Prompt writing and database querying share fundamental parallels in their approach to extracting relevant information efficiently and effectively. Both processes rely heavily on the formulation of carefully constructed inputs that serve to guide the respective systems toward producing the desired outputs, with a strong emphasis on clarity, precision, and intentionality. In the context of prompt writing, this involves the deliberate crafting of questions or statements designed to direct an AI model to generate responses that are not only accurate but also contextually appropriate and meaningful. Similarly, effective database querying depends on the creation of well formed queries that specify precise criteria to retrieve relevant data from extensive and often complex datasets. Mastery in both domains requires a deep understanding of how to frame requests in a manner that balances the level of detail and the scope of the inquiry, ensuring that the system interprets the input as intended without ambiguity, misdirection, or unintended bias(Zadrozny et al., 2008).

Optimizing information retrieval through prompts or queries involves employing sophisticated techniques that carefully balance specificity and generality to maximize the relevance, usefulness, and comprehensiveness of the results. When prompts or queries are overly specific, they may inadvertently restrict the range of results returned, potentially overlooking valuable related information that falls outside narrowly defined parameters or constraints. This can result in missed opportunities to discover broader insights or connected data points that would enrich understanding or analysis. Conversely, if the inputs are excessively general or vague, the system may produce an overwhelming volume of outputs, many of which may be irrelevant, tangential, or too broad to be actionable. This can lead to inefficiencies in processing and require additional effort to sift through the results to identify pertinent information. Therefore, effective prompt construction requires calibrating the level of detail to suit the particular task at hand, often by including essential context, background information, or constraints that guide the system without unduly limiting its flexibility or creative potential(X. Chen et al., 2020; Phan et al., 2007).

This nuanced balance enhances the system’s ability to generate meaningful, targeted responses or data sets, thereby improving both the efficiency and accuracy of information retrieval processes across AI driven and database environments. By mastering these principles, users can leverage the full potential of both prompt writing and database querying to obtain precise, relevant, and actionable information in a timely manner. Furthermore, understanding the interplay between the structure of the input and the nature of the output allows for iterative refinement, where prompts or queries can be adjusted based on initial results to progressively hone in on the most valuable data. This iterative approach is critical in complex or dynamic contexts where the scope of inquiry may evolve or where initial assumptions need validation(Suresh et al., 2024; Z. Wang et al., 2024).

In addition to balancing specificity and generality, effective prompt writing and database querying also require consideration of the linguistic and syntactic elements that influence system interpretation. For AI prompts, this means using clear, unambiguous language and avoiding jargon or overly complex constructs unless necessary, to ensure that the model’s natural language understanding capabilities are optimally engaged. For database queries, proper syntax, use of operators, and logical structuring are essential to accurately filter and retrieve data. Both domains benefit from an awareness of system limitations and capabilities, enabling users to craft inputs that align with the strengths of the underlying technology while mitigating potential weaknesses or biases(Shlain et al., 2020).

Ultimately, the shared goal of prompt writing and database querying is to facilitate efficient knowledge discovery and decision making by transforming user intent into precise, system readable commands that yield high quality outputs. Mastery of these skills empowers users to navigate vast information landscapes with confidence, extracting insights that drive innovation, research, and informed action(Sykes, 2013).

c. Linguistic Theory and Discourse Analysis

The application of pragmatics and semantics in prompt design is indispensable for ensuring that AI models interpret prompts as intended, thereby enabling the generation of responses that accurately reflect user intentions. Semantics deals primarily with the literal meanings of words and phrases, providing a stable and explicit foundation for understanding the content embedded within a prompt. This semantic foundation allows AI systems to decode the direct, surface level meanings conveyed by language, ensuring that the fundamental message is recognized. Pragmatics, in contrast, extends beyond the literal meanings to incorporate the broader communicative context, including the speaker's intent, situational variables, and discourse cues. This enables interpretation that captures implied meanings, subtleties, and nuances that are not explicitly stated but are essential for correctly understanding and responding to prompts. By integrating semantics and pragmatics, prompt designers can craft inputs that not only communicate explicit information but also implicitly guide AI systems to generate outputs that are coherent, contextually relevant, and aligned with the user's underlying communicative goals(Lo, 2023; Velásquez Henao et al., 2023).

In prompt writing, the role of context and discourse structure is particularly critical, as these elements fundamentally shape how prompts are understood and how potential ambiguities are resolved. Context encompasses the background knowledge, environmental factors, and situational parameters that frame the interpretation of a prompt, while discourse structure governs the organization and flow of information, clarifying relationships among ideas and guiding the interpretation process. Effective prompt design strategically leverages these components to manage ambiguity and polysemy—the occurrence of multiple meanings for a single word or phrase by embedding contextual clues and carefully organizing discourse to minimize confusion. Such strategies may include explicitly providing relevant background information, clearly articulating the speaker's intent, and employing disambiguating expressions or qualifiers that narrow down possible interpretations. These approaches not only enhance clarity but also improve the accuracy and reliability of AI generated responses, ensuring that outputs are closely attuned to the user's expectations and communication objectives. Ultimately, this careful and deliberate application of pragmatics and semantics fosters interactions between users and AI systems that are more natural, meaningful, and contextually appropriate, thereby enhancing the overall effectiveness and usability of AI driven communication tools(Blakemore, 2002; Yarowsky, 1994).

Moreover, the nuanced interplay between semantics and pragmatics in prompt design facilitates the handling of complex linguistic phenomena such as implicature, presupposition, and speech acts, which are crucial for interpreting indirect meanings and intentions. For instance, a prompt may contain an implicit request or suggestion that is not overtly stated but must be inferred from the context and the way the prompt is formulated. By accounting for these pragmatic aspects, prompt designers ensure that AI systems can detect and respond to such subtleties, leading to more sophisticated and human like interactions. Additionally, the dynamic nature of discourse means that prompts often function within larger conversational frameworks, where prior exchanges influence interpretation. Effective prompt design therefore also involves considering the sequential and relational aspects of discourse, allowing AI models to maintain coherence and relevance across multiple turns of interaction(Levinson, 2000; Searle, 1980).

Furthermore, the strategic application of pragmatics and semantics in prompt design contributes to reducing errors arising from lexical ambiguity and syntactic complexity. By carefully selecting vocabulary and constructing sentences that are semantically precise and pragmatically clear, prompt designers can minimize misunderstandings that might otherwise lead to irrelevant or incorrect AI responses. This includes the use of contextually appropriate synonyms, explicit referents, and well defined scope markers, which help disambiguate meaning at both the word and sentence levels. In doing so, prompt designers enhance the interpretability of prompts, enabling AI models to process inputs more effectively and generate outputs that are both accurate and context sensitive. This meticulous approach ultimately supports the development of AI applications that are reliable and user centric, capable of adapting to diverse communicative scenarios and user needs(Velásquez Henao et al., 2023).

III. PRACTICAL APPLICATIONS OF PROMPT WRITING

a. Text Generation and Content Creation

Strategies for developing prompts for creative writing tasks are fundamentally centered on crafting instructions that are open ended, evocative, and sufficiently flexible to inspire originality and diverse narrative outcomes. These prompts often employ vivid, sensory rich language and intriguing scenarios designed to stimulate the writer's imagination, encouraging deep exploration of characters, settings, and plot developments. By incorporating thematic elements or constraints, such prompts provide a guiding framework that balances direction with creative freedom, enabling writers to produce unique and compelling stories while maintaining narrative focus and coherence. Furthermore, effective creative writing prompts may challenge writers to experiment with different perspectives, narrative voices, or temporal structures, thereby fostering innovative storytelling techniques and expanding their expressive range. The inclusion of emotional or philosophical themes can also deepen the engagement, prompting writers to reflect on complex human experiences and societal issues through their narratives(Healey, 2024).

In contrast, prompts designed for technical documentation and reports prioritize clarity, precision, and structured communication to effectively convey complex information. These prompts emphasize identifying the target audience and clearly outlining the objectives of the document, ensuring that the content is organized logically using headings, bullet points, and data visualizations where appropriate. Best practices in this domain include defining the scope of the document to set clear boundaries, maintaining consistency in terminology to avoid ambiguity, and integrating relevant technical standards and detailed information to enhance the accessibility and usability of the documentation for its intended users. Additionally, such prompts often encourage the inclusion of summaries, glossaries, or appendices to support comprehension and provide quick reference points. Attention to formatting and style guidelines is also critical to ensure professional presentation and facilitate efficient communication within technical or specialized fields(M. H. Chen et al., 2013).

Similarly, prompt based summarization and paraphrasing strategies focus on distilling the essential meaning of the original text into concise, fresh language that remains faithful to the source material. Effective prompts in this context encourage isolating key points, removing redundancies, and restructuring sentences to improve readability and flow without altering the intended message. Techniques include focusing on the main ideas, employing synonyms and varied sentence structures, and maintaining the original tone and context, which together ensure that summaries and paraphrases are accurate, coherent, and adaptable for various academic or professional purposes. These strategies also emphasize the importance of preserving the logical relationships between ideas and avoiding distortion or oversimplification, thereby producing outputs that can be confidently used for literature reviews, reports, or presentations. Incorporating iterative review and refinement processes further enhances the quality and precision of the final text, making it suitable for diverse audiences and communication goals(Raundale & Shekhar, 2021).

b. Data Analysis and Decision Support

Designing prompts for extracting insights from large datasets involves the careful construction of precise, targeted queries that enable the identification of meaningful patterns, trends, and relationships within complex and often high dimensional data. This process demands a comprehensive understanding of the dataset's structure, including its variables, data types, and inherent noise, as well as the specific analytical objectives to be achieved. By tailoring prompts to focus on relevant variables and contextual factors, analysts can guide AI or analytical models to explore the data efficiently, thus enhancing both the accuracy and the relevance of the insights extracted. Additionally, well crafted prompts help to mitigate the limitations of AI models, such as biases or overfitting, by clearly defining the scope and constraints of the analysis, which ultimately leads to more robust and actionable findings(Sabeghi et al., 2024).

Formulating prompts for predictive modeling and forecasting centers on designing inputs that effectively direct AI models to generate reliable and valid predictions about future events or behaviors based on historical data. These prompts must integrate key predictive features that capture the underlying drivers of the phenomena being modeled, temporal components that reflect the dynamic nature of the data, and scenario parameters that allow for the exploration of different conditions or assumptions. The inclusion of these elements ensures that the predictive models can learn the complex relationships and temporal dependencies necessary for accurate forecasting. Moreover, prompt design in this context often involves iterative refinement to balance model complexity and generalizability, enabling the generation of predictions that are both precise and interpretable(D Kulkarni, 2024).

Similarly, prompt driven scenario analysis and planning utilize structured prompts to simulate a range of potential future outcomes, providing decision makers with a framework to evaluate risks, uncertainties, and opportunities under varying assumptions. This approach leverages the flexibility of AI models to model complex systems and interactions, allowing for the exploration of “what if” scenarios that inform strategic decision making. By carefully crafting prompts that incorporate relevant variables, constraints, and hypothetical conditions, organizations can use AI driven scenario analysis to anticipate challenges, optimize resource allocation, and develop contingency plans. Together, these approaches demonstrate how prompt engineering serves as a critical tool in maximizing the utility of AI for data driven decision making and strategic planning, enabling users to harness the full potential of large datasets and advanced modeling techniques(Cotroneo & Hutson, 2023; D Kulkarni, 2024).

c. Code Generation and Software Development

Techniques for writing prompts to generate code snippets involve crafting instructions that are clear, concise, and rich in context to guide the AI in producing accurate and relevant code. Effective prompts typically specify the programming language to be used, the desired functionality or purpose of the code, the expected input and output behavior, and any constraints such as performance considerations, library dependencies, or coding style preferences. This level of precision reduces ambiguity and increases the likelihood that the generated code aligns closely with the user’s requirements. Furthermore, including examples or partial code segments within the prompt can help the AI grasp the intended coding style, logic flow, or specific implementation details, thereby refining the output to better meet the user’s expectations. Iteratively refining prompts based on initial outputs also enhances accuracy by providing corrective feedback and clarifying any misunderstood requirements(Li et al., 2024; Tony et al., 2025).

Strategies for using prompts in debugging and error analysis focus on leveraging the AI’s capabilities to identify, explain, and resolve issues within code efficiently. Users should frame prompts to include relevant error messages, descriptions of unexpected behaviors, or problematic code snippets that highlight where the issue occurs. Best practices recommend isolating the error context by providing minimal, reproducible code examples that concentrate on the problematic section, avoiding extraneous or unrelated code. Additionally, prompting the AI to offer stepwise reasoning, alternative approaches, or explanations of underlying concepts helps deepen understanding and facilitates more effective troubleshooting. Including specific questions about potential causes or requesting suggestions for code improvements enables the AI to provide targeted assistance. This approach maximizes the AI’s utility in diagnosing bugs, optimizing code, and improving overall code quality(Rahman et al., 2020; Shah, 2024).

Best practices for prompt based software documentation emphasize generating content that is clear, complete, and user friendly. Prompts should encourage the production of documentation covering essential elements such as function descriptions, parameter explanations, expected return values, usage examples, and potential edge cases or error conditions. Structuring prompts to request these specific documentation components ensures that the output is comprehensive and standardized, which aids in easier understanding, maintenance, and onboarding for new developers. Additionally, tailoring prompts to the intended audience—whether they are developers, end users, or stakeholders—ensures that the documentation meets diverse needs, balancing technical depth with accessibility. Including requests for formatting style, tone, and organization within the prompt further enhances usability and consistency. This method supports better software adoption, collaboration, and long term project sustainability.

d. Natural Language Interfaces and Chatbots

Designing conversational prompts for human like interactions involves a multifaceted approach that ensures dialogue not only feels natural and engaging but also aligns precisely with the user's input and contextual background. This process demands a profound understanding of human communication nuances, including tone modulation, pacing, emotional subtleties, and the intricate variations in language that convey meaning beyond the literal text. Effective prompts are meticulously crafted to anticipate user intentions and needs, guiding conversations in a manner that is both intuitive and seamless. This anticipatory design encourages sustained interaction, fosters rapport, and builds trust—key components for a positive and meaningful user experience. Moreover, these prompts must be flexible and adaptable to diverse user profiles and conversational goals, ensuring that each interaction remains relevant, personalized, and contextually appropriate(Lin et al., 2023; Spillner & Wenig, 2021).

Techniques for managing context and maintaining coherence in dialogues are essential for sustaining meaningful, logical, and fluid exchanges over multiple conversational turns. This involves the continuous tracking and integration of previous messages and user inputs to preserve the conversation's thread, recognizing shifts in topics or changes in user intent, and dynamically adjusting responses to reflect the accumulated information accurately. Maintaining coherence also requires avoiding redundancy and ensuring that responses are contextually relevant, thereby enhancing clarity and user satisfaction. Advanced approaches leverage memory mechanisms and context aware models that enable the system to recall and integrate pertinent past interactions, supporting complex multi turn dialogues that feel purposeful and natural. These mechanisms also help in managing long term dependencies within conversations, allowing the AI to reference earlier points without losing focus or confusing the user(Chhikara et al., 2025; Umirzakova et al., 2024).

Strategies for handling edge cases and unexpected user inputs are critical for building robust conversational AI systems capable of gracefully managing ambiguity, off topic remarks, or erroneous inputs. These strategies include detecting when user inputs deviate from expected patterns or when the system lacks sufficient information to provide an accurate response. In such cases, the system employs fallback mechanisms such as posing clarifying questions, gently steering the conversation back on course, or delivering informative responses that acknowledge the input without causing confusion or frustration. This flexibility is vital for maintaining a resilient and user centric dialogue, preventing breakdowns in communication, and sustaining engagement over time. Additionally, these strategies often incorporate error recovery techniques and adaptive learning components that improve the system's responses based on interaction history and evolving user behavior(Chandra et al., 2022; Chaudhry & Debi, 2024).

Together, these elements—carefully designed prompts, sophisticated context management, and robust handling of edge cases—form the foundation of advanced conversational AI systems. Such systems are capable of sustaining engaging, coherent, and contextually relevant interactions that not only meet user expectations but also adapt dynamically to the complexities of human communication. This comprehensive approach ensures that conversational AI can function effectively across a wide range of applications, from customer service to academic assistance, delivering interactions that feel genuinely human like in both form and function(Ahmed et al., 2024).

IV. EMERGING TRENDS AND FUTURE DIRECTIONS

a. Multi modal Prompt Writing

The integration of text, image, and audio inputs in prompt design signifies a profound evolution in the way AI systems interact with users, enabling communication that is increasingly natural, intuitive, and versatile. By simultaneously leveraging multiple modalities, AI models achieve a more comprehensive and nuanced grasp of user intent and context. This multimodal understanding empowers the generation of responses that are not only more precise but also enriched with contextual details and emotional subtleties. Each modality plays a distinct and complementary role: text conveys explicit semantic content and detailed instructions; images provide essential visual context, spatial relationships, and environmental cues; audio inputs capture prosodic elements such as tone, pitch, rhythm, and emotional inflection. Together, these inputs create a synergistic effect that significantly enhances the interpretative capacity of AI, allowing it to disambiguate potentially vague or incomplete information from any single modality and tailor responses that better align with the user's communicative style, preferences, and situational needs(Liu et al., 2001).

However, the integration of heterogeneous data streams from different modalities introduces a set of complex technical challenges that must be addressed to realize the full potential of multimodal AI systems. One primary difficulty lies in effectively aligning and synchronizing inputs that vary widely in format, temporal dynamics, and resolution. For instance, audio data unfolds over time with fine grained temporal variations, images capture spatial information at high resolution, and text is inherently symbolic and discrete. Developing preprocessing pipelines that normalize these diverse inputs while preserving their unique informational content is essential. Furthermore, input quality can fluctuate significantly—audio may be corrupted by background noise, images might suffer from low resolution or occlusions, and textual input can be ambiguous or incomplete—requiring robust noise handling and uncertainty modeling techniques. Architecturally, designing AI models capable of fusing these disparate signals demands innovative approaches such as cross modal attention mechanisms, joint embedding spaces, and hierarchical fusion strategies that respect both modality specific features and shared semantic representations. These models must learn to balance the influence of each modality dynamically, depending on context and input reliability, to optimize overall interpretive accuracy(Jandoubi & Akhloufi, 2025; P. Wang et al., 2025).

Overcoming these challenges unlocks transformative opportunities across a broad spectrum of applications. Multimodal prompt systems can drive the development of highly immersive and context aware virtual assistants that understand and respond to a richer array of human communication cues, including emotional states and environmental context. In accessibility, such systems offer alternative interaction pathways for individuals with sensory impairments, enabling, for example, visually impaired users to receive detailed auditory descriptions of images or hearing impaired users to interact through text and images seamlessly. Content creation platforms stand to benefit by integrating multimodal inputs to produce sophisticated multimedia outputs that combine narrative text, visuals, and sound in coherent and creative ways. Beyond these, sectors such as education can leverage multimodal AI to provide personalized learning experiences that adapt to students' engagement signals across modalities, healthcare can use multimodal data for more accurate diagnostics and patient monitoring, and entertainment can create interactive experiences that respond dynamically to user inputs across text, voice, and gesture. As AI continues to advance, the fusion of text, image, and audio inputs will be instrumental in bridging the subtle nuances of human communication with machine understanding, catalyzing innovation and expanding the boundaries of human computer interaction across diverse domains(Dritsas et al., 2025; Mitra, R. et al).

b. Prompt Engineering and Optimization Automated

Prompt Engineering and Optimization Automated techniques for prompt refinement harness sophisticated algorithmic processes to systematically analyze, evaluate, and improve prompts used in AI and machine learning systems. These methods begin with the collection of extensive prompt performance data, which is then rigorously examined to identify patterns and trends that correlate with either successful or suboptimal outcomes. Through iterative adjustments to prompt phrasing and structure, these techniques enhance clarity, specificity, and contextual relevance, thereby optimizing the interaction between users and AI models. By minimizing human intervention in this refinement process, automated systems reduce the introduction of subjective biases and errors, resulting in prompts that more consistently elicit accurate, coherent, and context aware responses from AI(Lee, 2024).

Building on these foundational automated methods, machine learning approaches further advance prompt generation and selection by leveraging data driven models trained on vast datasets comprising effective prompt response pairs. These models apply techniques such as supervised learning, where labeled examples guide the system to recognize patterns of high quality prompts, reinforcement learning, which optimizes prompts based on feedback and reward signals, and generative modeling, which creates novel prompts tailored to specific tasks or domains. This dynamic adaptability allows AI systems to predict and generate prompts that maximize desired outcomes across diverse applications, improving both efficiency and effectiveness in AI human communication. Additionally, these machine learning models can continuously learn from new interactions, enabling prompt refinement to evolve in real time and adapt to emerging contexts or user needs(D Kulkarni, 2024; Suresh Babu & Akshara, 2023).

Beyond technical optimization, ethical considerations are integral to the design and deployment of prompt engineering techniques. Ensuring fairness involves actively identifying and mitigating biases that may be embedded in training data or introduced during automated refinement, thereby preventing the perpetuation of stereotypes or discriminatory content. Transparency demands that the processes and criteria used for prompt generation are understandable and auditable by stakeholders, fostering trust in AI systems. Accountability requires mechanisms to monitor and address harmful or misleading outputs resulting from prompt design. Responsible prompt engineering thus balances the pursuit of enhanced AI performance with the imperative to uphold ethical standards, safeguarding against misinformation, bias, and other adverse impacts in AI generated interactions. This holistic approach to prompt refinement ensures that improvements in AI communication are aligned with broader societal values and norms(Maphosa, 2024; Williamson & Prybutok, 2024).

Moreover, the integration of explainable AI techniques within prompt refinement frameworks aids in demystifying how prompts are generated and adjusted. By providing interpretable insights into algorithmic decisions, developers and users can better assess the reliability and fairness of prompts. This transparency supports regulatory compliance and ethical auditing, which are increasingly critical as AI applications expand into sensitive domains such as healthcare, legal services, and education. Furthermore, incorporating stakeholder feedback loops into automated prompt refinement processes enhances the alignment of AI generated content with user expectations and ethical guidelines, promoting responsible innovation(Agu et al., 2024; Belghachi, 2023).

c. Collaborative and Interactive Prompt WritingTools

Collaborative and Interactive Prompt WritingTools and platforms for collaborative prompt design empower multiple users to jointly create, refine, and optimize AI prompts within a shared digital environment, significantly enhancing both creativity and operational efficiency. These platforms commonly incorporate features such as real time editing, which allows contributors to see and respond to changes instantaneously; version control, enabling teams to track modifications, revert to previous iterations, and maintain a clear development history; and integrated feedback mechanisms that facilitate constructive critique and iterative improvement. By supporting these collaborative workflows, such platforms effectively bridge the divide between domain experts who bring subject matter knowledge and AI practitioners who understand model capabilities and limitations. This synergy results in higher quality prompt engineering that is more precisely tailored to the nuances of specific AI models and adaptable to evolving project requirements or user needs(Lee & Suh, 2024).

Techniques for real time prompt refinement based on AI feedback involve an interactive, iterative process where prompts are dynamically adjusted during the course of user AI interaction to enhance the relevance, accuracy, and contextual appropriateness of AI generated responses. This refinement process leverages immediate outputs from the AI, combined with user insights and preferences, to fine tune the prompt's wording, structure, or contextual framing. Often, this is supported by automated suggestions or performance metrics that guide users toward more effective prompt formulations. Over time, these iterative adjustments contribute to the development of robust prompt templates that consistently yield high quality results across a range of use cases. When integrated with collaborative platforms, these real time refinement techniques enable a more fluid and adaptive workflow, where prompt design evolves responsively to ongoing feedback and testing(Mahafdah et al., 2024).

Together, collaborative design tools and real time refinement techniques are shaping a future in which human AI collaboration in prompt writing becomes increasingly interactive, adaptive, and productive. This convergence pushes the boundaries of AI assisted creativity and problem solving by fostering environments where human expertise and AI capabilities complement each other seamlessly. As a result, teams can generate more nuanced, context aware prompts that maximize AI performance, ultimately driving innovation and efficiency in applications ranging from academic research and content creation to complex decision support systems. The continuous interplay between collaborative input and AI feedback is thus poised to transform prompt engineering into a highly iterative, co creative discipline that evolves alongside advances in AI technology(J. F. Chen et al., 2024; Luther et al., 2024).

CONCLUSION

a. Summary of Key Concepts

The theoretical foundations constitute a vital and indispensable framework that supports a comprehensive, detailed, and nuanced understanding of the fundamental principles, mechanisms, and underlying dynamics intrinsic to the subject matter. By systematically synthesizing a broad range of well established theories, conceptual models, and foundational paradigms, these theoretical bases elucidate the essential ideas and constructs that collectively define and shape the field. This synthesis empowers researchers, scholars, and practitioners alike to approach complex, multifaceted problems with a methodical, coherent, and structured perspective. Such a perspective not only enhances analytical rigor and intellectual clarity but also ensures that practical applications and interventions are firmly anchored in rigorously validated, empirically supported knowledge. This anchoring significantly enhances the reliability, validity, and overall effectiveness of applied solutions. Consequently, rooting practical efforts in these theoretical bases fosters the creation and refinement of innovative, contextually adaptable, and methodologically sound solutions. These solutions effectively bridge the divide between abstract conceptualization and tangible, real world challenges, thereby facilitating a productive interplay between theory and practice. This alignment fosters a continuous and dynamic feedback loop whereby theoretical insights inform practical applications, and practical outcomes, in turn, refine, challenge, and expand theoretical understanding, promoting an iterative process of knowledge advancement and refinement(Hasija & Esper, 2022).

Simultaneously, emerging trends underscore the dynamic, fluid, and ever evolving character of the field, which is driven by rapid and transformative technological advancements, methodological refinements, and an increasing emphasis on interdisciplinary collaboration and integration. These trends illuminate promising new pathways and innovative directions that hold considerable potential to broaden the scope, deepen the impact, and diversify the range of current practices and applications. Frequently characterized by the incorporation of state of the art tools, sophisticated data driven methodologies, and integrative, cross disciplinary approaches, these developments enable more precise, scalable, efficient, and effective problem solving strategies. Moreover, they facilitate the exploration of previously uncharted or underexplored areas, encouraging a culture of innovation, experimentation, and diversification within the discipline. Looking forward, future directions emphasize the paramount importance of sustained inquiry, intellectual flexibility, and responsiveness to ongoing change and emerging challenges. They advocate for persistent and adaptive research endeavors aimed at resolving enduring and complex challenges, leveraging novel opportunities, and anticipating shifts in the broader scientific, technological, and societal landscapes. By weaving together these emerging trends with forward thinking, strategic approaches, the field is equipped with a comprehensive, coherent, and strategic roadmap that supports continuous advancement, sustained growth, and long term relevance. This synthesis ensures that theoretical insights and practical applications evolve in a complementary, synergistic, and mutually reinforcing manner, adeptly addressing the shifting demands, complexities, and opportunities characteristic of the discipline’s future trajectory(Bahi et al., 2024).

b. Implications for AI Research and Development

Effective prompt writing significantly enhances AI system performance by enabling clearer, more precise communication between users and AI models. Well crafted prompts serve as a crucial interface, guiding the AI to accurately interpret the user's intent, which reduces ambiguity and minimizes the risk of irrelevant or off target responses. This clarity not only improves the relevance and quality of the generated outputs but also streamlines the interaction process, making AI tools more efficient and user friendly. In academic writing, where precision and contextual understanding are paramount, effective prompt design maximizes the utility of AI capabilities, allowing users to obtain tailored assistance that aligns closely with their specific research questions, writing style, and disciplinary conventions. As prompt engineering continues to evolve, it is becoming an indispensable skill for users aiming to leverage AI tools effectively, ensuring that outputs meet high standards of coherence, clarity, and academic rigor.

Looking ahead, future research directions in prompt writing and engineering are poised to explore several innovative avenues. One key area involves the development of automated methods for optimizing prompt structures, which could enable AI systems to suggest or refine prompts dynamically based on the task at hand, thereby enhancing user experience and output quality. Additionally, adaptive prompt generation tailored to individual user styles promises to personalize AI interactions, accommodating diverse preferences and expertise levels. Context aware prompt refinement, which adjusts prompts in real time based on ongoing dialogue or feedback, represents another promising frontier, fostering more interactive and responsive AI assisted writing workflows. Furthermore, integrating domain specific knowledge into prompt frameworks will be essential for improving the accuracy and depth of AI responses in specialized academic fields. Establishing standardized evaluation metrics for prompt effectiveness will also play a critical role in benchmarking advancements and guiding best practices. Collectively, these research directions will deepen the understanding of human AI collaboration, expanding the practical applications of AI in academic and professional writing environments and ultimately enhancing the quality, efficiency, and accessibility of scholarly communication.

c. Ethical Considerations and Responsible AI Use

Transparency and accountability in prompt design are fundamental to ensuring that AI systems function in an ethical, reliable, and trustworthy manner. Transparent prompt design allows both users and developers to clearly understand how specific inputs shape AI behavior and outputs. This clarity is vital for effectively diagnosing errors, identifying and mitigating biases, and fostering a deeper trust in the AI’s responses. When the mechanisms behind prompt construction are open and comprehensible, it becomes easier to evaluate the fairness and accuracy of AI generated content. Accountability mechanisms complement this transparency by holding prompt creators and deployers responsible for the outcomes of AI interactions. Such responsibility encourages adherence to ethical standards, helps prevent misuse or harmful applications, and ensures that any negative consequences can be addressed appropriately. Together, transparency and accountability uphold the integrity of AI systems, enabling the development of applications that are fair, unbiased, explainable, and aligned with societal values.

Advanced prompt writing techniques carry profound societal implications because they directly influence how AI systems interpret, process, and generate information. Expert prompt engineering can significantly enhance AI’s effectiveness and adaptability across a wide range of fields, from education and healthcare to business and creative industries. However, these powerful techniques also raise critical concerns, such as the potential for manipulation of AI outputs, propagation of misinformation, and the creation of unequal access to sophisticated AI capabilities. The societal impact of these developments includes shifts in communication norms, where AI generated content increasingly shapes public discourse and decision making processes. Moreover, the opacity of advanced prompt techniques can challenge transparency in automated systems, complicating efforts to hold AI accountable. Therefore, it is essential to strike a careful balance between fostering innovation in prompt design and embedding strong ethical considerations. By doing so, society can harness the transformative benefits of AI while minimizing risks related to misuse, bias, and inequity, thus promoting responsible and equitable AI deployment.

References

Agu, E., Efunniyi, C., Adeniran, I., Obiki Osafiele, A., Abhulimen, A., & Osundare, O. (2024). Discussing ethical considerations and solutions for ensuring fairness in AI driven financial services. *International Journal of Frontline Research in Multidisciplinary Studies*, *3*(2), 001–009. https://doi.org/10.56355/ijfrms.2024.3.2.0024

Ahmed, M., Munir, E. U., & Khan, H. U. (2024). Conversational AI: An Explication of Few Shot Learning Problem in Transformers Based Chatbot Systems. *IEEE Transactions on Computational Social Systems*, *11*(2), 1888–1906. https://doi.org/10.1109/tcss.2023.3281492

Bahi, A., Gahi, Y., & Gharib, J. (2024). Integrating Generative AI for Advancing Agile Software Development and Mitigating Project Management Challenges. *International Journal of Advanced Computer Science and Applications*, *15*(3). https://doi.org/10.14569/ijacsa.2024.0150306

Belghachi, M. (2023). A Review on Explainable Artificial Intelligence Methods, Applications, and Challenges. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *11*(4). https://doi.org/10.52549/ijeei.v11i4.5151

Benlalia, Z., Al Kateb, G., Diamanti, K., & Mzili, M. (2025). Reimagining Intelligence: A Comprehensive Review of Human Centric AI Systems and Their Societal and Healthcare Integration. *Mesopotamian Journal of Artificial Intelligence in Healthcare*, *2025*, 173–186. https://doi.org/10.58496/mjaih/2025/017

Blakemore, D. (2002). *Relevance and Linguistic Meaning*. Cambridge University. https://doi.org/10.1017/cbo9780511486456

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To Be or Not to Be …Human? Theorizing the Role of Human Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems*, *39*(4), 969–1005. https://doi.org/10.1080/07421222.2022.2127441

Chaudhry, B. M., & Debi, H. R. (2024). User perceptions and experiences of an AI driven conversational agent for mental health support. *mHealth*, *10*, 22. https://doi.org/10.21037/mhealth 23 55

Chen, J. F., Lin, R., Lin, P. H., & Ni, C. C. (2024). Designing the Future: A Case Study on Human AI Co Innovation. *Creative Education*, *15*(03), 474–494. https://doi.org/10.4236/ce.2024.153028

Chen, M. H., Huang, S. T., Chang, J. S., & Liou, H. C. (2013). Developing a corpus based paraphrase tool to improve EFL learners' writing skills.

Computer Assisted Language Learning, *28*(1), 22–40. https://doi.org/10.1080/09588221.2013.783873

Chen, X., Wang, J., & Ding, S. X. (2020). Complex System Monitoring Based on Distributed Least Squares Method. *IEEE Transactions on Automation Science and Engineering*, *18*(4), 1892–1900. https://doi.org/10.1109/tase.2020.3022924

Chhikara, P., Aryan, S., Singh, T., Khant, D., & Yadav, D. (2025). *Mem0: Building Production Ready AI Agents with Scalable Long Term Memory*. Institut F R Ost Und S Dosteuropaforschung. https://doi.org/10.3233/faia251160

Cotroneo, P., & Hutson, J. (2023). Generative AI tools in art education: Exploring prompt engineering and iterative processes for enhanced creativity. *Metaverse*, *4*(1), 14. https://doi.org/10.54517/m.v4i1.2164

D Kulkarni, N. (2024). Crafting Effective Prompts: Enhancing AI Performance through Structured Input Design. *JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING*, *12*(5), 1–10. https://doi.org/10.70589/jrtcse.2024.5.1

Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555. https://doi.org/10.1016/j.artint.2021.103555

Dritsas, E., Mylonas, P., Trigka, M., & Troussas, C. (2025). Multimodal Interaction, Interfaces, and Communication: A Survey. *Multimodal Technologies and Interaction*, *9*(1), 6. https://doi.org/10.3390/mti9010006

Flores Romero, P., Fung, K. N. N., Rong, G., & Cowley, B. U. (2025). Structured human LLM interaction design reveals exploration and exploitation dynamics in higher education content generation. *Npj Science of Learning*, *10*(1). https://doi.org/10.1038/s41539 025 00332 3

Mitra, R. et al. (2023) Learning from data with structured missingness. Nature Machine Intelligence, 5, pp. 13 23. (doi: 10.1038/s42256 022 00596 z)

Swathi. H., & Veeramanju, K. T. (2024). Predictive Models for Optimal Irrigation Scheduling and Water Management: A Review of AI and ML Approaches. *International Journal of Management, Technology, and Social Sciences*, 94–110. https://doi.org/10.47992/ijmts.2581.6012.0346

Hasyim, H., & Bakri, M. (2024). Advancements in Human Computer Interaction: A Review of Recent Research. *Advances: Jurnal Ekonomi & Bisnis*, *2*(4). https://doi.org/10.60079/ajeb.v2i4.327

Healey, B. (2024). Teaching an Embodied Thinking Process for Narrative Writing 'With Grammar in Mind': An Analysis of Writing Conferences and Teacher Talk. *Literacy Research and Instruction*, *64*(2), 203–228. https://doi.org/10.1080/19388071.2024.2345898

Jandoubi, B., & Akhloufi, M. A. (2025). Multimodal Artificial Intelligence in Medical Diagnostics. *Information*, *16*(7), 591. https://doi.org/10.3390/info16070591

Kishore, H. S., Solomon Paul Raj, D., & Senthilkumar, K. R. (2024). *AI for Accessibility* (pp. 214–225). Igi Global. https://doi.org/10.4018/979 8 3693 5593 0.ch015

Lee, C. P. (2024). *Design, Development, and Deployment of Context Adaptive AI Systems for Enhanced User Adoption*. 1–5. https://doi.org/10.1145/3613905.3638195

Lee, J., & Suh, S. (2024). AI Technology Integrated Education Model for Empowering Fashion Design Ideation. *Sustainability*, *16*(17), 7262. https://doi.org/10.3390/su16177262

Levinson, S. C. (2000). *Presumptive Meanings*. Massachusetts Institute Of Technology. https://doi.org/10.7551/mitpress/5526.001.0001

Li, Y., Shi, J., & Zhang, Z. (2024). An Approach for Rapid Source Code Development Based on ChatGPT and Prompt Engineering. *IEEE Access*, *12*, 53074–53087. https://doi.org/10.1109/access.2024.3385682

Lin, C. C., Yang, S. J. H., & Huang, A. Y. Q. (2023). A Review of AI Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability*, *15*(5), 4012. https://doi.org/10.3390/su15054012

Liu, J., Pastoor, S., Hurtienne, J., & Seifert, K. (2001). *Three dimensional PC: toward novel forms of human computer interaction*. *10298*, 102980C. https://doi.org/10.1117/12.419783

Lo, L. S. (2023). The Art and Science of Prompt Engineering: A New Literacy in the Information Age. *Internet Reference Services Quarterly*, *27*(4), 203–210. https://doi.org/10.1080/10875301.2023.2227621

Luther, T., Cress, U., & Kimmerle, J. (2024). Teaming Up with an AI: Exploring Human–AI Collaboration in a Writing Scenario with ChatGPT. *AI*, *5*(3), 1357–1376. https://doi.org/10.3390/ai5030065

Mahafdah, R., Bouallegue, S., & Bouallegue, R. (2024). Enhancing e learning through AI: advanced techniques for optimizing student performance. *PeerJ Computer Science*, *10*(1), e2576. https://doi.org/10.7717/peerj cs.2576

Maphosa, V. (2024). The Rise of Artificial Intelligence and Emerging Ethical and Social Concerns. *AI, Computer Science and Robotics Technology*, *3*. https://doi.org/10.5772/acrt.20240020

Molla, M. (2024). AI in Creative Arts: Advancements and Innovations in Artificial Intelligence. *International Journal of Advanced Research in Science, Communication and Technology*, 513–517. https://doi.org/10.48175/ijarsct 19163

Park, D., An, G. T., Kim, C. G., & Kamyod, C. (2024). A Study on Performance Improvement of Prompt Engineering for Generative AI with a Large Language Model. *Journal of Web Engineering*, 1187–1206. https://doi.org/10.13052/jwe1540 9589.2285

Park, J., & Choo, S. (2024). Generative AI Prompt Engineering for Educators: Practical Strategies. *Journal of Special Education Technology*, *40*(3), 411–417. https://doi.org/10.1177/01626434241298954

Phan, N., Wilkinson, R., & Bailey, P. (2007). *Understanding the relationship of information need specificity to search query length*. 709–710. https://doi.org/10.1145/1277741.1277870

R, P., Rathika, S., Sanjaya, K., Venkatesh, S., Hussein Alawadi, A., Rajalakshmi, B., & Makhzuna, K. (2023). Human Computer Interaction: Enhancing User Experience in Interactive Systems. *E3S Web of Conferences*, *399*, 04037. https://doi.org/10.1051/e3sconf/202339904037

Rahman, M. M., Watanobe, Y., & Nakamura, K. (2020). Source Code Assessment and Classification Based on Estimated Error Probability Using Attentive LSTM Language Model and Its Application in Programming Education. *Applied Sciences*, *10*(8), 2973. https://doi.org/10.3390/app10082973

Raundale, P., & Shekhar, H. (2021). *Analytical study of Text Summarization Techniques*. 1–4. https://doi.org/10.1109/asiancon51346.2021.9544804

Romeo, G., & Conti, D. (2025). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. AI & Society. https://doi.org/10.1007/s00146 025 02422 7

Sabeghi, P., Castaneda, G. D. R., Fields, B. K. K., Patel, D. B., Kinkar, K. K., Varghese, B. A., Eibschutz, L. S., & Gholamrezanezhad, A. (2024). Artificial intelligence and machine learning applications for the imaging of bone and soft tissue tumors. *Frontiers in Radiology*, *4*. https://doi.org/10.3389/fradi.2024.1332535

Searle, J. R. (1980). *Speech Act Theory and Pragmatics*. Springer Netherlands. https://doi.org/10.1007/978 94 009 8964 1

Shafik, W. (2024). *Human Computer Interaction (HCI) Technologies in Socially Enabled Artificial Intelligence* (pp. 121–150). Igi Global. https://doi.org/10.4018/979 8 3693 5533 6.ch005

Shah, J. K. (2024). Explainable AI In Software Engineering: Enhancing Developer AI Collaboration. *The American Journal of Engineering and Technology*, *06*(07), 99–108. https://doi.org/10.37547/tajet/volume06issue07 11

Shlain, M., Goldberg, Y., Taub Tabib, H., & Sadde, S. (2020). *Syntactic Search by Example*. 17–23. https://doi.org/10.18653/v1/2020.acl demos.3

Spillner, L., & Wenig, N. (2021). *Talk to Me on My Level – Linguistic Alignment for Chatbots*. 1–12. https://doi.org/10.1145/3447526.3472050

Suresh Babu, C. V., & Akshara, P. M. (2023). *Revolutionizing Conversational AI* (pp. 228–248). Igi Global. https://doi.org/10.4018/979 8 3693 0502 7.ch011

Suresh, K., Kackar, N., Schleck, L., & Fanelli, C. (2024). Towards a RAG based summarization for the Electron Ion Collider. *Journal of Instrumentation*, *19*(07), C07006. https://doi.org/10.1088/1748 0221/19/07/c07006

Sykes, J. M. (2013). *Technology in Interlanguage Pragmatics Research and Teaching*. John Benjamins Company. https://doi.org/10.1075/lllt.36

Talreja Wassan, J., & Ghuriani, V. (2024). *Perspective Chapter: Recent Trends in Deep Learning for Conversational AI*. Intechopen. https://doi.org/10.5772/intechopen.113250

Tony, C., Mutas, M., Scandariato, R., Dhif, S., & Díaz Ferreyra, N. E. (2025). Prompting Techniques for Secure Code Generation: A Systematic Investigation. *ACM Transactions on Software Engineering and Methodology*, *34*(8), 1–53. https://doi.org/10.1145/3722108

Torricelli, M., Martino, M., Baronchelli, A., & Aiello, L. M. (2024). *The Role of Interface Design on Prompt mediated Creativity in Generative AI*. *4*, 235–240. https://doi.org/10.1145/3614419.3644000

Umirzakova, S., Muksimova, S., Mardieva, S., Sultanov Baxtiyarovich, M., & Cho, Y. I. (2024). MIRA CAP: Memory Integrated Retrieval Augmented Captioning for State of the Art Image and Video Captioning. *Sensors (Basel, Switzerland)*, *24*(24), 8013. https://doi.org/10.3390/s24248013

Velásquez Henao, J. D., Franco Cardona, C. J., & Cadavid Higuita, L. (2023). Prompt Engineering: a methodology for optimizing interactions with AI Language Models in the field of engineering. *DYNA*, *90*(230), 9–17. https://doi.org/10.15446/dyna.v90n230.111700

Wang, P., Chen, T., Zhou, Q., Hu, J., & Wu, Y. (2025). DLF: Disentangled Language Focused Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39*(20), 21180–21188. https://doi.org/10.1609/aaai.v39i20.35416

Wang, Z., Shi, W., & Gan, B. (2024). *Multimodal Query Suggestion with Multi Agent Reinforcement Learning from Human Feedback*. *33*, 1374–1385. https://doi.org/10.1145/3589334.3645365

Williamson, S. M., & Prybutok, V. (2024). The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information*, *15*(6), 299. https://doi.org/10.3390/info15060299

Yarowsky, D. (1994). *Decision lists for lexical ambiguity resolution*. 88–95. https://doi.org/10.3115/981732.981745

Zadrozny, S., De Tré, G., Kacprzyk, J., & De Caluwe, R. (2008). *An Overview of Fuzzy Approaches to Flexible Database Querying* (pp. 34–54). Igi Global. https://doi.org/10.4018/978 1 59904 853 6.ch002

Chapter 14

Synthesis of Copper Incorporated Porous Clay Heterostructures (PCH) and FTIR Study

Funda Turgut BASOGLU[1]

Abstract

In the field of nanomaterials, copper-supported porous clay heterostructures (PCH) stand out due to their distinctive properties and versatile applications. This study focuses on the synthesis and characterization of PCH, particularly through FTIR analysis. The determination of the functional groups in the material structure and the changes caused by metal loading on the functional groups and surface acid centers were interpreted with FTIR (Fourier Transform Infrared Spectroscopy) spectra. The addition of cerium along with copper to the structure appears to increase and change the vibration properties of the clay matrix, especially in regions associated with metal-oxygen interactions. This may be due to a synergistic effect in which cerium alters the electronic environment of copper, leading to changes in bond lengths, bond angles, or types of bonds formed. Titanium addition also alters the structural and chemical environment of the clay matrix, affecting interactions and formations typically associated with copper. This can have effects on the physical properties and chemical stability of the material. These changes are crucial for applications where the electronic and structural properties of the material are critical, such as catalysis or electronic materials. After pyridine adsorption and desorption at room temperature, it was obtained that the bands attributed to pyridine coordinately bound to Lewis sites (1446 cm^{-1}) Lewis and Brønsted sites (1491 cm^{-1}) and Brønsted sites (1546 cm^{-1}) were observed for all copper supported PCH samples.

Keywords: Porous Clays Heterostructures (PCH), Copper, FTIR Analysis, Synthesis, Characterization

---

[1] Gazi University, Engineering Faculty, Chemical Engineering Department,
Ankara, Türkiye, tfunda@gazi.edu.tr, ORCID No: 0000-0002-9874-7393.

1. INTRODUCTION

In the field of nanomaterials, copper-supported porous clay heterostructures (PCH) are notable for their unique properties and diverse applications. This study emphasizes the synthesis and characterization of PCH, with a particular focus on FTIR analysis to examine structural features and thermal stability. This research highlights the practical applications of PCH, advocating for their use in improving catalytic performance and environmental cleanup. By building on previous advancements in porous materials and characterization methods, it offers guidance for utilizing copper-supported porous clay heterostructures for a sustainable future (Ishizaki et. al., 1998)

1.1. Porous Material

Porous materials are solids that contain pores. Porosity, which generally ranges from 20% to 95%, refers to the ratio of pore volume to total volume. These materials are used in various applications, from water purification to industrial filtration. Pores can be classified as open or closed. Open pores connect to the outside and are preferred in industrial applications. In porous materials, open pores reduce density and increase specific surface area (Ishizaki et. al., 1998). Different applications require different pore sizes. Atomic-scale pores are suitable for gas separation and catalysis, while micron-scale pores are used for water filtration. Recently, porous ceramics with high thermal stability and corrosion resistance have been developed. Porous materials can be classified by pore size, shape, and material. According to IUPAC, they are divided into microporous (< 2 nm), mesoporous (2-50 nm), and macroporous (> 50 nm). Porous materials are used in gas separation, catalysis, fuel cells, and environmental cleanup (Ishizaki et. al., 1998).

1.2. Clay and Pillared Clay

Pillared clays, also known as pillared interlayered clays, are a class of materials that have garnered significant attention in the fields of materials science and catalysis. These materials are created by inserting inorganic pillars between the layers of clay minerals. This process leads to the expansion of the interlayer spacing and the formation of a porous structure. The pillars, typically composed of metal hydroxides or oxides, prevent the restacking of the clay layers, thus maintaining the expanded structure and creating a high-surface-area material with enhanced porosity (Schoonheydt et al., 1999). The unique structural characteristics of pillared clays make them highly versatile and suitable for a wide range of applications. One of the key properties of pillared clays is their large surface area, which provides numerous active sites for various chemical reactions. This property makes them valuable as catalysts in numerous industrial processes, including hydrocarbon conversion, fine chemical synthesis, and environmental remediation. Additionally, the increased porosity of pillared clays allows them to effectively adsorb and remove pollutants from air and water, making them useful for environmental protection and purification applications (Pires and Carvalho, 2022).

1.3. Porous Clay Heterostructures (PCH)

Porous Clay Heterostructures (PCH) refer to materials that consist of a combination of different types of clay with porous properties. These heterostructures are designed to have specific functionalities, such as adsorption or catalytic properties. They are of interest for various applications, including environmental remediation, catalysis, and nanotechnology. The synthesis and characterization of such materials have been reported in scientific studies, highlighting their potential for use in adsorption and detoxification processes (Jaramillo-Fierro et al., 2020). The use of clays in general, including porous clay heterostructures, is motivated by their abundance, cost effectiveness, and eco-friendly nature (Mao et al., 2022).

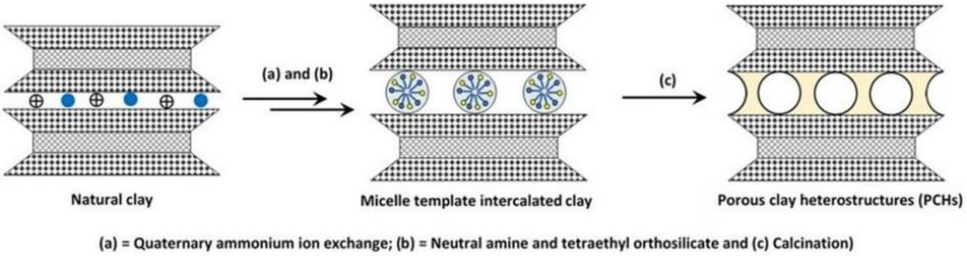


Figure 1. Schematic representation of porous clay heterostructures (PCH)s (Nagendrappa and Chowreddy, 2021)

Porous Clay Heterostructures (PCH) are produced through the insertion of an organic bulky cation into the interlayer spacing of a smectite, a type of clay mineral. This process leads to the formation of a layered material with a hierarchical porous structure, consisting of both mesopores and micropores, which provides a large surface area for adsorption and catalysis (Cecilia et al., 2018). The synthesis of PCH has been the subject of various scientific studies, highlighting the diverse range of methods and modifications used to tailor their properties for specific applications (Polverejan et al., 2002).

1.4. Literature Survey

Clay materials offer unique properties, but researchers are looking for ways to improve them. One approach is to incorporate copper (Cu) into the clay structure and form composites called pillared clay heterostructures (PCHs). Literature studies on Cu-PCH studies are summarized below.

Guimarães et al., (2023) aimed at evaluating strategies for the adsorption and recovery of anthocyanins present in purple cabbage using natural and modified clays as adsorbent. In the batch adsorption experiments, the anthocyanin extracts were put in contact with the adsorbents, and different parameters were evaluated to determine the best conditions for their adsorption and recovery. It was noted that the highest levels of adsorption (28.0mg g^{-1}) occurred using a porous clay heterostructure (PCH) material as adsorbent, with a mass of 25mg and 120min of contact. Under the same conditions, the sepiolite only presented an adsorption capacity of 14.0mg g^{-1}. The desorption results showed that the 60% methanolic solution recovered 60% of the anthocyanins adsorbed on PCH, while the 80% ethanolic solution recovered 35% of those adsorbed on sepiolite. The eluted anthocyanin solutions showed a 98% lower sugar concentration than the crude extract, indicating the low affinity of the adsorbents for sugars. Six types of acylated cyanidins were identified via UPLC-QToF-MSE in the extract, and it was confirmed from the FTIR analyses that the highest affinity of the clays occurred with the anthocyanins that presented more organic acid in their structure. The results show that PCH and sepiolite have high selectivity for anthocyanins and low affinity for the sugars present in the plant extract, facilitating the process of partial purification and application of these pigments.

In Kashif et al., (2022) study, the impregnation method was used to prepare a new Mn/GaPCH catalyst by using montmorillonite as a host clay for the synthesis of porous clay heterostructures (PCH) as catalyst support, and low-temperature selective catalytic reduction (SCR) of NO by C_3H_6 over PCH-based catalyst was explored. The effect of the Mn and Ga- porous clay heterostructures supported catalysts were studied by ICP, XRD, SEM, N_2 adsorption-desorption, XPS, and Py-FTIR. The 8Mn/3Ga-PCHs (8 wt% Mn and 24 wt% Ga loading) catalyst exhibited 95% NO conversion at 300 °C and noticed that above 70% conversion of NO was achieved at low-temperature (150–250 °C). The active Mn metal addition resulted in high NO conversion at low temperature for the 8Mn/3Ga-PCH catalyst due to more surface acidic sites, amorphous Ga and Mn formation, and Mn^{4+} and Mn^{3+} ratio. In addition, the catalytic activity sequence is consistent with Mn^{3+}/Mn^{4+}, implying that high-valence manganese may represent the active sites for NO conversion at low temperature.

In Muñoz et al., (2023) study, nickel salts of Keggin heteropolytungstates with the general formula NixAyW12−yO39or40 (A = Si/P) were synthesized and studied as bulk catalytic materials or supported ones by deposition on modified and functionalized clay minerals (pillared layered clay and porous clay hetero structure). Characterizations by Raman, ^{31}P and ^{29}Si-NMR, and ESEM-EDS techniques showed that pure and supported systems preserved the Ni/W ratio and the expected structural properties of heteropolyanions. These materials were evaluated as catalysts in the selective oxidation of sulfides to sulfoxides or sulfones, using aqueous hydrogen peroxide and mild reaction conditions. The bulk materials, with a higher content of Ni, displayed a remarkable catalytic behavior in the oxidation of diphenyl sulfide ($Ni_3PW_{11}NiO_4OH$, 90% conversion in 15 min at 75°C, 100% sulfone selectivity in 3h). Supported catalysts, particularly the non-functionalized PCH ($Ni_2SW_{12}O_{40}$/PCH), showed excellent activity, with also being selective in the oxidation of sulfide to sulfoxide (87% conversion, 88.9% sulfoxide selectivity). The reuse of these materials was studied in the optimum reaction conditions, resulting in similar activity and selectivity.

The study conducted by Nicola et al.,(2022), in order to obtain linear alpha-olefins (LAOs) which are high-value intermediates for fine chemicals and petrochemicals, olefin oligomerization was applied. The use of heterogenized catalytic systems based on anchoring of organometallic precursors on inorganic materials has the advantages of increasing the selectivity of the reaction and the possibility of reusing the catalyst. Porous clay nanostructures with bi (2D) and three-dimensional (3D) structures were synthesized to act as catalytic supports. A Ni-β-diimine organometallic complex was anchored to these materials giving rise to heterogenized nickel complexes that were tested as catalytic precursors for the ethylene oligomerization reaction, and they were compared with the analogous complex in homogeneous medium. The results showed the influence of the heterogenization of the materials on the thermal stability of the catalyst and the selectivity of the products. Also, the results demonstrated a noticeable influence of the different structures of the materials on catalytic activity. The best result was obtained with the catalyst anchored on the support with a 3D structure, presenting a T.O.F. (turnover frequency) of 754×10^3 h^{-1} . Finally, this material was subjected to recycling tests that qualitatively showed that the catalyst can be reused while maintaining a high selectivity of 95% for alpha-olefins.

In Sato et al., (2022) study, embedding functionalized mesopores in silica materials is a straightforward approach for the design of adsorption materials suitable for volatile compounds such as acetaldehyde molecules. In the study, mesopores with protonated interior surfaces for porous clay nanoheterostructures (PCHs) originating from cost effective and eco-friendly smectites were synthesized. The textural properties of both micro- and mesopores in the PCHs were analyzed by N_2 gas adsorption/desorption at −196 °C. Their chemical information was complemented by open-space analyses using positronium. The protonated mesopore surfaces enhanced the adsorption performance of acetaldehyde molecules through hydrogen bonds despite the reduction in mesoporosity. An approach to the tunability of protonation on the mesopore surfaces in PCHs is presented, focusing on the layer charge of base smectites caused by a few compositional percentage modifications.

In Chmielarz et al., (2018) study, montmorillonites intercalated with silica, silica-alumina, silica-titania and silica-zirconia pillars by surfactant directed method were studied as catalysts for dehydration of methanol and ethanol to dimethyl ether as well as diethyl ether and ethylene, respectively. Moreover, the samples of montmorillonite intercalated with silica pillars were doped with aluminum using template ion-exchange method. In the case of all the samples deposited metal (Al, Ti, Zr) species were present in highly dispersed forms and resulted in generation of acid sites catalytically active in dehydration of methanol and ethanol. The best catalytic results in both studied processes were obtained for the samples doped with aluminum, especially introduced by using template ion-exchange method (first time applied for deposition of metal cations into PCHs), which were more catalytically active in the process of methanol to dimethyl ether conversion than γ-Al_2O_3 (one of the most important commercial catalysts of this process). High catalytic activity of the aluminum doped samples was related to the nature and strength of acid sites active in the processes of alcohols dehydration.

In Besghaier et al., (2021) study, the synthesis of PCH from natural bentonite produces a porous heterostructure material effective for the adsorption of glyphosate from water. The adsorption process takes place through an interaction between the silanol group of montmorillonites and/or the PCH adsorbent with the functional groups of glyphosates. The glyphosate adsorption isotherms, recorded for all the studied samples, have been established to be of Langmuir type. The kinetic of the herbicide adsorption on the PCH was best described by the pseudo-second-order model. With the increase in temperature from 25 to 50 °C, the sorption capacities of the materials studied towards glyphosate increased. The process of glyphosate adsorption was found to be endothermic and spontaneous in nature, as indicated by positive values of ΔH and negative values of ΔG. According to the results obtained, the herbicide sorption was more effective in a basic environment. The maximum amount of adsorbed glyphosate is almost doubled with PCH from 13.5 mg/g of natural clay to 27.5 mg/g of PCH.

In Essih et al., (2022) study, the layered structure of two smectites (montmorillonite and saponite) are modified by the insertion of SiO_2–ZrO_2 nanoparticles, with a Si/Zr molar ratio of 5, to form porous clay heterostructures

(PCHs). These PCHs exhibit a clear improvement of the textural and acid properties in comparison to the starting clays, due to the formation of a pillared structure and the presence of Zr species, which provide an increase in the amount of Lewis acid sites. These materials are studied in one-pot catalytic processes to transform furfural into valuable products, such as furfuryl alcohol, alkyl furfuryl ethers, alkyl levuli nates or γ-valerolactone, depending on the experimental conditions. Thus, the use of milder reaction temperature (110 °C) favors the formation of furfuryl alcohol and alkyl furfuryl ethers, while higher temperatures (170 °C) promote the formation of alkyl levulinate and γ-valerolactone, as well as an increase in nondetected products associated to the formation of humins due to the polymerization of furfural and/or furfuryl alcohol.

In Chmielarz et al., (2007) another study, porous clay heterostructures (PCH), obtained on the basis of synthetic saponite, were modified with copper and iron ions by an ion-exchange method and tested as catalysts for the selective reduction of NO with ammonia (NH_3- SCR). Transition metal ions were introduced into H^+ and NH_4^+ forms of PCH. The catalysts were characterized with respect to their texture (BET), composition (EPMA), coordination of transition metals (UV–vis-DRS) and surface acidity (FT-IR, NH_3- TPD). The PCH-based catalysts have been found to be active, selective and stable in the NH_3-SCR process. The Cu-containing catalysts effectively operated at temperatures significantly lower than the Fe-modified samples. The catalyst obtained by exchanging copper ions in the ammonium form of PCH was considerably more active compared to the sample obtained from H^+ -PCH.

In Yuan et al., (2019) study, selective catalytic reduction (SCR) of NO with C_3H_6 was investigated on porous clay heterostructures modified with $CuFe_2O_4$ (CuFe-PCH). The catalysts were characterized by XRD, HRTEM, H_2-TPR, XPS, Py-FTIR and in situ DRIFTS technologies. CuFe-PCH showed high SCR activity at low temperature. Among all the catalysts in this study, Cu_1Fe-PCH with the molar ratio of Cu:Fe=1:1 exhibited the best catalytic performance, which reached 58.5% NO removal at 300°C (0.1%C_3H_6) and 100% NO removal at 350°C (0.3%C_3H_6) respectively. The $CuFe_2O_4$ species on Cu_1Fe-PCH had high crystallinity, good redox ability, large proportion of lattice oxygen, high density of acid sites and strong Brönsted acidity, which were all responsible for the highest C_3H_6-SCR activity. By in situ DRIFTS, more isocyanates, including Cu-NCO and Fe-NCO, were found on Cu_1Fe-PCH, which were beneficial to enhance C_3H_6-SCR activity. Finally, a possible reaction path was proposed.

In Chmielarz et al., (2011) study, natural montmorillonite was used as a raw material for the preparation of titania pillared clays (Ti-PILC) and porous clay heterostructures (Ti-PCH). Ti-PILC was produced by direct exchange of interlayer sodium cations in montmorillonite (sodium form) for titanium oligocations. Ti-PCH was obtained by the surfactant directed method. Tetraethylorthosilicate and titanium isopropoxide were used as the source of silica and titania, respectively. Copper and iron were introduced to Ti-PILC and Ti-PCH by the ion exchange method. The obtained materials were characterized with respect to chemical composition (EPMA), structure (XRD) and texture (BET), coordination and aggregation of transition metal species (UV–vis-DRS) and surface acidity (FTIR). Modified montmorillonites were found to be active and selective catalysts for the reduction of NO with ammonia (DeNOx process) and their catalytic performance depended on the type of support (Ti-PILCor Ti-PCH) as well as deposited transition metal (Cu or Fe) species.

In Chmielarz et al., (2006) study, porous clay heterostructures (PCH) obtained on the basis of the synthetic saponite was used as a catalytic support for the deposition of transition metal (Cu, Fe) oxides. Transition metal oxides were introduced into the parent as well as ammonia treated PCH support by the ion-exchange method. The obtained catalysts were characterized with respect to their textural parameters (BET), composition (EPMA), chemical nature of the introduced transition metal species (UV–vis–DRS) and surface acidity (NH_3-TPD, FTIR). The modified PCH materials were tested as catalysts for the selective oxidation of ammonia to nitrogen and water. The studied catalysts have shown a very high selectivity towards N_2, which did not drop below 90% in a large, studied temperature range (150–550 °C) However, the Cu-containing samples were active at temperatures significantly lower than PCH modified with iron.

2. EXPERIMENTAL STUDY

2.1. PCH Synthesis

In the synthesis, hexadecyl trimethyl ammonium chloride (HTAC) was used as a common surfactant, tetra ethyl orthosilicate (TEOS) as a silica source, and ammonia solution as a base source. Octylamine was used as a cosurfactant for micellar formation and silica interaction of the surfactant. The studies were based on the SPC (Silica Pillared Clay) synthesis recipe of Dincer et al., (2020). In the recipe prepared according to the determined amounts of substances, 1 gram of clay was suspended in 100 ml of distilled water (1% by mass) and left at room temperature overnight. For 1 g of clay, 1.1 ml of HTAC was added to the clay suspension and mixed at room temperature in a magnetic stirrer for 8 hours. Then, it was filtered and washed to remove the chlorine ions in it. The obtained modified clay (Q+-MMT) was dried at room temperature and then kept in an oven at 100 °C for 24 hours. First of all, after mixing homogeneously in some distilled water, octylamine was added to 1 gram of Q+-MMT and mixed in a magnetic stirrer for 15 minutes. Then, TEOS and copper source Cu $(NO_3)_2$*$3H_2O$ were mixed and added. Final solution was mixed at room temperature in a magnetic stirrer for 8 hours. The resulting product was dried at room temperature and treated with ammonia solution. It was kept at room temperature for one day. Then, it was left to mix under cover for 24 hours at 38 °C. The product was filtered and dried at room temperature. Then, the obtained sample was kept in an oven at 100 °C overnight. Then, SPC structures were obtained by calcining in a tube furnace at 550 °C with a heating rate of 1 °C/min for 6 hours under 60 cm^3/min air flow. While the synthesis recipe is compatible with the recipe used by B. Yeter Dinçer et al. (2020), the addition of the metal to the structure was carried out simultaneously with the silica welding. In copper-loaded SPC syntheses, the metal/silicon molar ratio Cu/Si: 0.04 was used. In the in-situ synthesis method, Silica/clay molar ratio is 50. The sample was obtained by determining the silica/octylamine molar ratio as 2.5 (Dincer et al., 2020).

2.2. FTIR Analysis

FTIR analysis of clay mineral and synthesized SPC structures were carried out. The determination of the functional groups in the material structure and the changes caused by metal loading on the functional groups and surface acid centers were interpreted with FTIR (Fourier Transform Infrared Spectroscopy) spectra.

ATR (Attenuated Total Reflectance) of pure samples in the range of 400-4000 cm^{-1} wavenumbers and DRIFT (Diffuse Reflectance Infrared) of powder samples diluted to 1g sample/100mg KBr in the Bruker Vertex 70/70v FT-IR Spectrophotometer Device were obtained. Transmittance peaks were monitored using Fourier Transform. Samples were kept in the oven at 100 °C for 1 night to remove moisture before measurement. Bronsted and Lewis acid centers were evaluated with the help of spectra taken after pyridine adsorption/desorption. The samples were kept in pyridine vapor in a fume hood for a 1 week. Then, spectra were taken at room temperature, 100°C, 150°C, 200°C and 250°C.

3. RESULTS AND DISCUSSION

The FTIR spectrum taken with the ATR unit for copper added porous clay heterostructure (Cu-PCH) is given below (Figure 2). Several different peaks are shown, each corresponding to different types of molecular vibrations and therefore different bonds. A detailed analysis of the important peaks and their possible missions is mentioned below.

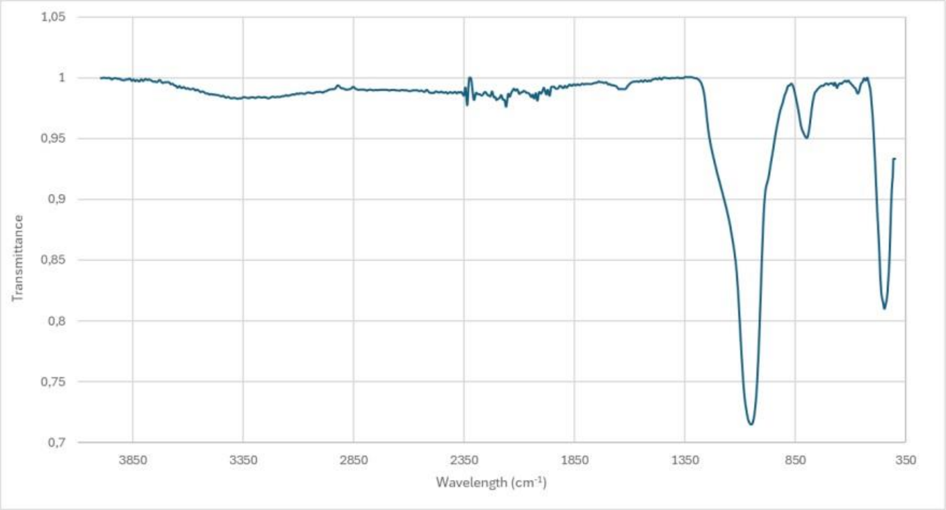


Figure 2. FTIR spectrum of Cu-PCH.

Peaks around 3620 cm^{-1} is typically indicative of O-H stretching vibrations commonly found in hydroxyl groups. In clays, these groups are often part of the structural water or hydroxyl units within the clay minerals. Broad and somewhat intense band around 3400 cm^{-1} is characteristic of O-H stretching vibrations from water molecules. This indicates the presence of water in the clay, either as moisture or as part of the structure of the clay. Peak around 1630 cm^{-1} is generally attributed to the bending vibrations of water (H-O-H bending). It confirms the presence of water in the clay matrix, which could be interlayer water or water adsorbed on the surface. Peaks around 1030 cm^{-1} and 1000 cm^{-1} are typically associated with Si-O stretching vibrations in silicates. This is a common feature in clay, indicating the silicate layers that are fundamental to the clay structure. Peak around 620 cm^{-1} can be attributed to the bending modes of metal–oxygen bonds, possibly Cu–O, given the addition of copper to the clay structure. This suggests that copper interacts with oxygen in the clay matrix, potentially forming copper oxide species or being incorporated into the clay lattice. Peak around 470 cm^{-1} is probably related to Si-O bending or rocking modes in the silicate structure of the clay. The presence of these peaks suggests a typical clay structure with additional properties due to the incorporation of copper. The addition of copper appears to affect the structure of the clay, possibly through the formation of copper-oxygen bonds that can change the physical and chemical properties of the clay.

When the FTIR spectrum was analyzed (Figure 3), spectral differences of the samples with copper (Cu) and copper-cerium (Cu-Ce) added to the porous clay heterostructure were observed.

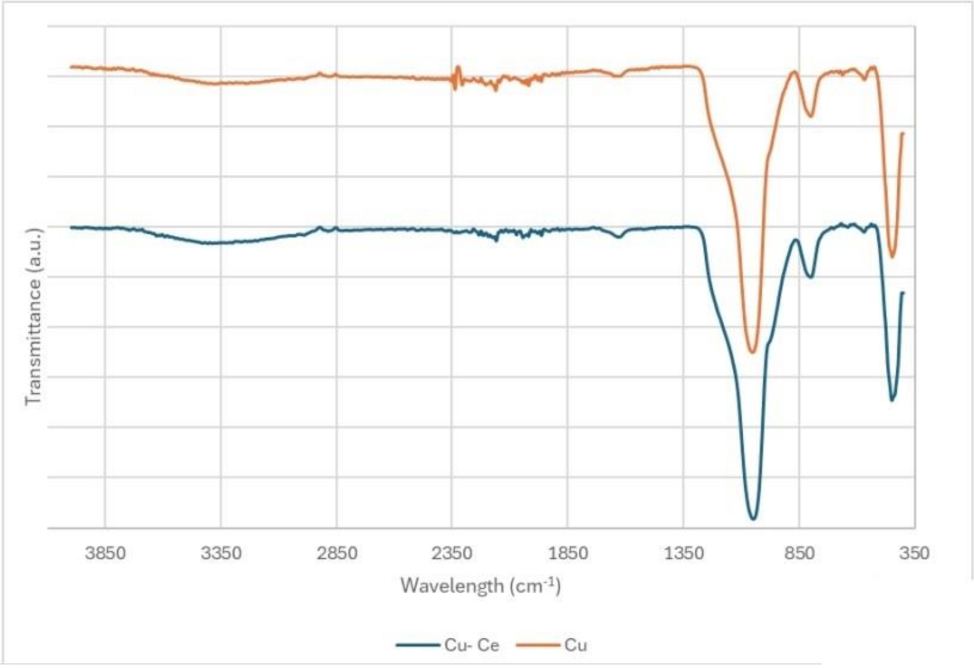


Figure 3. FTIR spectra taken with the ATR unit for Copper doped Cu-PCH and Copper and Cerium doped Cu-Ce-PCH.

At approximately 2350 cm^{-1}, Copper (Cu) Spectrum shows a relatively flat and stable pattern in this region, indicating minimum absorption in this region for the copper added sample. Copper and Cerium (Cu-Ce) Spectrum similar to the copper spectrum, the copper and cerium added sample also shows minimum absorption around 2350 cm^{-1}. This suggests that the addition of cerium does not significantly affect the molecular vibrations absorbed in this region, which typically contains the molecular vibrations of atmospheric CO_2. At approximately 850 cm^{-1}, there is a noticeable decrease in transmittance in the Copper (Cu) Spectrum, indicating a significant absorption peak. This suggests the presence of bond vibrations, possibly involving copper interactions with oxygen or other ligands in the clay matrix. Copper and Cerium (Cu-Ce) Spectrum absorption peak is more pronounced in the CuCe sample, indicating stronger or additional vibration modes at this wavelength. The presence of cerium may be affecting the electronic environment of copper, or it may be directly involved in the binding affecting this spectral region. At approximately 450 cm^{-1}, the Copper (Cu) Spectrum has a sharp and deep absorption peak typical of metal-oxygen bond vibrations in metal oxides or complexes. Copper and Cerium (CuCe) Spectrum shows that the absorption peak is deeper and slightly shifted, indicating a change in the metal-oxygen binding environment. The addition of cerium likely alters the copper-oxygen binding dynamics, possibly through a change in bond lengths or angles or through the formation of mixed metal oxides. The addition of cerium along with copper to the structure appears to increase and change the vibration properties of the clay matrix, especially in regions associated with metal-oxygen interactions. This may be due to a synergistic effect in which cerium alters the electronic environment of copper, leading to changes in bond lengths, bond angles, or types of bonds formed. These changes are crucial for applications where the electronic and structural properties of the material are critical, such as catalysis or electronic materials.

By analyzing the FTIR spectrum provided in the Figure 4, significant differences were observed between the two samples: one with copper (Cu) added to the structure and the other with both copper (Cu) and titanium (Ti) added.

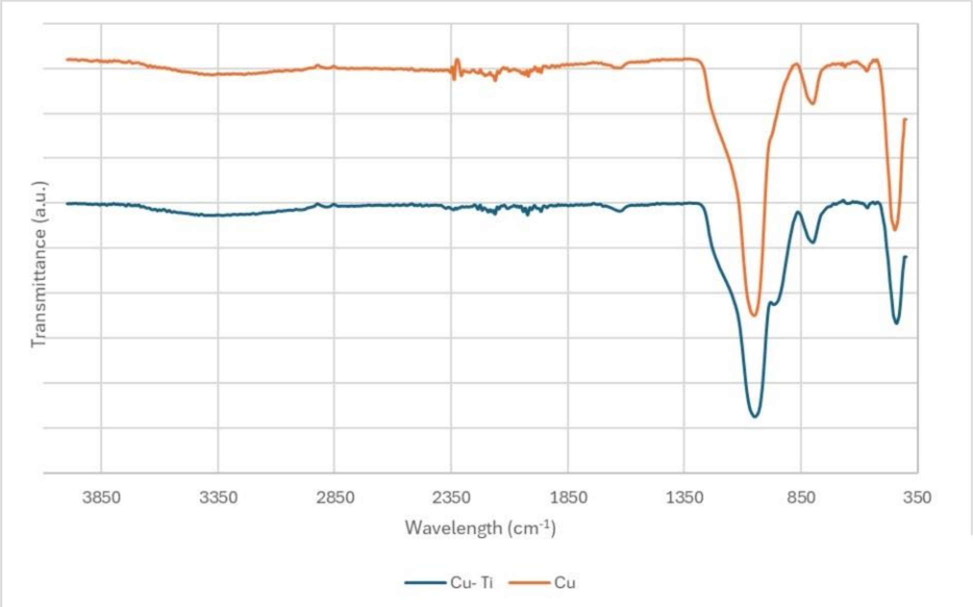


Figure 4. FTIR spectra taken with the ATR unit for Copper doped Cu-PCH and Copper and Titanium doped Cu-Ti-PCH.

Peak at 2350 cm^{-1} generally corresponds to atmospheric CO_2 absorption, which may not be related to material properties. The Copper (Cu) Spectrum has a visible peak at around 900 cm^{-1}, which may be indicative of Cu-O vibrations, suggesting the presence of copper oxide formations within the clay matrix. In the spectrum of

Copper and Titanium (Cu-Ti), peak at around 900 cm^{-1} is less pronounced compared to the Cu-only sample. This suggests that the addition of titanium may affect the formation of copper oxide or alter its environment, possibly forming different complex structures or stabilizing the matrix differently. In Copper (Cu) Spectrum, peak at approximately 450 cm^{-1} is sharp and deep trough, indicating strong absorption. This is typically characteristic of metal-oxygen bonds and probably indicates a strong interaction between copper ions and the clay matrix. In Copper and Titanium (Cu-Ti) Spectrum, peak at approximately 450 cm^{-1} is significantly less intense compared to that in Cu-Spectrum. This could mean that the presence of titanium as well as copper alters the metal–oxygen binding environment, potentially due to the formation of Ti–O bonds or a change in the structural arrangement of the clay matrix.

Copper (Cu) supports strong Cu-O binding, as evidenced by sharp peaks at lower wavelengths (about 450 cm^{-1}). The presence of copper oxide is also evident from the peak around 900 cm^{-1} . Titanium (Ti) addition appears to attenuate the intensity of copper-related peaks, suggesting a complex interaction that may stabilize or alter typical copper oxide formations. The reduced density at 450 cm^{-1} with Cu-Ti suggests a different bonding environment, possibly with Ti incorporation. These observations indicate that titanium addition alters the structural and chemical environment of the clay matrix, affecting interactions and formations typically associated with copper. This can have effects on the physical properties and chemical stability of the material.

By analyzing FTIR spectra for copper incorporated porous clay heterostructure (Cu-PCH) after pyridine adsorption and desorption at various temperatures, changes in transmittance peaks indicative of changes in the Browsted-Lewis acidity of the material were observed (Figure 5).

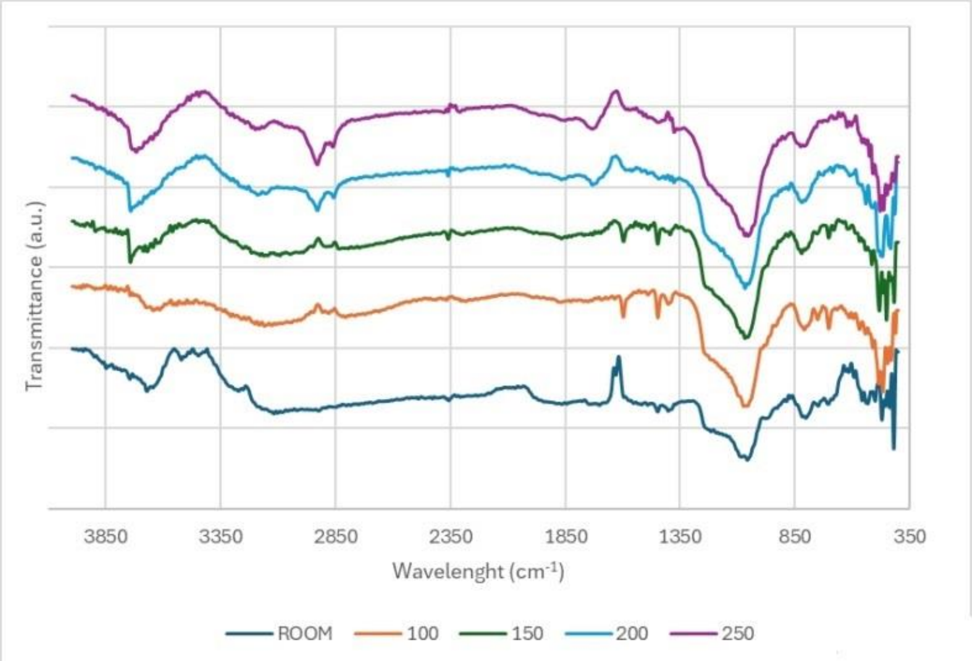


Figure 5. FTIR spectrum at different temperatures taken with the DRIFT unit for Cu-PCH.

At room temperature, the bands attributed to pyridine coordinately bound to Lewis sites (1446 cm^{-1}) Lewis and Brønsted sites (1491 cm^{-1}) and Brønsted sites (1546 cm^{-1}) were observed for Cu-PCH. The spectrum shows distinct peaks around 3350 cm^{-1} and 1630 cm^{-1}, which are associated with O-H stretching and bending vibrations, respectively. These peaks indicate the presence of hydroxyl groups, which are active sites for Browsted acidity. At 100°C, the intensity of the O-H stretching peak at 3350 cm^{-1} decreases slightly; this indicates a possible decrease in hydroxyl group availability or a change in hydrogen bonding dynamics due to temperature increase. At 150°C, the further decrease in the O-H stretching peak at 3350 cm^{-1} is more evident. Additionally, there is a slight shift in the peak towards lower wave numbers; this suggests a change in the chemical environment of the hydroxyl groups, possibly due to increased deprotonation or interaction with copper ions. At 200°C, the decreasing trend of peak intensity at 3350 cm^{-1} continues; this may be indicative of continued loss of hydroxyl groups or their conversion to other functional groups that are less active in Browsted acidity. The peak at 1630 cm^{-1} also shows a decrease and this indicates changes in water content or molecular interactions within the clay. At 250°C, the peaks at both 3350 cm^{-1} and 1630 cm^{-1} decrease significantly, indicating a significant reduction in hydroxyl groups. This decrease means lower Browsted acidity at higher temperatures as the availability of active acidic sites decreases. FTIR spectra show a consistent decrease in the intensity of the O–H stretching and bending peaks as the temperature increases from room temperature to 250 °C, this is associated with a decrease in Browsted-Lewis acidity. This trend can be attributed to the thermal deactivation of hydroxyl groups, which are crucial for acidity, and possibly their interaction or complexation with copper ions within the clay matrix. These changes are crucial for applications where temperature-dependent acidity plays a role in catalysis or other chemical processes.

The FTIR spectrum of the pyridine adsorbed clay samples modified with two different metals: one with copper (Cu) and the other with a combination of copper and cerium (Cu-Ce) were shown in the Figure 6. A detailed analysis of the observed changes in the spectrum and their potential impact on the Brønsted-Lewis acidity of the clay structure was performed.

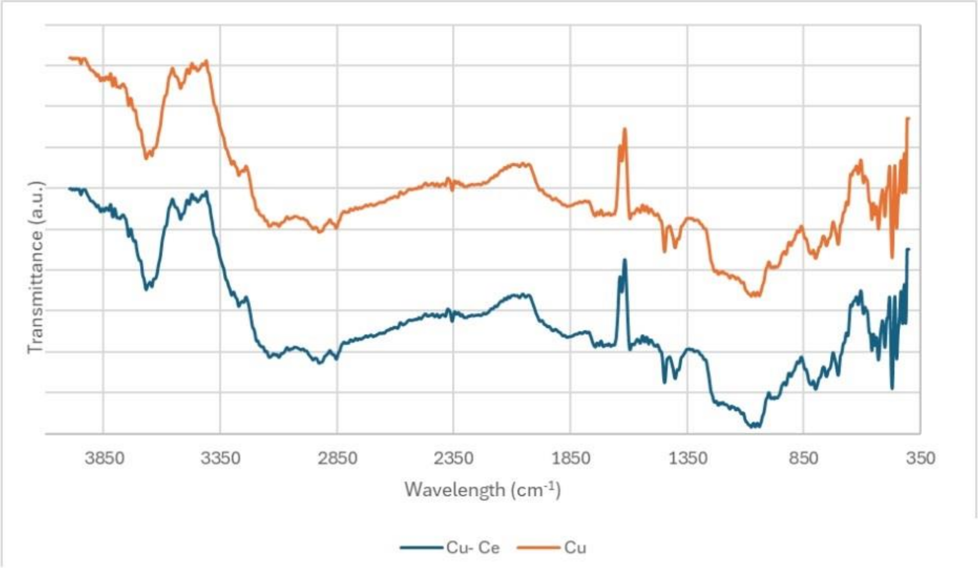


Figure 6. FTIR spectra taken with DRIFT unit for Copper doped Cu-PCH and Copper and Cerium doped Cu-Ce-PCH.

Band around 3850-3350 cm^{-1} corresponds to O-H stretching vibrations. The Cu-Ce sample shows a softer and less intense peak compared to the Cu sample; This suggests a possible interaction or modification of hydroxyl groups in the presence of cerium, which could affect the hydrogen bonding environment. Approximately 2850 cm^{-1} usually related to C-H stretching vibrations, shows significant differences between the two samples. The Cu sample exhibits sharper and more defined peaks indicating less interaction or modification of these groups compared to the CuCe sample. Around 2350 cm^{-1} peak may be associated with CO_2 or other triple bonded groups. The presence and intensity of peaks in this region may vary depending on the sample environment and preparation. Around 1850-1350 cm^{-1} large region can contain various functional groups, including C=O (carbonyl groups). The Cu-Ce sample shows more distinct peaks, indicating enhanced interactions or modifications involving these groups. Below 1350 cm^{-1} region generally contains bending vibrations of various ligaments. Significant differences in peak shapes and densities suggest different metal interactions affecting the network of the clay structure. The bands attributed to pyridine coordinately bound to Lewis sites (1446 cm^{-1}) Lewis and Brønsted sites (1491 cm^{-1}) and Brønsted sites (1546 cm^{-1}) were observed copper- and copper cerium- supported PCH samples. Modifications in the hydroxyl site (3850-3350 cm^{-1}) indicate changes in the availability and reactivity of proton donors, which directly affects Brønsted acidity. The softer peaks in the Cu-Ce sample may indicate a decrease in Brønsted acidity due to the addition of cerium. Changes in the regions associated with metal-oxygen bonds (around 1850-1350 cm^{-1}) indicate changes in the electron-accepting capacity of the clay. The increased peaks in the Cu-Ce sample indicate an increase in Lewis acidity due to the incorporation of cerium, which can potentially act as a strong Lewis acid. In summary, the addition of cerium in combination with copper appears to alter both the Brønsted and Lewis acidity of the clay by altering the interaction and environment of functional groups within the clay structure. These changes can significantly affect the catalytic, adsorptive and reactive properties of the modified clay.

The FTIR spectra (Figure 7) were observed for two different pyridine adsorbed clay samples modified with metals: one with copper-titanium (Cu-Ti) and the other with copper only (Cu).

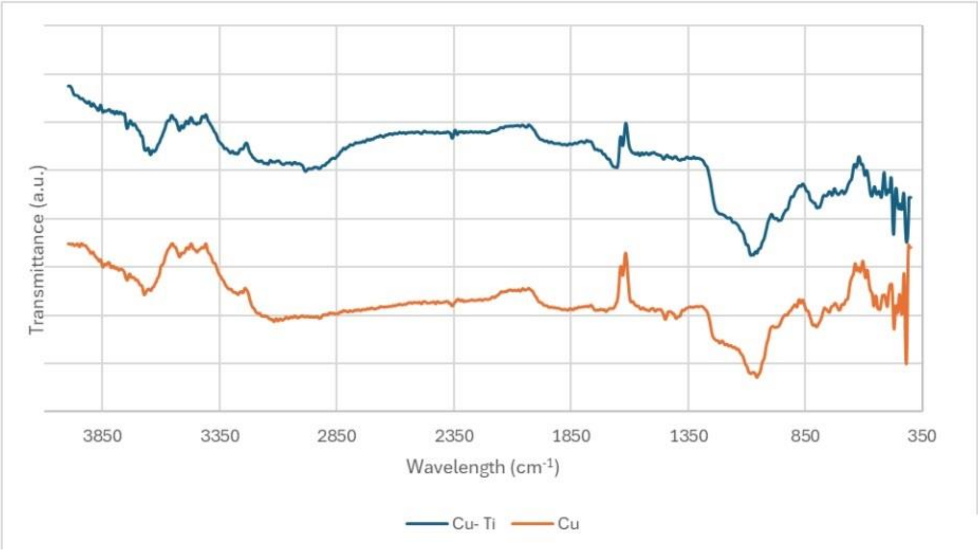


Figure 7. FTIR spectra taken with DRIFT unit for Copper doped Cu-PCH and Copper and Titanium doped Cu-Ti-PCH

Cu-Ti Spectrum generally shows smoother transitions with fewer sharp peaks compared to the Cu spectrum. Notable peaks appear around 3350 cm^{-1} and 1630 cm^{-1} , which are typically associated with O-H stretching and bending vibrations and suggest the presence of hydroxyl groups or water molecules. Cu Spectrum is rougher with sharper and more distinct peaks, especially noticeable around 2850 cm^{-1} and 1350 cm^{-1} . These may be indicative of C-H stretching and bending vibrations of O-H groups in organic compounds, respectively. The bands attributed to pyridine coordinately bound to Lewis sites (1446 cm^{-1}) Lewis and Brønsted sites (1491 cm^{-1}) and Brønsted sites (1546 cm^{-1}) were observed for Cu-Ti-PCH. Brønsted Acidity, typically associated with the presence of hydroxyl groups (O-H), sharper peaks around hydroxyl-related regions in the Cu spectrum indicate a possible increase in Brønsted acidity compared to the Cu-Ti spectrum. Lewis's acid sites are usually associated with metal centers that can accept electron pairs. The softer Cu-Ti spectrum may indicate a change in the electronic environment around the metal centers, potentially altering the Lewis acidity. The presence of titanium along with copper may affect the electron density and nature of the metal domains compared to clay containing only copper. Adding titanium to the Cu-Ti sample appears to stabilize the structure, as evidenced by smoother transitions in the spectrum. This may mean a more uniform distribution of metal ions within the clay matrix. The distinct differences in the spectra indicate that the addition of different metals (Cu and CuTi) significantly affects the chemical environment and acidity of the clay. This may affect the catalytic properties, adsorption properties and reactivity of modified clays. In conclusion, FTIR spectra show that adding titanium along with copper to the clay structure not only changes the physical appearance of the IR spectrum but also potentially increases stability and changes the acidity of the clay. This may have implications for applications where specific acidity and stability are crucial, such as catalysis or removal of contaminants.

This study investigated the synthesis and characterization of copper supported porous clay heterostructures (PCHs), with a particular focus on utilizing Fourier Transform Infrared Spectroscopy (FTIR) to examine their structural features and acidity. The findings demonstrate the potential of PCHs for various applications, particularly in improving catalytic performance and environmental cleanup. Key observations from the research include; Successful synthesis of PCHs: The in-situ synthesis method effectively produced PCHs using HTAC, TEOS, and a copper source. The addition of copper occurred simultaneously with silica welding. FTIR analysis reveals structural details: The obtained FTIR spectra provided valuable insights into the functional groups present in the PCHs and how they are affected by metal incorporation. Impact of temperature on Brønsted-Lewis acidity: The analysis of FTIR spectra at different temperatures using DRIFT indicated a decrease in Brønsted-Lewis acidity with increasing temperature. This suggests the deactivation of hydroxyl groups, which are crucial for acidity. Influence of cerium and titanium on PCH properties: The addition of cerium and titanium alongside copper showed significant changes in the FTIR spectra, suggesting their influence on the electronic environment and binding dynamics within the clay matrix. The FTIR analysis successfully identified various functional groups within the PCHs and how they interact with incorporated metals like copper, cerium, and titanium. The observed changes in peak intensities and positions provided valuable information about the impact of these metals on the structural and chemical properties of the PCHs. The decrease in Brønsted-Lewis acidity with increasing temperature suggests that PCHs might be more suitable for applications at lower temperatures. However, further investigations are needed to determine the optimal temperature range for specific applications. The addition of cerium and titanium alongside copper resulted in interesting changes in the FTIR spectra. These observations suggest that cerium and titanium can potentially modify the electronic environment and binding behavior within the PCHs, potentially leading to tailored properties for desired applications. Future studies should explore the impact of these modifications on the catalytic activity and adsorption capacity of PCHs. Overall, this study demonstrates the successful synthesis and characterization of PCHs. The FTIR analysis provided valuable insights into their structure and how metal incorporation affects their properties. The findings also suggest promising applications of PCHs in catalysis, adsorption, and environmental cleanup. Further research is recommended to explore these potential applications in greater detail. This study focused on the characterization of PCHs using FTIR. Further studies could involve techniques like X-ray diffraction (XRD) and electron microscopy to obtain a more comprehensive understanding of the structural properties. The catalytic activity and adsorption capacity of PCHs for various applications require further investigation. Optimizing the synthesis process to achieve specific pore sizes and metal loadings could be explored to tailor PCHs for desired functionalities.

3. CONCLUSION

This study successfully synthesized and characterized copper-supported porous clay heterostructures (PCHs) using an in-situ method with HTAC, TEOS, and a copper source. Fourier Transform Infrared Spectroscopy (FTIR) provided valuable insights into the structural features and acidity of the PCHs. Key findings include: The in-situ synthesis method effectively produced PCHs with simultaneous copper incorporation during silica welding. The FTIR spectra revealed the presence of various functional groups and demonstrated how metal incorporation, specifically copper, cerium, and titanium, affects the PCHs' structural and chemical properties. FTIR analysis at different temperatures showed a decrease in Brønsted-Lewis acidity with increasing temperature, indicating the deactivation of crucial hydroxyl groups. The addition of cerium and titanium alongside copper resulted in significant changes in the FTIR spectra, suggesting modifications in the electronic environment and binding dynamics within the clay matrix. The study indicates that PCHs hold promise for applications in catalysis, adsorption, and environmental cleanup. The observed changes in FTIR spectra suggest potential for tailoring PCH properties by incorporating different metals. Future research should explore these modifications' impact on catalytic activity and adsorption capacity and consider techniques like X-ray diffraction (XRD) and electron microscopy for a more comprehensive understanding of structural properties.

Acknowledgement

The authors gratefully acknowledge the financial support provided by the Scientific Research Project Department of Gazi University through research grant BAP 06/2009-49.

References

Besghaier, S., Cecilia, J. A., Chouikhi, N., Vilarrasa-García, E., Rodríguez-Castellón, E., Chlendi, M., & Bagane, M. (2021). Glyphosate adsorption onto porous clay heterostructure (PCH): kinetic and thermodynamic studies. *Brazilian Journal of Chemical Engineering*, 1-15.

Cecilia, J. A., García-Sancho, C., Vilarrasa-García, E., Jiménez-Jiménez, J., & Rodriguez- Castellón, E. (2018). Synthesis, characterization, uses and applications of porous clays heterostructures: a review. *The Chemical Record*, 18(7-8), 1085-1104.

Chmielarz, L., Kuśtrowski, P., Drozdek, M., Dziembaj, R., Cool, P., & Vansant, E. F. (2006). Selective catalytic oxidation of ammonia into nitrogen over PCH modified with copper and iron species. *Catalysis Today*, 114(2-3), 319-325.

Chmielarz, L., Kuśtrowski, P., Dziembaj, R., Cool, P., & Vansant, E. F. (2007). Selective catalytic reduction of NO with ammonia over porous clay heterostructures modified with copper and iron species. *Catalysis Today*, 119(1-4), 181-186.

Chmielarz, L., Piwowarska, Z., Kuśtrowski, P., Węgrzyn, A., Gil, B., Kowalczyk, A., ... & Michalik, M. (2011). Comparison study of titania pillared interlayered clays and porous clay 27 heterostructures modified with copper and iron as catalysts of the DeNOx process. *Applied Clay Science*, 53(2), 164-173.

Chmielarz, L., Kowalczyk, A., Skoczek, M., Rutkowska, M., Gil, B., Natkański, P., ... & Ryczkowski, J. (2018). Porous clay heterostructures intercalated with multicomponent pillars as catalysts for dehydration of alcohols. *Applied Clay Science*, 160, 116-125.

Dincer, B. Y., Balcı, S., & Tomul, F. (2020). In-situ mesoporous silica pillared clay synthesis and effect of titanium and iron incorporation to structural properties. *Microporous and Mesoporous Materials*, 305, 110342.

Essih, S., Cecilia, J. A., Jiménez-Gómez, C. P., García-Sancho, C., García-Mateos, F. J., Rosas, J. M., & Maireles-Torres, P. (2022). Synthesis of Porous Clay Heterostructures Modified with SiO2–ZrO2 Nanoparticles for the Valorization of Furfural in One-Pot Process. *Advanced Sustainable Systems*, 6(5), 2100453.

Guimarães, D. T., Mendes, L. M. R., de Sousa Sabino, L. B., de Brito, E. S., VilarrasaGarcía, E., Rodríguez-Castellón, E., ... & da Silva Junior, I. J. (2023). Partial Purification of Anthocyanins (Brassica oleracea var. Rubra) from Purple Cabbage Using Natural and Modified Clays as Adsorbent. *Adsorption Science & Technology*, 2023, 2724122.

Ishizaki, K., Komarneni, S., & Nanko, M. (1998). Porous Materials: Process technology and applications, 1-10.

Jaramillo-Fierro, X., González, S., Jaramillo, H. A., & Medina, F. (2020). Synthesis of the ZnTiO3/TiO2 nanocomposite supported in ecuadorian clays for the adsorption and photocatalytic removal of methylene blue dye. *Nanomaterials*, 10(9), 1891.

Kashif, M., Khan, M. N., Su, Y., & Heynderickx, P. M. (2022). Most efficient mesoporous Mn/Ga-PCH catalyst for low-temperature selective catalytic reduction of NO with C3H6. *Vacuum*, 198, 110879.

Mao, J., Zhou, Y., Lv, G., & Zhou, R. (2022). Simultaneous detoxification of aflatoxin B1, zearalenone and deoxynivalenol by modified montmorillonites. *Molecules*, 27(1), 315.

Muñoz, M., Greber, M., Tayeb, K. B., Lamonier, C., Cabello, C. I., & Romanelli, G. P. (2023). Catalysts based on nickel salt heteropolytungstates for selective oxidation of diphenyl sulfide. *Green Processing and Synthesis*, 12(1), 20230026.

Nagendrappa, G., & Chowreddy, R. R. (2021). Organic reactions using clay and claysupported catalysts: a survey of recent literature. *Catalysis Surveys from Asia*, 25(3), 231-278.

Nicola, B. P., Schwanke, A. J., & Bernardo-Gusmão, K. (2022). Porous clay nanoarchitectures as supports for heterogenized nickel complexes in catalytic reactions of ethylene oligomerization. *Catalysis Today*, 394, 256-267.

Pires, J., & Carvalho, S. (2022). Adsorbents from natural clays for the separation of hexane isomers. *Journal of Environmental Chemical Engineering*, 10(3), 107689.

Polverejan, M., Liu, Y., & Pinnavaia, T. J. (2002). Aluminated derivatives of porous clay heterostructures (PCH) assembled from synthetic saponite clay: properties as supermicroporous to small mesoporous acid catalysts. *Chemistry of materials*, 14(5), 2283-2288.

Sato, K., Fujikawa, E., & Cecilia, J. A. (2022). Increased protonation of a mesopore surface in a porous clay nanoheterostructure. *The Journal of Physical Chemistry C*, 126(30), 12615- 12622.

Schoonheydt, R., Pinnavaia, T., Lagaly, G., & Gangas, N. (1999). Pillared Clays and Pillared Layered Solids. *Pure and Applied Chemistry,* 71, 2367-2371.

Yuan, M., Su, Y., Deng, W., & Zhou, H. (2019). Porous clay heterostructures (PCHs) modified with copper ferrite spinel as catalyst for SCR of NO with C3H6. *Chemical Engineering Journal*, 375, 122091.

Chapter 15

Evaporation Prediction Using Machine Learning Models Enhanced with cVAE–cWGAN-GP Based Synthetic Data Generation

Özlem TERZİ[1],Emine Dilek TAYLAN[2]

ABSTRACT

This study proposes an evaporation estimation framework that integrates machine learning (ML) models with synthetic data generation to address data scarcity in hydrometeorological time series. In the first stage, evaporation was modeled using observed monthly precipitation, temperature, and lagged variables. The dataset was chronologically divided into two groups: 80% training (1975-1999) and 20% testing (2000-2012). The ML models (Random Forest (RF), XGBoost, LightGBM, CatBoost, Support Vector Regression (SVR), and Multilayer Perceptron (MLP)) and a stacking ensemble (SE) were developed. Among the ML_BASE models developed using observed variables, XGBoost achieved the best performance (RMSE = 33.24 mm, $R^2 = 0.891$).

In the second stage, data augmentation was performed to expand the training set and improve generalization. Synthetic meteorological sequences were generated using a log1p-v3 architecture based on Conditional Variational Autocoder (cVAE) and Conditional Wasserstein GAN with Gradient Penalty (cWGAN-GP). Statistical similarity between observed and synthetic samples was assessed using Kolmogorov-Smirnov (KS) and Wasserstein Distance (WD). All KS p-values exceeded 0.10, indicating no significant distributional difference. WD values ranged from 0.29 to 2.30, and the Frobenius correlation difference (0.393) confirmed that multivariate dependence structures were preserved.

In the final stage, the ML_AUG models were retrained using the combined observed and synthetic data. XGBoost again showed the highest accuracy (RMSE = 33.12 mm, $R^2 = 0.892$). Seasonal analysis revealed that RF and XGBoost remained consistently stable in both summer and winter. Feature

---

[1] Prof. Dr.; Department of Civil Engineering, Faculty of Technology, Isparta University of Applied Sciences, Isparta, Turkey. ozlemterzi@isparta.edu.tr ORCID No:0000-0001-6429-5176 (Corr. author)

[2] Prof.Dr.; Department of Civil Engineering, Faculty of Engineering and Natural Sciences, Suleyman Demirel University, Isparta, Turkey. dilektaylan@sdu.edu.tr ORCID No:0000-0003-0734-1900

importance rankings remained unchanged, demonstrating that augmentation did not disrupt physical relationships.

Overall, the findings demonstrate that cVAE + cWGAN-GP-based augmentation effectively enriches hydrometeorological datasets, preserves statistical properties, and increases model robustness in data-limited environments.

Keywords – Evaporation prediction; Machine learning; Data augmentation; Eğirdir Gölü; Turkey.

INTRODUCTION

Evaporation is a fundamental component of the hydrological cycle and plays a critical role in water budget modeling, flood–drought assessment, agricultural irrigation planning, and basin-scale water resources management (Lange & Sippel, 2020). Accurate estimation of evaporation is strategically important, particularly in water-stressed regions, where it affects reservoir operation, drinking water supply, and the evaluation of climate change impacts. However, evaporation exhibits a nonlinear structure shaped by complex interactions of meteorological, hydrological, and atmospheric processes (Deo et al., 2016). For this reason, classical statistical methods often fail to provide sufficient accuracy, especially when observation records are short or incomplete (Khandelwal et al., 2020).

Traditional estimation approaches rely on physically based models such as Penman, Priestley–Taylor, Makkink, and Thornthwaite. Yet, the direct measurement of evaporation is difficult, and the absence of long-term station records makes calibration of these models challenging in practice. Additionally, the lack of complementary meteorological variables—such as wind speed, solar radiation, humidity, and sunshine duration—limits the reliability of physical methods (Lange & Sippel, 2020). Consequently, alternative techniques for evaporation prediction have gained attention in recent years, and machine learning (ML)–based models are increasingly used in the literature (Deo et al., 2016; Al-Mukhtar et al., 2022; Zerouali et al., 2025; Shabani et al., 2020).

Machine learning techniques offer the ability to capture nonlinear relationships, represent complex interactions among multiple variables, and learn irregularities in data distributions. Methods such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), LightGBM, CatBoost, Support Vector Regression (SVR), and artificial neural networks (ANNs) have demonstrated strong performance in evaporation prediction across different hydroclimatic conditions (Deo et al., 2016). However, the success of ML models largely depends on data quantity and variability. In many regions, meteorological stations have short observation periods, missing measurements, sensor errors, or discontinued operation—factors that reduce generalization ability and increase uncertainty, especially in extreme event prediction (Khandelwal et al., 2020).

This challenge has made data augmentation and synthetic data generation important and increasingly popular research directions in hydrological modeling. Generative Adversarial Networks (GAN) and Variational Autoencoders (VAE) have been successfully applied in various scientific domains due to their ability to learn complex data distributions and generate statistically consistent synthetic samples. However, producing synthetic data that accurately represents hydrometeorological variables requires preserving not only individual data distributions, but also multivariate dependency structures, skewness, density patterns, and extreme value behavior. Therefore, maintaining the correlation structure is just as essential as reproducing marginal distributions (Bakhshipour et al., 2023).

To address these limitations, this study proposes an evaporation prediction framework that integrates ML models with a cVAE–cWGAN-GP–based synthetic data generation process. The main objectives of the study are:

- To develop and evaluate ML-based evaporation prediction models using observed meteorological data,
- To address data scarcity through cVAE–cWGAN-GP-based synthetic data generation,
- To assess the statistical similarity between synthetic and real data using KS tests, Wasserstein Distance, and correlation matrix comparison,
- To retrain ML_AUG models using the augmented dataset and compare their predictive performance,
- To analyze model behavior through time-series evaluation, seasonal performance metrics, and feature-importance analysis.

In this context, the study evaluates not only the predictive performance of machine learning models, but also the hydrological validity and statistical reliability of the synthetic data itself. The findings indicate that synthetic samples do not degrade model performance; instead, they enhance stability while preserving key distributional characteristics of the original dataset. Such an approach is particularly valuable for basins where long-term hydrometeorological records are limited or incomplete.

Therefore, the primary objective of this study is to develop a machine learning–based evaporation prediction framework supported by a cVAE–cWGAN-GP synthetic data generation model, and to assess how data augmentation influences prediction accuracy, seasonal performance, variable sensitivity, and temporal dynamics for Eğirdir Lake. By integrating statistical similarity analyses with machine learning evaluation, the study aims to demonstrate that synthetic data can serve as a reliable substitute for observational gaps and a practical tool in regional water resources management.

STUDY AREA

This study focuses on Eğirdir Lake, a prominent freshwater body situated in the Mediterranean Region of Türkiye, within the Isparta province, at geographic coordinates of approximately 37.87° N latitude and 30.85° E longitude (Figure 1). With a surface area of approximately 482 km², Eğirdir ranks as the fourth largest natural lake in Türkiye, following Lakes Van, Tuz, and Beyşehir (Şener et al., 2013). The lake lies in a tectonic depression in the Isparta Angle, a transitional zone between Central Anatolia and the Mediterranean basin (Karaman, 2010) and has an average depth ranging from 8 to 14 meters (Atilgan et al., 2022). Its elevation is approximately 915–920 meters above sea level, and it exhibits a narrow, elongated morphology oriented primarily along the north–south axis (Kazancı & Roberts, 2019)

Eğirdir Lake is located within a Mediterranean climatic zone, characterized by hot, dry summers and mild, wet winters, which significantly influences its hydrological balance. The average annual precipitation ranges between 500 and 700 mm, largely concentrated in the winter and spring months, while evaporation rates peak during the summer, driven by high temperatures and low relative humidity (Atilgan et al., 2022). Seasonal variations in air temperature, wind speed, and solar radiation contribute to fluctuations in lake water level, making evapotranspiration a key component in the lake's water budget dynamics. These climatic patterns are especially critical given the region's reliance on the lake for agricultural irrigation, notably for apple and cherry orchards surrounding the basin (Şener et al., 2013).

In this study, monthly evaporation, precipitation, and temperature data were used to predict evaporation for Eğirdir Lake. All meteorological variables were obtained from the Turkish State Meteorological Service (MGM) and span the period 1975–2012, providing a sufficiently long record to characterize the hydroclimatic regime of the lake. The descriptive statistics of the meteorological parameters are given in Table 1.

Table 1: The descriptive statistics of monthly meteorological parameters

Parameters	**Station**	**Mean**	**Max**	**Min**
Evaporation (mm)	Eğirdir	128.46	443.48	0.00
Precipitation (mm)	Eğirdir	64.52	457.73	0.00
Temperature (^{0}C)	Eğirdir	12.38	29.07	-5.74

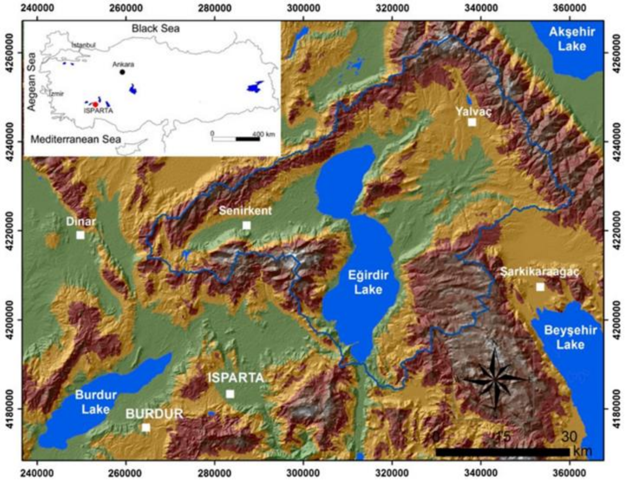


Figure 1: Study area (Şener et al., 2013)

MACHINE LEARNING

Machine learning (ML) has become increasingly prevalent in hydrological modeling due to its ability to capture nonlinear relationships, identify complex patterns, and generate robust predictions from high-dimensional data. In evaporation dynamics, where atmospheric drivers interact in nonlinear and time-dependent ways, ML provides significant advantages over traditional physically-based modeling (Liang et al., 2023).

In this study, six primary ML algorithms and one ensemble model were employed for evaporation forecasting: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP). In addition, a Stacking Ensemble (SE) model was constructed to combine the strengths of the base learners. The descriptions of these models and their advantages for hydroclimatological prediction are provided below.

Random Forest (RF)

RF is a tree-based ensemble method that uses bootstrapped sampling and random variable selection to construct multiple independent decision trees. The final prediction is obtained by averaging outputs from all trees. This mechanism reduces overfitting and improves generalization, making RF highly effective with noisy meteorological data. Moreover, RF offers a straightforward estimation of feature importance, providing insight into the relative influence of climatic variables on evaporation (Kumar et al., 2023).

Extreme Gradient Boosting (XGBoost)

XGBoost is an optimized form of gradient boosting, designed for computational efficiency, regularization, and reliable handling of missing values. The model sequentially builds trees that correct errors from previous learners, improving accuracy at each iteration. XGBoost is widely used in hydrological and environmental forecasting due to its high predictive power, fast training capability, and strong performance with structured tabular datasets (Kumar et al., 2023).

Light Gradient Boosting Machine (LightGBM)

LightGBM is another gradient boosting algorithm, optimized through a leaf-wise tree growth strategy that reduces training time and minimizes memory requirements. This structure allows deeper trees with fewer nodes, improving prediction accuracy and computational efficiency. Its suitability for continuous meteorological variables and high-dimensional inputs makes it valuable for representing evaporation dynamics (Liang et al., 2023).

Categorical Boosting (CatBoost)

CatBoost is a gradient boosting method designed to handle categorical variables without extensive preprocessing. Although the meteorological dataset used in this study mostly consists of continuous inputs, CatBoost’s regularization techniques and ordered boosting structure reduce overfitting and enhance stability in time-dependent hydrological modeling (Kumar et al., 2023).

Support Vector Regression (SVR)

SVR models nonlinear relationships by mapping input variables into a higher-dimensional feature space using kernel functions, most commonly the Radial Basis Function (RBF). The algorithm seeks an optimal hyperplane that fits the data within an acceptable tolerance margin. While effective for hydrological forecasting, the computational cost increases significantly with large datasets (Risha & Liu, 2025).

Multi-Layer Perceptron (MLP)

MLP represents a feed-forward artificial neural network composed of input, hidden, and output layers. Activation functions allow the model to learn complex nonlinear patterns within the data. Although widely applied in climate and water resources modeling, MLP requires careful tuning of hyperparameters and may experience overfitting if the dataset lacks sufficient diversity (Kumar et al., 2023).

Stacking Ensemble (SE)

The SE combines outputs from multiple models using a secondary learner. In this study, base models consisted of RF, XGBoost, LightGBM, CatBoost, SVR, and MLP. Their predictions were blended using a Ridge Regression meta-learner to generate the final output. Stacking aims to exploit complementary strengths of individual algorithms, reduce individual model biases, and enhance generalization performance (Liang et al., 2023).

SYNTHETIC DATA GENERATION: cVAE–cWGAN-GP

To address the limitations arising from short observational records and missing hydrometeorological measurements, a hybrid synthetic data generation strategy was implemented, combining a Conditional Variational Autoencoder (cVAE) and a Conditional Wasserstein GAN with Gradient Penalty (cWGAN-GP). This integrated approach is increasingly applied for hydroclimatic modeling, where maintaining physical consistency and distributional structure is critical (Kingma & Welling, 2014; Gulrajani et al., 2017; Jahangir & Quilty, 2023).

cVAE Component

The cVAE learns the underlying probability density of observed predictors and generates new realizations by resampling a latent probabilistic space. Observed variables X are encoded into latent vectors $z \sim (\mu, \sigma^2)$ and then decoded under the same conditioning. This mechanism reproduces realistic variability, ensuring that synthetic samples exhibit smooth, continuous probability distributions rather than deterministic replications (Rezende et al., 2014). In hydrological contexts, cVAE models have demonstrated robust performance in preserving stochastic variability and forecast uncertainty (Jahangir & Quilty, 2023).

cWGAN-GP Component

To refine the cVAE-generated samples, a cWGAN-GP model was applied. The generator produces synthetic data conditioned on meteorological predictors, while the critic evaluates discrepancies between real and synthetic distributions using the Wasserstein distance, which provides a continuous learning signal and mitigates training instability (Arjovsky et al., 2017; Gulrajani et al., 2017). The gradient penalty enforces Lipschitz continuity and prevents mode collapse, a common problem in classical GANs. The conditional structure guarantees that physically meaningful relations—such as temperature–evaporation and precipitation–evaporation interactions—are preserved (Yonekura et al., 2024).

Advantages of the Hybrid Design

In this study, the two architectures were combined sequentially: *(1)* cVAE captured the global statistical distribution of observed predictors and produced an initial synthetic dataset, *(2)* cWGAN-GP improved realism by preserving multivariate correlation structures and tail behavior.

This hybrid mechanism: preserved the probability distributions of individual variables; maintained multivariate correlation patterns; enhanced representation of rare or extreme events (Jahangir & Quilty, 2023; Yonekura et al., 2024).

STATISTICAL VALIDATION METRICS

To verify the realism of the synthetic dataset, a multi-metric evaluation was conducted between observed and generated samples:

Kolmogorov–Smirnov (KS) test:

Assesses whether real and synthetic distributions differ significantly. This test is particularly effective when validating distributional assumptions in small-sample hydrological datasets (Jahangir & Quilty, 2023).

Wasserstein Distance (WD):

Quantifies the transport cost between probability distributions. WD is useful for capturing differences in skewed or heavy-tailed climate variables, where traditional metrics may underestimate distributional divergence (Gulrajani et al., 2017).

Frobenius Norm of Correlation Matrices:

Measures structural similarity of multivariate dependencies. This metric is valuable when assessing whether the synthetic data preserves joint variable interactions critical to hydrological realism (Yonekura et al., 2024).

Together, these metrics confirm that the hybrid cVAE–cWGAN-GP system can reproduce both marginal distributions and multivariate structure of the original dataset, making the synthetic observations suitable for evaporation prediction in data-scarce environments.

RESULTS AND DISCUSSION

In the first stage of this study, evaporation was estimated using machine learning models (ML_BASE) trained solely on observed data. Subsequently, synthetic samples representing the model inputs were generated using a conditional Variational Autoencoder (cVAE) and a conditional Wasserstein Art-GAN with Gradient Penalty (cWGAN-GP) based (log1p)–v3 algorithm, and the dataset was expanded. In this way, the statistical distribution of the data was balanced, and additional samples were produced to enhance the generalization capability of the model. In the final stage, ML_AUG models were retrained using both observed and synthetic data, and their performances were compared.

The analyses were based on monthly precipitation (mm), temperature (°C), and evaporation (mm) observations. To represent the temporal dependency of evaporation, the lagged variables Temp_lag1, Precip_lag1, and Evap_lag1 were generated; and to capture short-term trends and seasonal effects, Temp_mean3 and Precip_sum3 were derived. Thus, the final input set consisted of seven features: Precipitation, Temperature, Temp_lag1, Precip_lag1, Evap_lag1, Temp_mean3, and Precip_sum3.

The dataset was split chronologically into 80% training (1975–1999) and 20% testing (2000–2012) subsets. During model development, input variables were normalized using the StandardScaler, while the target variable (evaporation) was scaled with the MinMaxScaler (Abadi et al., 2016). This preprocessing procedure aimed to balance variables measured on different units and improve the stability of the learning process.

To evaluate the performance of the developed models, Root Mean Square Error (RMSE) and Coefficient of Determination (R^2), were employed as performance metrics. The formulations of RMSE, and R^2 are provided in Equations (1-2). Here, N denotes the total number of data points, $E_{i(obs)}$ is the observed streamflow, and E is the simulated streamflow. $\bar{E}_{i(obs)}$ *and* $\bar{E}_{i(sim)}$ represent the mean values of the observed and simulated streamflows, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(E_{i(obs)} - E_{i(\mathrm{sim})}\right)^2} \qquad (1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(E_{i(obs)} - E_{i(sim)}\right)^2}{\sum_{i=1}^{N}\left(E_{i(obs)} - \bar{E}_{i(obs)}\right)^2} \tag{2}$$

During the training phase, six ML models and one ensemble model were used: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost), Support Vector Regression (SVR), and Multi-Layer Perceptron (MLP). These models are widely preferred in evaporation prediction studies due to their ability to capture nonlinear relationships and represent complex interactions (Shabani et al., 2020; Al-Mukhtar et al. 2022). In addition, a Stacking Ensemble (SE) model with a Ridge meta-learner was designed to combine the strengths of the base models, aiming to enhance generalization performance by exploiting the complementary learning patterns of different algorithms.

ML_BASE Models

In the first stage of the study, ML models developed using only observed data (ML_BASE) were evaluated, and the performance metrics of the test set are presented in Table 2.

Table 2: Performance metrics of ML_BASE models

Model	**RMSE (mm)**	**R²**
RF	35.95	0.873
XGBoost	33.24	0.891
LightGBM	35.09	0.879
CatBoost	39.96	0.843
SVR	49.47	0.760
MLP	40.65	0.838
SE	34.39	0.884

According to Table 2, XGBoost achieved the best performance with an RMSE of 33.24 mm and an R^2 of 0.891, while SE (RMSE = 34.39 mm, R^2 = 0.884) and LightGBM (RMSE = 35.09 mm, R^2 = 0.879) produced similar accuracy levels. Although RF yielded a slightly higher error (35.95 mm), it still demonstrated an acceptable performance. In contrast, CatBoost showed a noticeable decrease in accuracy (RMSE = 39.96 mm, R^2 = 0.843), suggesting that the algorithm may lose part of its advantage when applied solely to numerical meteorological data.

Non–tree-based models such as MLP and SVR reached lower accuracy levels, indicating that tree-based methods tend to be more stable for nonlinear hydrometeorological processes. Overall, the findings confirm that ensemble-based approaches—particularly XGBoost, SE, and LightGBM—provide high predictive performance in evaporation estimation. The performance of the models in summer and winter seasons was evaluated separately, and the corresponding results are given in Table 3.

Table 3. Seasonal Performance (Summer/Winter)

Model	**Summer**		**Winter**	
	RMSE (mm)	**R^2**	**RMSE (mm)**	**R^2**
RF	30.37	0.917	30.79	0.914
XGBoost	29.20	0.923	30.54	0.915
LightGBM	26.54	0.936	34.73	0.891
CatBoost	32.72	0.903	43.67	0.827
SVR	49.16	0.782	56.23	0.714
MLP	36.73	0.878	34.78	0.890
SE	28.17	0.928	32.20	0.906

In the summer period, the best performance was obtained by LightGBM with an RMSE of 26.54 mm and an R^2 of 0.936, while SE and XGBoost achieved similar accuracy levels. In winter, error values increased overall; however, XGBoost and RF produced more stable predictions with RMSE values around 30–31 mm and R^2 values between 0.914–0.915. CatBoost showed a notable loss of accuracy during winter, and SVR had the lowest predictive skill in both seasons. These findings indicate that LightGBM and SE perform more strongly during summer, whereas RF and XGBoost are more robust under winter conditions. Following the seasonal performance analysis presented in Table 3, Figure 2 illustrates the results of the feature importance analysis conducted to determine which meteorological variables the XGBoost model is most sensitive to.

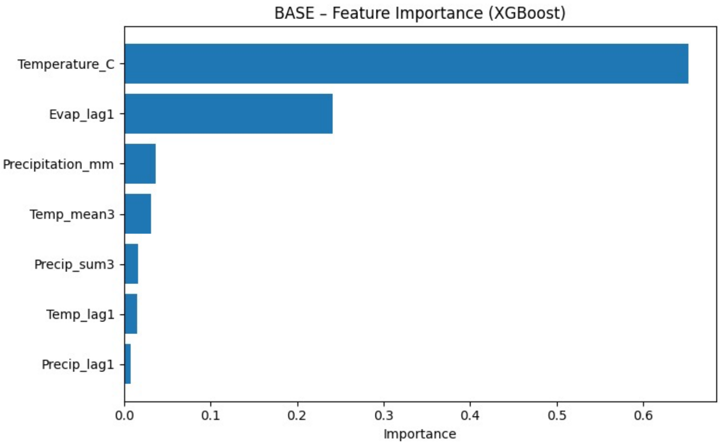


Figure 2: Feature importance for XGBoost_BASE

Figure 2 illustrates the feature importance distribution for the XGBoost_BASE model. The Temperature variable is the most dominant input with an importance value of 0.652, followed by Evap_lag1 with 0.240. This result confirms that temperature is the primary driver of evaporation and that the previous period’s evaporation reflects temporal memory in the system. The contributions of the remaining meteorological parameters are relatively limited: Precipitation (0.036), Temp_mean3 (0.031), Precip_sum3 (0.016), Temp_lag1 (0.015), and Precip_lag1 (0.007). In Figure 3, the evaporation estimates produced by the XGBoost model for the test set are presented together with the observed measurements over time.

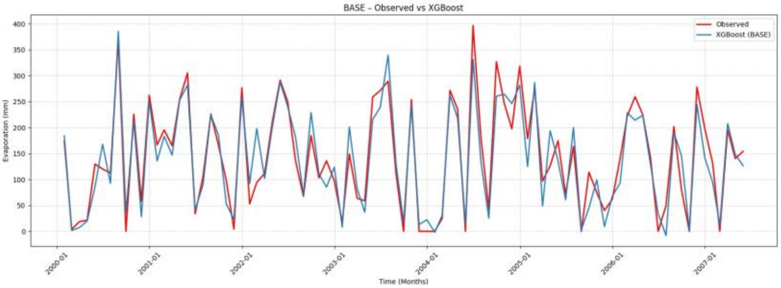


Figure 3. Time-series of the XGBoost_BASE model

As shown in Figure 3, the XGBoost predictions are presented together with the observed evaporation values. The prediction curve successfully captures the peak points during high-evaporation periods and aligns well with the observations across most of the temporal fluctuations. The limited deviations that appear in certain intervals may be associated with measurement uncertainties or sudden meteorological variations.

While the time-series representation in Figure 3 illustrates how effectively the XGBoost model reproduces the temporal dynamics and peak evaporation events, a deeper statistical evaluation is required to holistically assess model skill. For this purpose, Taylor diagrams were employed to jointly examine correlation, standard deviation, and centered RMS error, thereby providing a multidimensional perspective on model fidelity. The Taylor diagrams generated for both the test and training phases (Figures 4) allow for a rigorous comparison of the BASE models, highlighting their ability to replicate the statistical structure of the reference evaporation series and revealing potential differences in generalization and overfitting behavior. This integrated visualization thus complements the time-series analysis and offers a more robust interpretation of model performance.

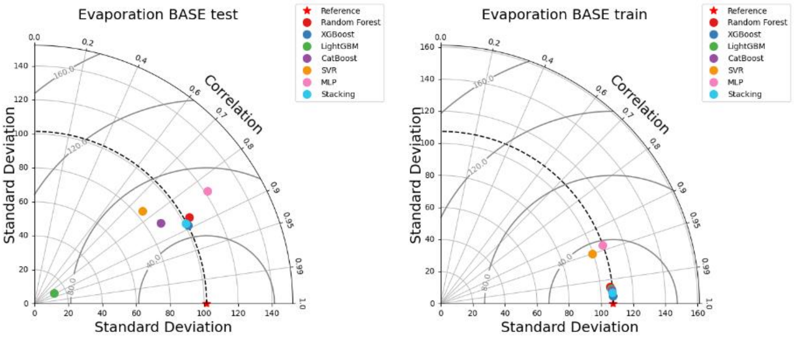


Figure 4. Taylor diagram illustrating the comparative performance of the BASE models

ML_AUG Models

Since meteorological observations often have limited temporal coverage, which may restrict the generalization ability of machine learning models, synthetic data generation was adopted in this phase. For this purpose, the observed input variables were resampled and expanded using a conditional Variational Autoencoder (cVAE) and a conditional Wasserstein Art-GAN with Gradient Penalty (cWGAN-GP) based (log1p)–v3 algorithm. This approach increased data diversity and broadened the learning space of the models, aiming to achieve a more stable evaporation prediction performance.

To evaluate the reliability of the synthetic data generation process, a statistical similarity analysis was performed between the observed (BASE) dataset and the synthetic (AUG) dataset (Table 4). For each variable, the

Kolmogorov–Smirnov (KS) test and Wasserstein Distance (WD) metrics were calculated. While the KS test determines whether the difference between the distributions is statistically significant, the WD metric quantifies the distance between probability densities. Accordingly, a comprehensive assessment was carried out in terms of both distributional shape and probability structure.

Table 4: Statistical similarity between observed and synthetic datasets

Variable	KS p-value	WD	Interpretation
Precipitation (mm)	0.996	1.205	Similar
Temperature (°C)	0.218	0.545	Similar
Temp_lag1	0.567	0.444	Similar
Precip_lag1	1.000	0.958	Similar
Evap_lag1	0.999	2.207	Similar
Temp_mean3	0.632	0.297	Similar
Precip_sum3	1.000	2.284	Similar
Real dataset size: (443, 7)			
Synthetic dataset size: (800, 7)			
Correlation matrix Frobenius difference: 0.393			

The results indicate that the synthetic dataset exhibits a high level of statistical similarity to the real data. The p-values of the KS test being well above 0.10 for all variables demonstrate that there is no statistically significant difference between the distributions and that the synthetic data successfully preserves the distributional characteristics of the real data. The WD values range from 0.29 to 2.30, which falls within acceptable deviation limits for hydrometeorological variables. The highest WD values are observed for Evap_lag1 (2.207) and Precip_sum3 (2.284), suggesting that time-lagged processes tend to generate slightly higher variance in the synthetic samples.

The Frobenius difference of 0.393, obtained from the comparison of correlation matrices, confirms that the inter-variable dependency structure is largely preserved. This indicates that statistically coherent relationships among physically related meteorological variables (e.g., temperature–evaporation, precipitation–evaporation) are maintained in the synthetic dataset as well. Therefore, the proposed synthetic data generation framework successfully reproduces both the marginal distributions and the multivariate correlation structure of the real dataset. This finding supports the potential use of the model as a reliable synthetic data generator for data augmentation and privacy-preserving applications.

To visually assess the statistical similarity between the real and synthetic datasets, the distribution of each variable was compared. Figure 5 presents the histogram of the precipitation variable as an example. Overlaying the real and synthetic distributions provides a visual confirmation of how accurately the model reproduces the original data patterns. This comparison not only supports the numerical test results but also demonstrates the successful preservation of distributional shape, further validating the effectiveness of the synthetic data generation process.

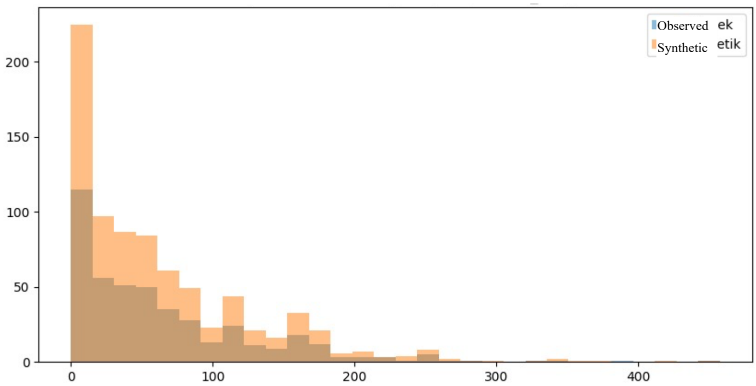


Figure 5: Distribution comparison of observed and synthetic precipitation data

As illustrated in Figure 5, both datasets exhibit a right-skewed distribution, which reflects a typical meteorological pattern in which low precipitation values occur frequently, while higher precipitation events are relatively rare. In the 0–100 mm range, the density of the synthetic samples closely matches that of the observed dataset, indicating that the model successfully learned the sampling behavior of low-to-moderate rainfall conditions. The slight deviations observed at the upper tail (above 200 mm) are likely due to the rarity of extreme precipitation events and the limited ability of generative models to reproduce very sparse observations. This visual finding is consistent with the high KS p-value ($p = 0.996$) and the low WD distance ($WD = 1.205$), both of which confirm a strong statistical similarity for the precipitation variable.

Therefore, when Table 4 and Figure 5 are evaluated together, both the numerical test results and the histogram comparison demonstrate that the proposed synthetic data generation approach is capable of accurately reproducing the probability distributions of individual variables. This confirms that the model can generate physically meaningful data while preserving the structural characteristics of hydrometeorological processes.

Following the validation of the synthetic dataset, the ML_AUG models were retrained, and their performance metrics are presented in Table 5.

Table 5: Performance metrics of ML_AUG Models

Model	RMSE (mm)	R²
RF	35.81	0.874
XGBoost	33.12	0.892
LightGBM	35.76	0.874
CatBoost	39.57	0.846
SVR	57.17	0.679
MLP	38.78	0.852
SE	38.34	0.856

According to Table 5, XGBoost achieved the highest accuracy after data augmentation, with an RMSE of 33.12 mm and an R^2 of 0.892. RF and LightGBM exhibited similar performance levels, both yielding approximately 35.8 mm RMSE and 0.874 R^2. The CatBoost model produced lower accuracy (RMSE = 39.57 mm, $R^2 = 0.846$), while MLP and SE models demonstrated moderate performance with RMSE values of around 38 mm and R^2 values close to 0.85. The lowest accuracy was obtained by the SVR model (RMSE = 57.17 mm, $R^2 = 0.679$). These results show that, following data augmentation, XGBoost emerged as the most accurate and stable algorithm, and the performance gap between models became more pronounced.

To examine the effect of data augmentation under different climatic conditions, the performance of the models was also evaluated separately for summer and winter seasons. The corresponding results are presented in Table 6.

Table 6: Seasonal Performance (Summer/Winter)

Model	Summer		Winter	
	RMSE (mm)	R²	RMSE (mm)	RMSE
RF	29.06	0.924	29.46	0.921
XGBoost	32.26	0.906	30.80	0.914
LightGBM	29.45	0.922	33.84	0.896
CatBoost	30.57	0.916	41.22	0.846
SVR	60.25	0.672	63.53	0.634
MLP	40.11	0.855	35.18	0.888
SE	31.30	0.912	31.74	0.909

According to Table 6, RF (RMSE = 29.06 mm, $R^2 = 0.924$) and LightGBM (RMSE = 29.45 mm, $R^2 = 0.922$) achieved the highest accuracy during the summer period, followed by CatBoost and XGBoost. SVR was clearly the weakest model in summer. In winter, although error values increased overall, RF (RMSE = 29.46 mm, $R^2 = 0.921$) and XGBoost (RMSE = 30.80 mm, $R^2 = 0.914$) produced comparatively more stable results. CatBoost experienced a marked decline in performance during winter (RMSE = 41.22 mm, $R^2 = 0.846$), while SVR remained the lowest-performing model in both seasons. Overall, these findings indicate that, after data augmentation, RF and XGBoost are the most seasonally consistent and stable models.

The feature importance analysis presented for ML_BASE was repeated for the augmented version, XGBoost_AUG (Figure 5), and the ranking of variables remained largely unchanged. The dominance of Temperature as the most important predictor demonstrates that a major portion of the model's explanatory power originates from temperature. The second-highest contribution belongs to Evap_lag1, confirming the role of temporal persistence in evaporation prediction. The relatively small contributions of the remaining variables—Temp_mean3 (0.087), Precipitation_mm (0.020), Precip_sum3 (0.0079), Temp_lag1 (0.0054), and Precip_lag1 (0.0035)—are clearly observable in Figure 6. Thus, despite the application of data augmentation, the model behavior remained strongly temperature-driven, indicating that the inclusion of synthetic samples did not disrupt the original variable importance structure.

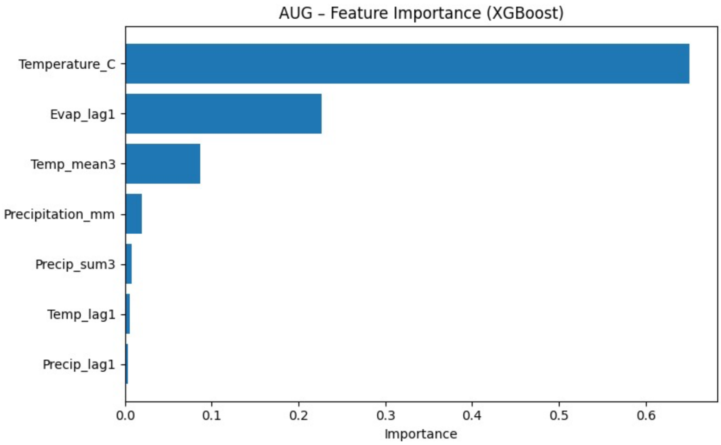


Figure 6: Feature importance for XGBoost_AUG

Following these analyses, the model's ability to capture dynamic behavior was evaluated through a time-series examination. Figure 7 presents the temporal comparison between the evaporation estimates produced by the XGBoost_AUG model and the observed measurements.

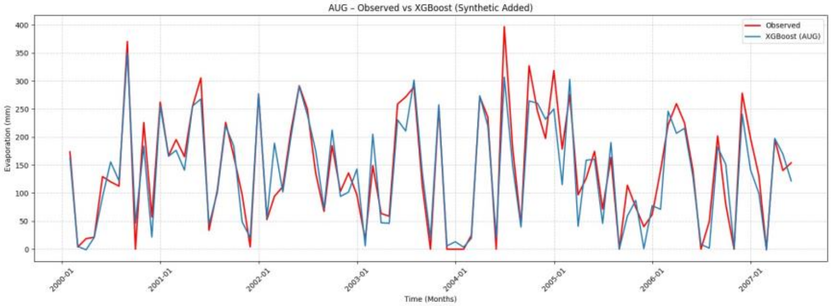


Figure 7: Time-series of the XGBoost_AUG model

As illustrated in Figure 7, the predicted and observed evaporation curves progress in close alignment, with the model successfully capturing peak values during high-evaporation periods. Although minor deviations are observed in certain months, the model generally follows the overall trend and preserves the temporal variability of the series. These findings indicate that the data augmentation process did not disrupt the dynamic structure of the model; on the contrary, it contributed to a more stable representation of temporal behavior.

Although the time-series evaluation in Figure 7 highlights that the XGBoost_AUG model maintains the temporal coherence of the evaporation signal, such visual assessments alone cannot fully describe the statistical alterations introduced by the augmentation strategy. For a more structured examination, the models were further analyzed using Taylor diagrams, which summarize key statistical attributes such as variability, linear association, and error structure within a unified framework. This approach enables the identification of subtle shifts in model dispersion and correlation patterns that may not be directly observable in time-domain analyses. Accordingly, Figure 8 presents the Taylor diagrams for the AUG configuration, offering insight into how data augmentation shaped the statistical behavior of the models during both the training and testing phases.

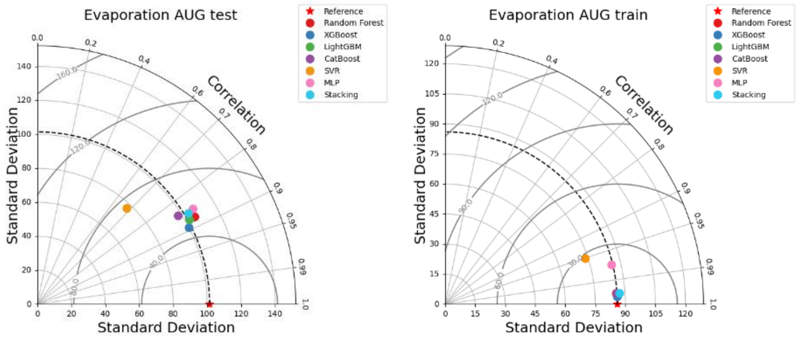


Figure 8: Taylor diagrams illustrating the
statistical performance of the AUG models

CONCLUSIONS

In this study, a hybrid approach was developed to address data scarcity in monthly evaporation prediction by combining machine learning models with synthetic data generation. In the first stage, ML_BASE models trained solely on observed data were evaluated, and XGBoost, SE, and LightGBM achieved the highest accuracy. XGBoost provided the best overall performance (RMSE = 33.24 mm, R^2 = 0.891), confirming that tree-based models are more stable for hydrometeorological processes.

In the second stage, a cVAE–cWGAN-GP–based synthetic data generation framework was applied to increase the statistical diversity of the dataset. KS and Wasserstein Distance analyses showed that the synthetic samples successfully reproduced the real data distributions, while the low Frobenius difference confirmed that the inter-variable correlation structure was largely preserved. Histogram comparisons also demonstrated a strong overlap between real and synthetic values, particularly in low-to-moderate evaporation ranges.

After augmentation, ML_AUG models were retrained, and XGBoost again achieved the highest performance (RMSE = 33.12 mm, R^2 = 0.892). Seasonal evaluations revealed that RF and XGBoost were the most stable models in both summer and winter periods. The preservation of feature-importance ranking after augmentation indicates that synthetic samples did not distort the physical relationships between variables but instead expanded the learning domain safely. Time-series comparisons further verified that peak evaporation events were accurately captured, and the temporal dynamics of the series were consistently represented.

Overall, the findings demonstrate that (i) tree-based models offer strong predictive skill for evaporation, (ii) cVAE–cWGAN-GP–based augmentation improves model robustness without altering the physical structure of the data, and (iii) synthetic data integration is an effective solution for increasing accuracy and stability in data-limited hydrological applications. This approach can serve as a reliable alternative in basins where data availability or data privacy is a constraint.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and & Zheng, X. (2016). {TensorFlow}: a system for {Large-Scale} machine learning. In 12th USENIX symposium on operating systems design and implementation (OSDI 16) (pp. 265-283).

Al-Mukhtar, M., Elbeltagi, A., Kushwaha, N. L., & Vishwakarma, D. K. (2022). Monthly pan evaporation modelling using hybrid machine learning algorithms in a semi-arid environment.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875.

Atilgan, A., Yucel, A., & Markovic, M. (2022). Statistical analyses and hydrometeorological parameters of long-year water levels and volumes: example of Lake Egirdir.

Bakhshipour, A. E., Koochali, A., Dittmer, U., Haghighi, A., Ahmad, S., & Dengel, A. (2023). A bayesian generative adversarial network (gan) to generate synthetic time-series data, application in combined sewer flow prediction. arXiv preprint arXiv:2301.13733.

Deo, R. C., Samui, P., & Kim, D. (2016). Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. Stochastic Environmental Research and Risk Assessment, 30(6), 1769-1784.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. Advances in neural information processing systems, 30.

Jahangir, M. S., & Quilty, J. (2023). Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. Journal of Hydrology, 624, 130498. https://doi.org/10.1016/j.jhydrol.2023.130498.

Karaman, M. E. (2010). The tectonic evolution of Lake Eğirdir, west Turkey. Geologos, 16(4), 223-234.

Kazancı, N., & Roberts, N. (2019). The lake basins of South-west Anatolia. In Landscapes and Landforms of Turkey (pp. 325-337). Cham: Springer International Publishing.

Khandelwal, A., Xu, S., Li, X., Jia, X., Stienbach, M., Duffy, C., Nieber, J., & Kumar, V. (2020). Physics guided machine learning methods for hydrology. arXiv preprint arXiv:2012.02854.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kumar, V., Kedam, N., Sharma, K. V., Mehta, D. J., & Caloiero, T. (2023). Advanced machine learning techniques to improve hydrological prediction: A comparative analysis of streamflow prediction models. Water, 15(14), 2572.

Lange, H., & Sippel, S. (2020). Machine learning applications in hydrology. In Forest-water interactions (pp. 233-257). Cham: Springer International Publishing.

Liang, Y., Feng, D., & Sun, Z. (2023). Comparison of four machine learning models for forecasting daily reference evaporation based on public weather forecast data. Hydrology and Earth System Sciences Discussions, 2023, 1-57.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning (pp. 1278-1286). PMLR.

Risha, M., & Liu, P. (2025). Data-Driven Facies Prediction: A Comparative Study of Random Forest, XGBoost, SVM, CatBoost, and K-Means.

Şener, Ş., Davraz, A., & Karagüzel, R. (2013). Evaluating the anthropogenic and geologic impacts on water quality of the Eğirdir Lake, Turkey. Environmental earth sciences, 70(6), 2527-2544.

Shabani, S., Samadianfard, S., Sattari, M. T., Mosavi, A., Shamshirband, S., Kmet, T., & Várkonyi-Kóczy, A. R. (2020). Modeling pan evaporation using Gaussian process regression, K-nearest neighbors, random forest and support vector machines; comparative analysis. Atmosphere, 11(1), 66.

Yonekura, K., Tomori, Y., & Suzuki, K. (2024). Airfoil Shape Generation and Feature Extraction Using the Conditional VAE-WGAN-gp. AI, 5(4), 2092-2103.

Zerouali, B., Bailek, N., Bouchouich, K., Mawloud, G., Kuriqi, A., Sami Khafaga, D., Alharbi, A.H.& El-kenawy, E. S. M. (2025). Enhancing water security through advanced modeling: integrating deep learning and a novel metaheuristic optimization algorithm for accurate pan evaporation prediction. AQUA—Water Infrastructure, Ecosystems and Society, 74(1), 18-35.

Chapter 16

Streamflow Prediction Using Multi-Frequency Analysis of Time Series: A Wavelet–LSTM Model

Emine Dilek TAYLAN[1], **Özlem TERZİ**[2]

ABSTRACT

In this study, Long Short-Term Memory (LSTM) networks and wavelet transform-based hybrid models (Wavelet–LSTM) were developed for monthly streamflow prediction at the E05A024 (Sazköy) streamflow gauging station located in the Gediz Basin. The model inputs consisted of precipitation, temperature, streamflow, and derived hydro-meteorological variables with 1–6-month lags and accumulated values. First, an LSTM+ model was constructed using a 24-month time window and a two-layer bidirectional LSTM architecture. The model performed well under low and medium flow conditions but showed limitations in predicting extreme flows (test NSE = 0.641, KGE = 0.520). To improve prediction accuracy, Wavelet–LSTM models were developed by decomposing the time series using five different wavelet types (Sym5, Db4, Dmey, Haar, Coif3). Performance comparisons revealed that the model constructed with the Sym5 wavelet achieved the highest accuracy on the test set (NSE = 0.963, RMSE = 0.545). Seasonal analyses showed that the model performed well during winter, spring, and summer, while its accuracy decreased during the transitional autumn period. The findings demonstrate that Wavelet–LSTM models provide a powerful alternative for monthly streamflow prediction by better representing seasonal dynamics.

Keywords – Streamflow prediction; LSTM+; Wavelet analysis; Gediz Basin; Turkey.

---

[1] Prof.Dr.; Department of Civil Engineering, Faculty of Engineering and Natural Sciences, Suleyman Demirel University, Isparta, Turkey. dilektaylan@sdu.edu.tr ORCID No:0000-0003-0734-1900

[2] Prof. Dr.; Department of Civil Engineering, Faculty of Technology, Isparta University of Applied Sciences, Isparta, Turkey. ozlemterzi@isparta.edu.tr ORCID No:0000-0001-6429-5176 (Corr. author)

INTRODUCTION

Accurate streamflow prediction plays a crucial role in the effective management of water resources, including flood control, drought mitigation, irrigation scheduling, and hydropower generation. Reliable forecasts are essential for supporting decision-making processes, particularly in basins where seasonal variability and extreme hydrological events are significant (Slater and Villarini, 2018; Troin et al., 2021). Traditional statistical approaches, while useful, often face limitations when capturing the non-linear and non-stationary characteristics of hydrological time series. In recent decades, machine learning and deep learning methods have gained increasing attention for their ability to model such complex patterns (Lin et al., 2021).

Among these, Long Short-Term Memory (LSTM) networks have emerged as powerful tools for sequential data modeling, providing accurate forecasts by learning temporal dependencies within time series. However, when applied to raw hydrological data, LSTM models may struggle to effectively capture multi-frequency fluctuations and sudden peaks, especially in regions with pronounced seasonal cycles (Xie et al., 2022). To address these challenges, hybrid modeling approaches that combine signal decomposition techniques with deep learning architectures have been proposed.

Wavelet transform is one of the most widely used decomposition methods, offering the ability to separate a time series into multiple frequency components and reveal both short-term variations and long-term trends. When integrated with LSTM networks, wavelet-based hybrid models can significantly improve forecasting performance by enabling the model to learn from different frequency scales independently (Wang et al., 2022; Guo et al., 2023). Numerous studies have reported that such hybrid models outperform conventional approaches in hydrological forecasting tasks (Wei and You, 2022; Muhammad et al., 2019).

Despite this growing body of research, limited attention has been given to the comparative performance of different wavelet types within hybrid Wavelet–LSTM models. Most studies select a single wavelet family without systematic evaluation, which may overlook the influence of wavelet characteristics on model accuracy. In this context, investigating how different wavelet types affect forecasting performance is essential for identifying the most suitable configurations for a given hydrological setting (Shoaib et al., 2014; Barzegar et al., 2017; Barzegar et al., 2021).

This study focuses on the Gediz Basin (Turkey) and aims to compare the performance of multiple wavelet types in monthly streamflow prediction using hybrid Wavelet–LSTM models. The novelty of this study lies in the limited application of hybrid Wavelet–LSTM models for monthly streamflow prediction in the Gediz Basin, incorporating different wavelet types and comprehensively evaluating their performance within the basin's seasonal hydrological dynamics. After introducing the methodological framework, the study evaluates various wavelet families (Sym5, Db4, Dmey, Haar, Coif3) and analyzes their performance based on multiple accuracy metrics and seasonal dynamics. The findings provide insights into the role of wavelet selection in hydrological forecasting and offer guidance for future applications in similar basins.

STUDY AREA

This study was conducted within the boundaries of the Gediz Basin, one of the major river basins in Turkey. The basin is in the central part of the Aegean Region in Western Anatolia and covers an area of approximately 17,500 km² (Kumanlioglu, 2020). It encompasses the provinces of Kütahya, Uşak, Manisa, and İzmir, and is bordered by the Bakırçay Basin to the north, the Afyonkarahisar Plateau to the east, the Büyük Menderes Basin to the south, and the Aegean Sea to the west. The main river system of the basin, the Gediz River, originates from Murat Mountain and flows for approximately 401 km, joining several tributaries along its course before discharging into the Gulf of İzmir and eventually the Aegean Sea. Geographically, the basin is situated between 38°04′–39°13′ N latitudes and 26°42′–29°45′ E longitudes.

The Gediz Basin has a Mediterranean climate, characterized by hot and dry summers and mild, rainy winters. Precipitation is concentrated mainly in the winter and spring months, while it is significantly lower during the summer period. Annual average precipitation ranges between 500–600 mm in the eastern parts of the basin and decreases westward toward the coastal areas, where it is around 400 mm. The maximum recorded annual mean precipitation is 622.5 mm. The highest monthly average precipitation generally occurs in December–January, whereas the lowest values are observed in July–August. The annual mean temperature ranges between 12–16 °C. This pronounced seasonality leads to a high degree of variability in the flow regime, with high flows in winter and low flows in summer (GBDMP, 2019; Taylan, 2024).

Within the scope of this study, the Uşak meteorological station and the Sazköy streamflow gauging station (E05A024), which are located close to each other in the eastern part of the basin, were evaluated (Figure 1). For these stations, monthly precipitation (P) and air temperature (Ta) data for the period 1985–2015 were obtained from the Turkish State Meteorological Service, while monthly streamflow (Q) data were obtained from the General Directorate of State Hydraulic Works. Statistics of the variables are given in Table 1.

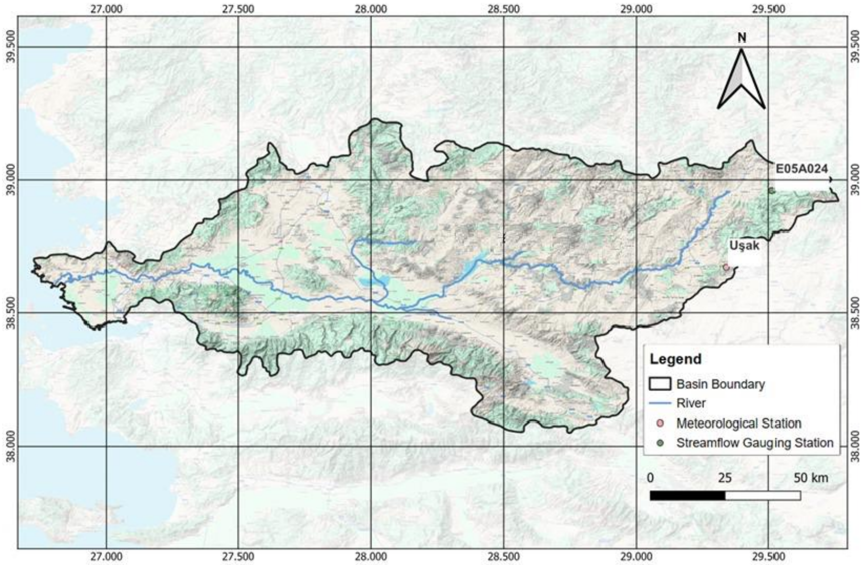


Figure 1: Study area and location of the stations (Taylan, 2024)

Table 1: The descriptive statistics for used parameters (Taylan, 2024)

Parameters	Station	Mean	Max	Min
Streamflow (m^3/s)	E05A024	1.56	17.20	0.00
Precipitation (mm)	Uşak	45.71	211.70	0.00
Air Temperature (^0C)	Uşak	12.70	27.00	-2.10

With its pronounced seasonal cycles, complex hydrological structure, and significant human influences, the Gediz Basin provides an ideal case study area for the development and testing of seasonal streamflow forecasting models. Furthermore, the availability of long-term hydro-meteorological datasets enhances the applicability of data-driven modeling approaches in the basin.

LSTM+

Long Short-Term Memory (LSTM) networks were developed to address the vanishing gradient problem that recurrent neural networks (RNNs) face when learning long-term dependencies in sequential data (Hochreiter and Schmidhuber, 1997). LSTM cells incorporate input, forget, and output gates along with a cell state, which enables the network to selectively retain, update, or discard information from past time steps.

This gating mechanism enables LSTM networks to maintain long-term contextual dependencies more effectively than conventional RNNs or feedforward architectures, leading to superior performance in modeling complex sequential data such as time series or natural language sequences (Krichen and Mihoub, 2025; Ohno and Kumagai, 2021). Consequently, LSTMs have become a popular choice for forecasting tasks that involve nonlinearity, seasonality, and lagged effects (Krichen and Mihoub, 2025; Jothi, 2025).

Although standard LSTM architectures are powerful, they often rely on a limited set of input features and operate with unidirectional information flow, which can limit their ability to accurately capture seasonal patterns, delayed effects, and high-magnitude events.

To overcome these limitations, an enhanced version of the LSTM model—referred to as LSTM+—was developed in this study. The LSTM+ model integrates several architectural and functional improvements, summarized as follows:

- Feature Engineering: Input variables were transformed and enriched with additional derived features to better represent underlying temporal structures.
- Dual-layer BiLSTM Structure: Bidirectional LSTM (BiLSTM) layers were employed to process information in both forward and backward directions, enabling more comprehensive learning of temporal dependencies.
- Optimized Loss Functions: Specialized loss functions were designed to improve the model's ability to capture extreme values and peak behaviors.

Previous studies have demonstrated that similar enhanced LSTM architectures can provide substantial improvements in predictive accuracy compared to standard models across various time series forecasting tasks (Guo et al., 2019; Eniola et al., 2025).

WAVELET ANALYSIS

Wavelet analysis is a powerful mathematical technique that enables a detailed examination of time series in both the time and frequency domains (Daubechies, 2002; Torrence and Compo, 1998). Unlike the classical Fourier transform, which provides only frequency information, the wavelet transform allows the identification of frequency components that vary over time. This feature makes wavelet methods particularly useful for analyzing nonlinear and nonstationary processes (Ogbonna et al., 2016).

Given a time series $x(t)$, the Continuous Wavelet Transform (CWT) can be expressed as follows:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^{*}\left(\frac{t-b}{a}\right) dt \tag{1}$$

where a is the scale parameter, b is the translation parameter, and ψ denotes the mother wavelet. Large scale values correspond to low-frequency (long-period) components, while small scales correspond to high-frequency (short-period) components (Mallat, 2002).

Wavelet analysis can be applied in two main forms:

- Continuous Wavelet Transform (CWT): Provides a continuous time-frequency representation and is particularly useful for wavelet coherence analysis.
- Discrete Wavelet Transform (DWT): Decomposes the time series into multiple scales and allows for reconstruction. Using the multiresolution analysis framework (Mallat, 2002; Percival and Walden, 2000; Hong-fa, 2012), the time series is decomposed into approximation (A) and detail (D) components.

The DWT enables decomposition up to J levels depending on the data length and the selected wavelet type. High-frequency short-term fluctuations are isolated at lower levels, while long-term structures are captured at higher levels. The decomposition can be expressed as Equation (2):

$$x(t) = A_J(\mathrm{t}) + \sum_{j=1}^{J} D_J(\mathrm{t}) \tag{2}$$

where $A_J(t)$ represents the lowest-frequency approximation component and $D_j(t)$ corresponds to the detail component at level j. These components can then be analyzed separately or used as inputs to predictive models, such as Wavelet–LSTM hybrids (Bunrit et al., 2018).

The choice of the mother wavelet plays a critical role in the success of the decomposition. Different wavelet families — such as Symlet, Daubechies, Coiflet, Haar, or Dmeyer — can yield different results depending on the characteristics of the time series (Addison, 2017). The selection is typically based on empirical testing or statistical performance criteria (e.g., RMSE, NSE, KGE) (Kaddar & Fizazi, 2017).

Different wavelet families exhibit distinct mathematical properties that influence the decomposition performance. Daubechies wavelets (dbN) provide compact support and increasing smoothness with order but are asymmetric (Daubechies, 2002). Symlets (symN) are nearly symmetric versions that reduce phase distortion while maintaining good localization (Mallat, 2002; Addison, 2017). Coiflets (coifN) offer high smoothness and excellent frequency localization due to vanishing moments of both scaling and wavelet functions (Percival & Walden, 2000). The Haar wavelet is simple and efficient for detecting abrupt changes but lacks smoothness. Finally, the Discrete Meyer (Dmey) wavelet is smooth and symmetric with well-defined frequency characteristics, making it suitable for analyzing long-term components (Addison, 2017).

By decomposing a signal into multiple components, wavelet analysis can isolate and capture underlying temporal patterns more effectively. When integrated with machine learning models, this hybrid approach often leads to more accurate predictions than using raw time series data alone (Partal & Kisi, 2007; Nourani et al., 2011; Deo & Şahin, 2015).

RESULTS AND DISCUSSION

In this study, LSTM+ and Wavelet–LSTM algorithms were applied for streamflow prediction at the E05A024 (Sazköy) gauging station, located in the Gediz Basin. For this purpose, precipitation, and temperature data from the Uşak meteorological station for the period 1986–2015 were used. The input dataset for the streamflow forecasting models consisted of precipitation (mm), air temperature (°C), monthly streamflow (m^3/s), and several derived hydro-meteorological variables. Specifically, the inputs included 1- and 2-month lagged values for each variable, 3- and 6-month cumulative precipitation, 3-month moving averages of temperature and streamflow, and the Antecedent Precipitation Index (API) calculated using a recession constant of k = 0.85. In addition, to capture seasonality, the month information was transformed into sine and cosine components using a trigonometric transformation method. In total, 16 input variables were used to train the models.

To evaluate the performance of the developed models, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), Nash–Sutcliffe Efficiency (NSE), and Kling–Gupta Efficiency (KGE) were employed as performance metrics. The formulations of RMSE, MAE, and NSE are provided in Equations (3)–(5). Here, N denotes the total number of data points, $Q_{i(obs)}$ is the observed streamflow, and $\mathrm{Q_{i(sim)}}$ is the simulated streamflow. $\bar{Q}_{i(obs)}$ *and* $\bar{Q}_{i(sim)}$ represent the mean values of the observed and simulated streamflows, respectively.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Q_{i(obs)} - Q_{i(\mathrm{sim})}\right)^2} \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left(Q_{i(obs)} - Q_{i(\mathrm{sim})}\right)^2 \tag{4}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(Q_{i(obs)} - Q_{i(sim)}\right)^2}{\sum_{i=1}^{N}\left(Q_{i(obs)} - \bar{Q}_{i(obs)}\right)^2} \tag{5}$$

The calculation of the Kling–Gupta Efficiency (KGE) is presented in Equation (6). In this equation, r represents the Pearson correlation coefficient, which measures the linear relationship between observed and simulated values; α denotes the variability ratio, calculated as the ratio of the standard deviation of the simulated values (σ) to that of the observed values; and β represents the bias ratio, calculated as the ratio of the mean of the simulated values (μ) to the mean of the observed values.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{6}$$

$$\alpha = \frac{\sigma^{\mathrm{sim}}}{\sigma^{obs}} \qquad \beta = \frac{\mu^{sim}}{\mu^{obs}} \tag{7}$$

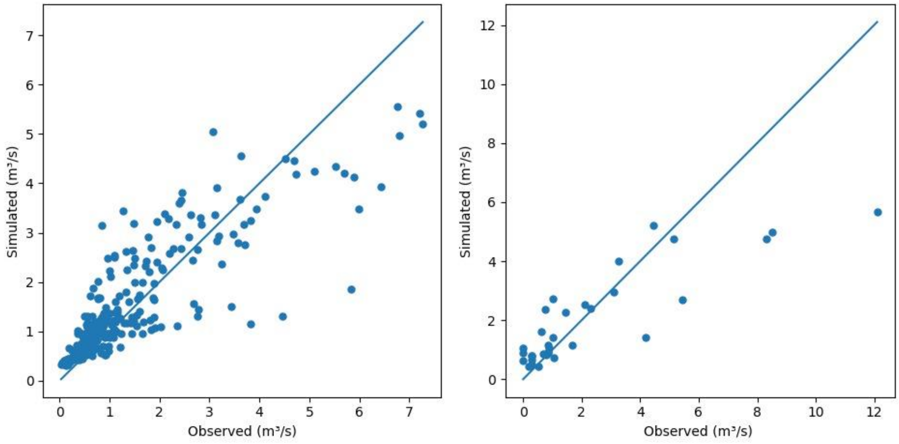
$$r = \frac{\sum_{t=1}^{N}(Q_{i(obs)} - \bar{Q}_{i(obs)})(Q_{i(sim)} - \bar{Q}_{i(sim)})}{\sqrt{\sum_{t=1}^{N}(Q_{i(obs)} - \bar{Q}_{i(obs)})^2 \sum_{t=1}^{N}(Q_{i(sim)} - \bar{Q}_{i(sim)})^2}} \tag{8}$$

LSTM+ Model

In this study, an LSTM+ model was first developed using an input structure with 16 variables and additional preprocessing steps. The model was designed to make predictions using a time window of 24 months (SEQ_LEN = 24). The LSTM+ model is based on a two-layer bidirectional biLSTM architecture. The first LSTM layer consists of 64 neurons and uses the parameter return_sequences=True to pass the entire sequence to the next layer. The second LSTM layer consists of 32 neurons. To prevent overfitting, a 20% dropout was applied after each layer. Following the LSTM layers, there is a dense layer with 32 neurons and ReLU activation, and a single-neuron output layer.

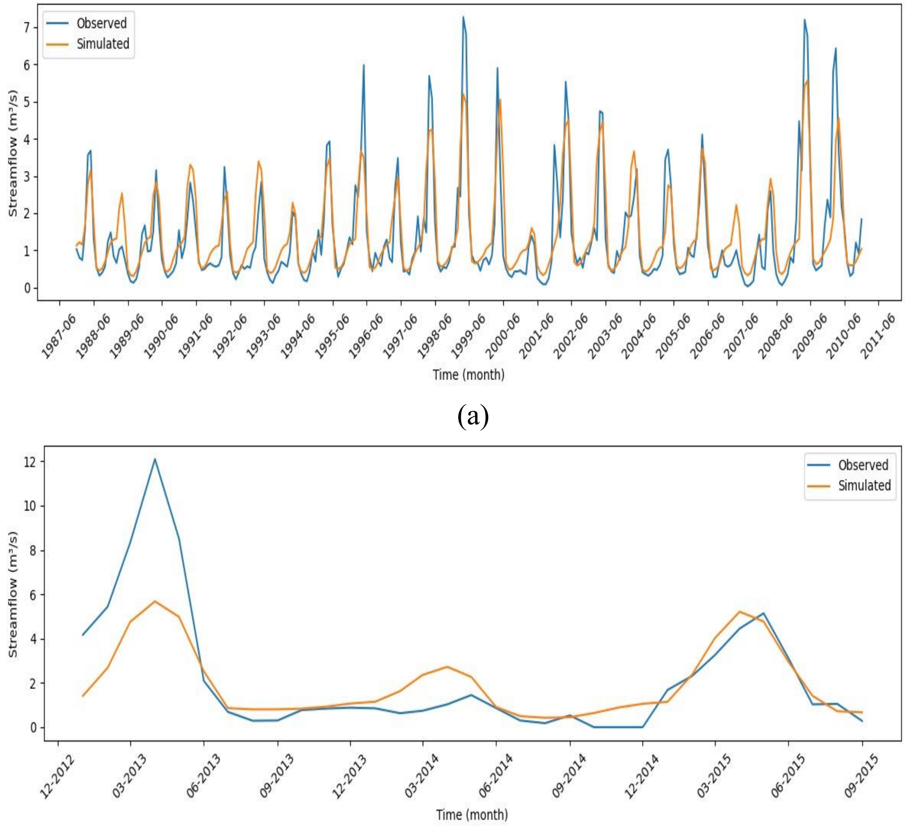
For optimization, the Adam algorithm was used with a learning rate of 0.001. To reduce the influence of extreme values, Huber Loss was chosen as the loss function. During training, EarlyStopping and ReduceLROnPlateau callbacks were applied to prevent overfitting and improve model generalization. All input data were normalized using Min–Max scaling. The dataset was split into a training period (1985–2010) and a testing period (2011–2015).

Because SEQ_LEN = 24, the first 24 months of data were used only as the input sequence, and no predictions were generated for this warm-up period. Due to this initial input window, up to 36 months of predictions were expected for the test period (2011–2015). However, because of the lagged input variables, 33 months of test values were available for evaluation. The scatter and time series plots of the LSTM+ model for the training and testing sets are presented in Figures 2 and 3, respectively.



(a) (b)

Figure 2: Scatter plot of the LSTM+ model: a) training set b) test set



(a)

(b)

Figure 3: Time series plot of the LSTM+ model: a) training set b) test set

In the scatter plots shown in Figure 2, most of the points are concentrated around the 1:1 line. While deviations are limited for low and medium flows, the scatter increases with higher flow values, and the model exhibits a tendency to underestimate extreme flows. The time series comparison in Figure 3 indicates that during the training period, the model successfully captures the seasonal fluctuations and overall trends of the streamflow, and it accurately follows low to medium flow levels, particularly during the summer and dry periods. However, for peak flows, the model tends to slightly underestimate the peak magnitudes.

These visual findings are consistent with the training performance metrics, which are RMSE = 0.747, MAE = 0.493, NSE = 0.712, and KGE = 0.750. This indicates that the model generally captures the mean, variance, and correlation components well, although uncertainty increases for extreme flows. The test metrics are RMSE = 1.689, MAE = 1.010, NSE = 0.641, and KGE = 0.520. The performance observed in the training set should be regarded as an upper bound for model skill, and methods such as Wavelet–LSTM approaches or peak-focused loss functions may be used to improve peak flow prediction accuracy.

Wavelet–LSTM Models

The results indicated that the LSTM+ model demonstrated strong predictive capability under low and medium flow conditions, but there was room for improvement in forecasting extreme flows. To enhance prediction performance, hybrid Wavelet–LSTM models were developed by combining wavelet transformations with the LSTM architecture. This hybrid approach enables the decomposition of time series into multiple frequency components, allowing both short-term fluctuations and long-term trends to be separately introduced to the model.

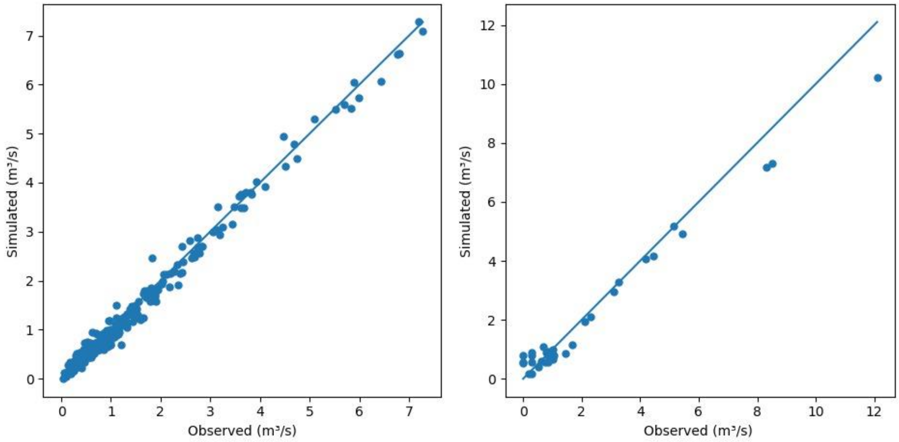
Before developing the Wavelet–LSTM models, the time series was decomposed using five different wavelet types—Sym5, Db4, Dmey, Haar, and Coif3—to capture its multi-frequency structure. For each wavelet type, the original streamflow series was decomposed into several sub-series representing low-, medium-, and high-frequency components. Correlation coefficients between each sub-series and the original series were then calculated to quantitatively assess the explanatory power of each component. Subsequently, the sub-series with the highest correlation values were selected as model inputs, and the LSTM architecture was configured accordingly. This procedure aimed to improve the model's ability to represent both short-term variability and seasonal patterns in the streamflow data. The performance metrics of the developed Wavelet–LSTM models are presented in Table 2.

Table 2: Performance metrics of the Wavelet–LSTM models

Wavelet types	**Training set**				**Testing set**			
	RMSE	**MAE**	**NSE**	**KGE**	**RMSE**	**MAE**	**NSE**	**KGE**
Sym5	0.153	0.114	0.988	0.965	0.545	0.374	0.963	0.841
Db4	0.633	0.407	0.793	0.806	1.318	0.925	0.781	0.604
Dmey	0.166	0.115	0.986	0.981	0.590	0.451	0.956	0.842
Haar	0.266	0.165	0.963	0.971	0.616	0.441	0.952	0.854
Coif3	0.563	0.339	0.837	0.881	1.730	1.105	0.623	0.476

Table 2 presents the performance metrics of the Wavelet–LSTM models developed using different wavelet types for both the training and testing sets. The results indicate that the choice of wavelet type has a significant impact on model accuracy. The Sym5 wavelet produced the lowest RMSE (0.153 and 0.545) and MAE (0.114 and 0.374) values in both the training and testing sets. In addition, the NSE values reached the highest levels with 0.988 for the training set and 0.963

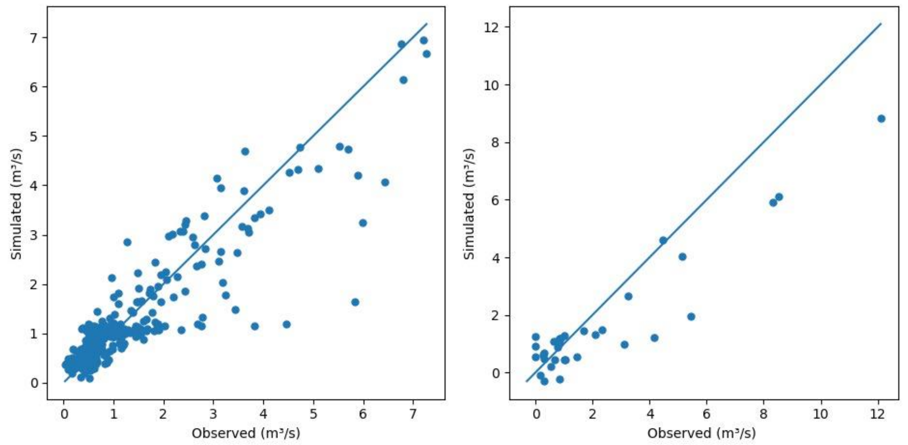
for the testing set. This demonstrates that the model constructed with the Sym5 wavelet outperformed the other wavelet types during both the learning and generalization stages. Although the Dmey and Haar wavelets also achieved relatively high NSE and KGE values, they did not perform as well as Sym5 in terms of error metrics. In contrast, models developed using Db4 and Coif3 wavelets exhibited high RMSE values in the test set, indicating lower generalization capability. Overall, the Sym5 wavelet stands out as the most suitable wavelet type for improving the time series forecasting performance of the Wavelet–LSTM model. Figures 4–8 provide a comparative visualization of the scatter plots for the Wavelet–LSTM models developed using these wavelet types for both the training and testing sets, allowing for a detailed performance evaluation.



(a) (b)

Figure 4: Scatter plot of the best model for the Sym5 wavelet type: a) training set b) test set

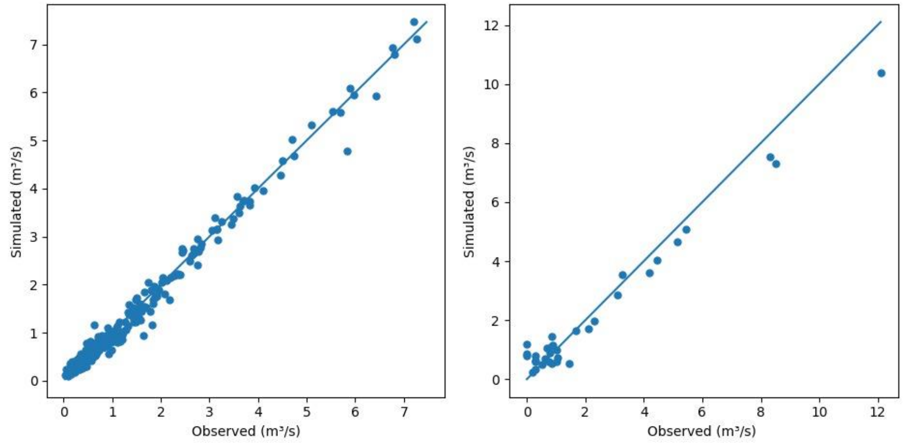
As shown in Figure 4, the Wavelet–LSTM model developed using the Sym5 wavelet demonstrated the best performance for both the training and testing sets. The scatter plots reflect a strong linear relationship between the observed and predicted values. The points are concentrated around the 1:1 line, and deviations, particularly for high flow values, remain at a minimum level.



(a) (b)

Figure 5: Scatter plot of the best model for the Db4 wavelet type: a) training set b) test set

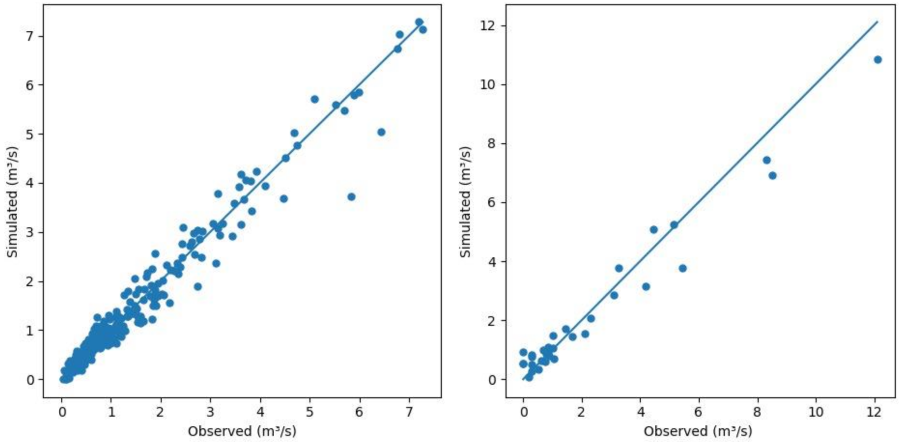
The results obtained for the Db4 wavelet (Figure 5) show a weaker performance compared to Sym5. In particular, the scatter plots reveal greater deviations from the 1:1 line in the test set, along with increased dispersion for low and medium flow values.



(a) (b)

Figure 6: Scatter plot of the best model for the Dmey wavelet type: a) training set b) test set

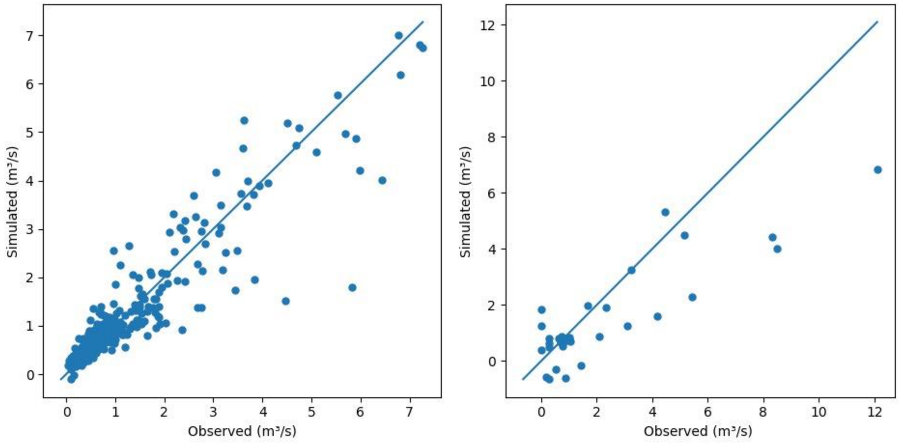
The model developed using the Dmey wavelet (Figure 6) shows very high accuracy for the training set and demonstrates good generalization performance for the test set as well. The scatter plot indicates that both low and high flow values are well represented, with data points distributed close to the 1:1 line.



(a) (b)

Figure 7: Scatter plot of the best model for the Haar wavelet type: a) training set b) test set

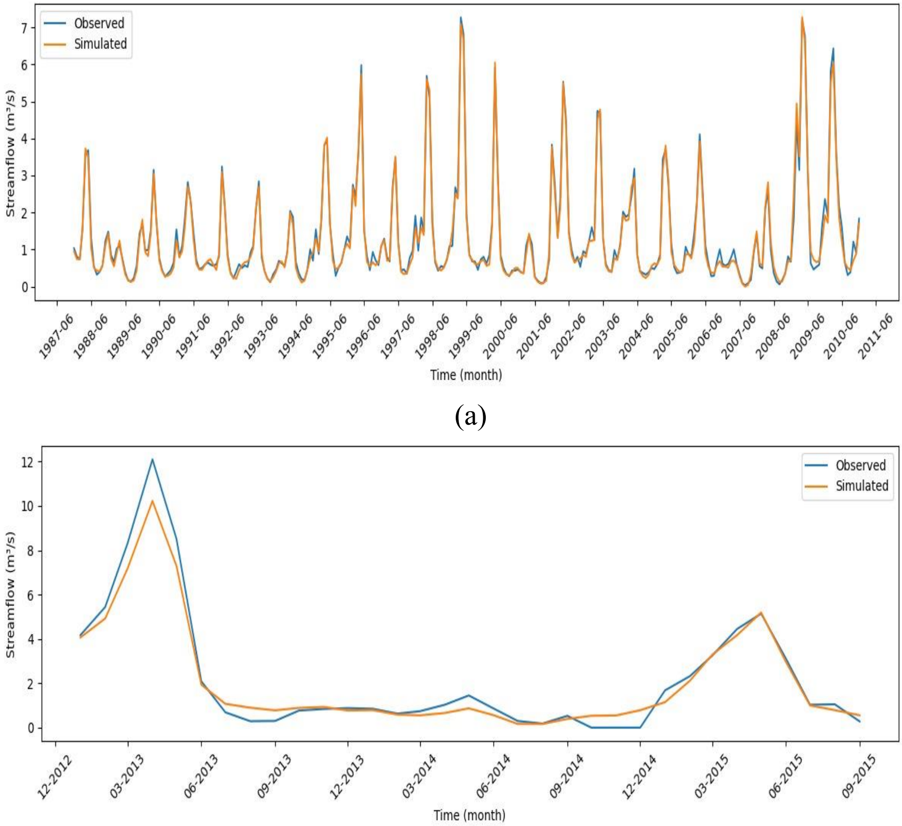
The results obtained for the Haar wavelet (Figure 7) indicate a stable performance, although not as strong as Sym5 and Dmey. The NSE values were found to be 0.963 for the training set and 0.952 for the test set (Table 2). Although the scatter plot shows slight deviations at high flow values, it generally reflects a linear relationship.



(a) (b)

Figure 8: Scatter plot of the best model for the Coif3 wavelet type: a) training set b) test set

The model developed using the Coif3 wavelet (Figure 8) showed the lowest performance for both the training and testing sets. The scatter plots reveal that the data points deviate significantly from the 1:1 line, with increasing prediction errors for high flow values in the test set. The test NSE value was 0.623 and the KGE value was 0.476, indicating a substantial decrease in performance compared to the other wavelet types.

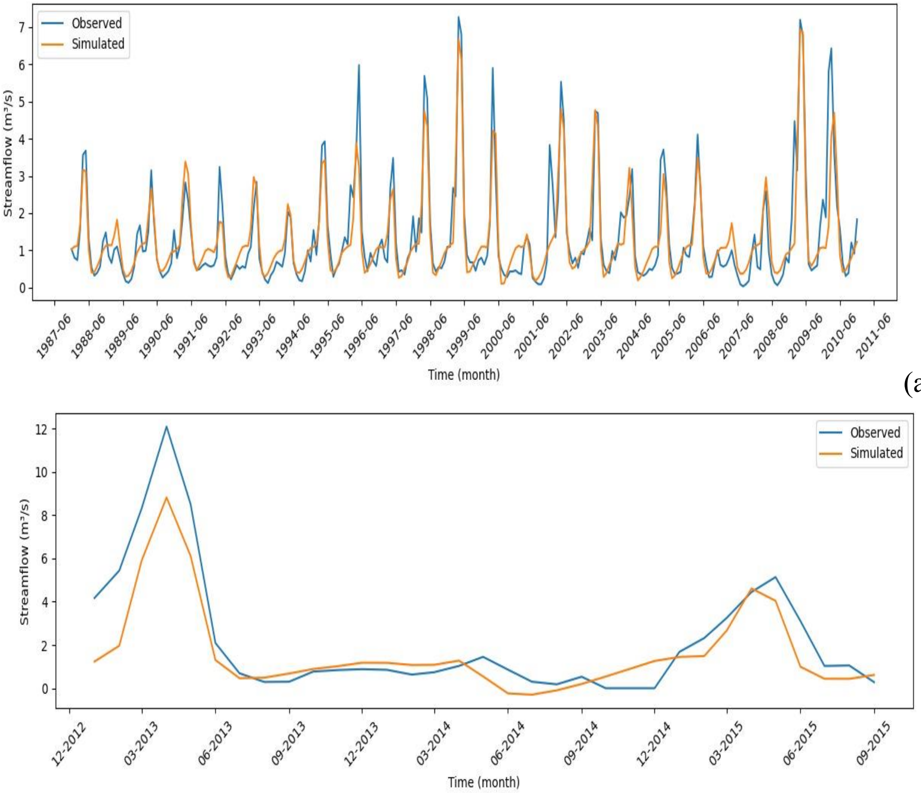


(a)

(b)

Figure 9: Time series plot of the LSTM model developed for Sym5: a) training set b) test set

The time series plots presented in Figure 9 show that the Sym5 model successfully captures both seasonal variations and flow changes during flood periods. The NSE values of 0.988 (training set) and 0.963 (testing set) further support this agreement.

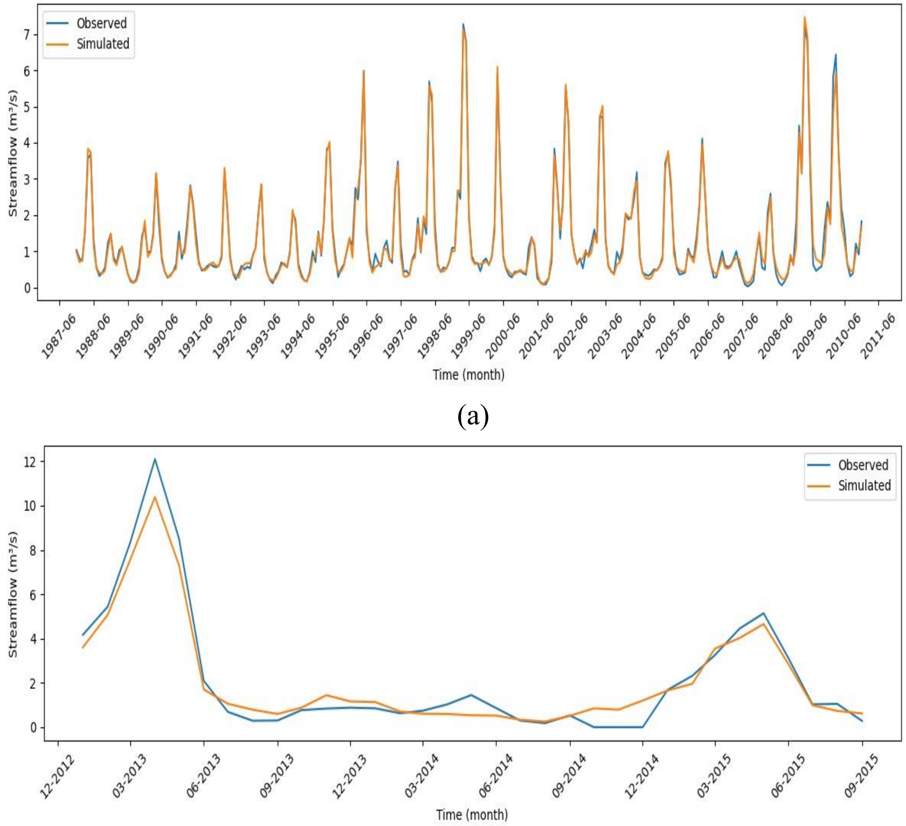


(a)

(b)

Figure 10: Time series plot of the LSTM model developed for Db4: a) training set b) test set

For the LSTM model developed using the Db4 wavelet, the test set NSE value was calculated as 0.781 and the KGE value as 0.604 (Table 2). Although the time series in Figure 10 generally capture the seasonal flows, it is observed that the model's response weakens at high flow values.

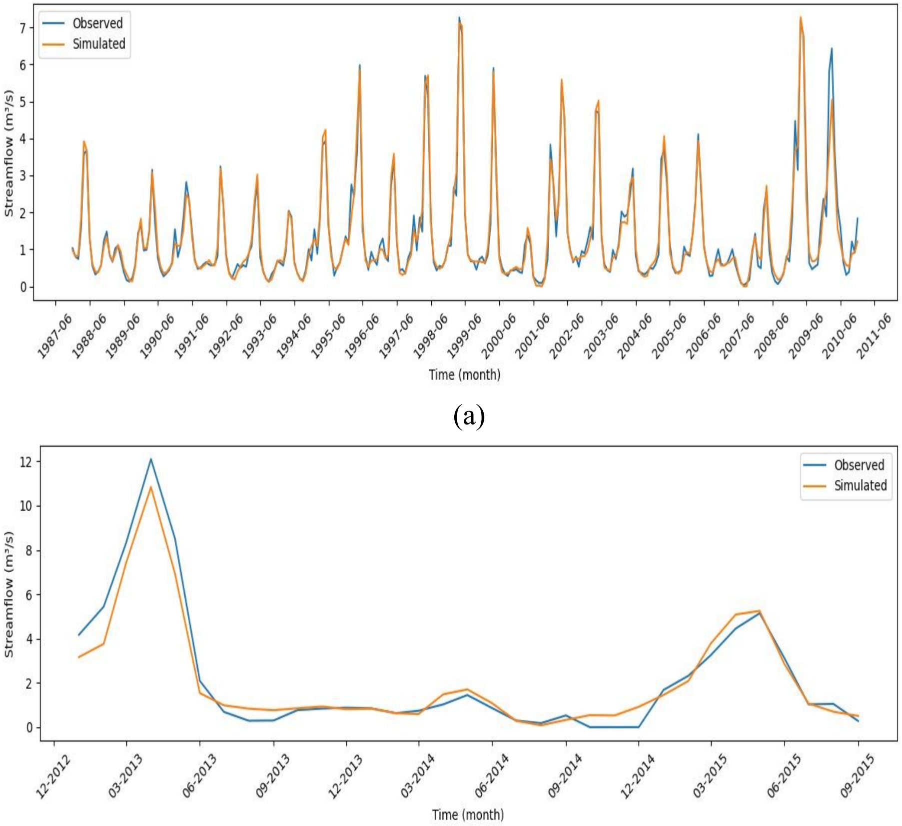


(a)

(b)

Figure 11: Time series plot of the LSTM model developed for Dmey: a) training set b) test set

The time series plots in Figure 11 clearly show that the model successfully captures flow peaks and performs very well during flood periods. This is supported by the NSE value of 0.986 for the training set and the NSE and KGE values of 0.956 and 0.842, respectively, for the test set, as shown in Table 2.

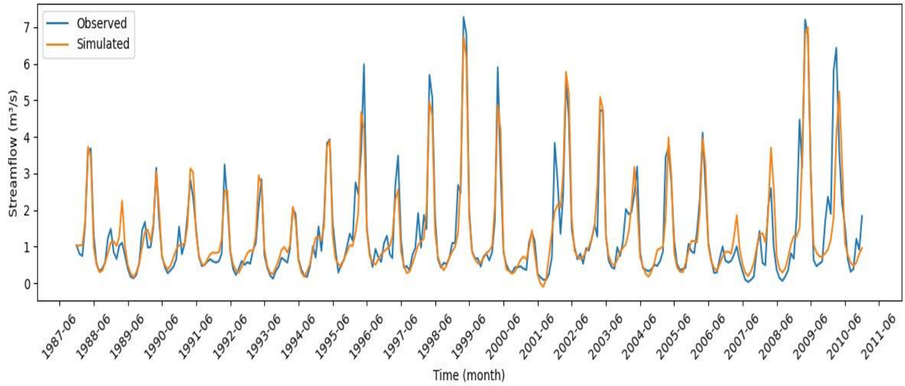


(a)

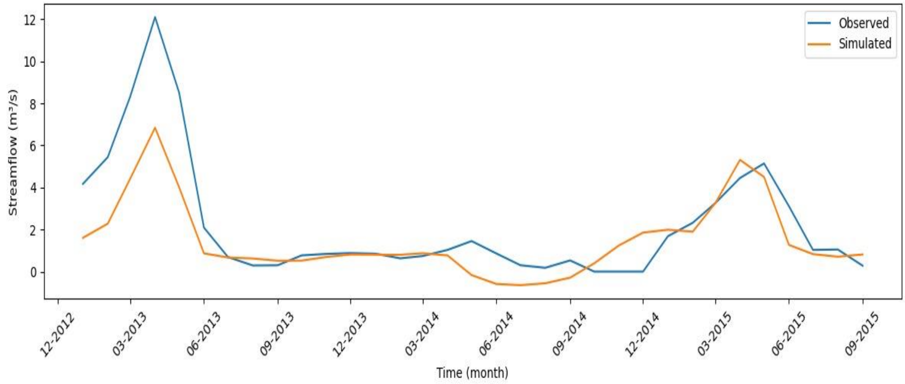
(b)

Figure 12: Time series plot of the LSTM model developed for Haar: a) training set b) test set

Figure 12 shows that the LSTM model developed with the Haar wavelet successfully captures the overall seasonal variability of streamflow during both the training and testing periods. The simulated hydrographs closely follow the observed series, especially under low and medium flow conditions, indicating a strong representation of seasonal and baseflow dynamics.



(a)



(b)

Figure 13: Time series plot of the LSTM model developed for Coif: a) training set b) test set

In the time series presented in Figure 13, the model's predictions fail to adequately represent the observed values, particularly during flood periods. To further investigate the seasonal behavior of the model, seasonal performance metrics were calculated for the Sym5 wavelet, and the results are summarized in Table 3.

Table 3: Seasonal performance metrics for the Sym5 wavelet

Season	RMSE	NSE	KGE
DJF	0.393	0.951	0.877
MAM	0.875	0.944	0.819
JJA	0.287	0.900	0.909
SON	0.365	0.350	0.116

To provide a more comprehensive assessment of the model's performance across different hydrological conditions, seasonal performance metrics were calculated for the Sym5 wavelet, and the results are presented in Table 3 along with a radar chart visualization in Figure 14. The radar chart offers a compact way to compare seasonal variations in RMSE, NSE, and KGE simultaneously, highlighting differences in model performance throughout the year.

When evaluating the seasonal performance metrics, it is observed that the model exhibits a generally balanced behavior across DJF, MAM, and JJA, while notable variability occurs in SON. During DJF, the model achieves very high accuracy (RMSE = 0.393, NSE = 0.951, KGE = 0.877), effectively capturing winter high-flow dynamics. Similarly, high performance is maintained during MAM (NSE = 0.944, KGE = 0.819), indicating that the model successfully represents seasonal flow transitions during spring. In JJA, representing dry summer months, the model performs well (NSE = 0.900, KGE = 0.909), capturing stable low flows.

In contrast, during SON, although RMSE remains moderate (0.365), NSE and KGE drop significantly (-0.350 and 0.116, respectively), reflecting the model's difficulty in capturing transitional flow regimes. The radar chart clearly illustrates this seasonal contrast, with a sharp contraction of the SON polygon compared to other seasons. Overall, the Sym5-based Wavelet–LSTM model demonstrates high skill in winter, spring, and summer, but reduced accuracy in autumn, emphasizing the value of seasonal performance evaluations in hydrological modeling.

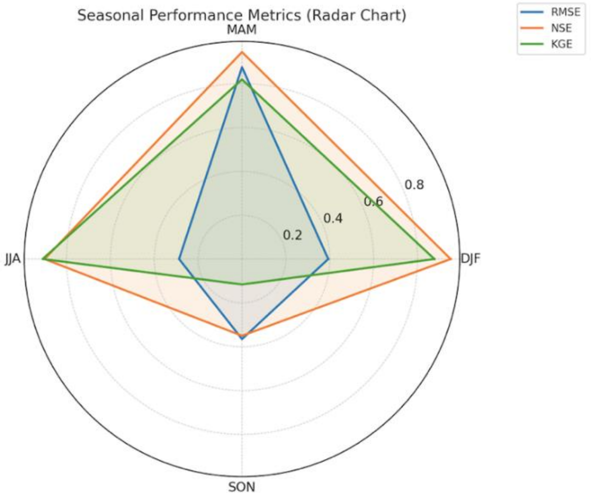


Figure 14: Seasonal performance metrics for the Sym5 wavelet represented as a radar chart

CONCLUSIONS

In this study, LSTM+ and Wavelet–LSTM models were developed for monthly streamflow prediction in the Gediz Basin, and the impact of different wavelet types on model performance was evaluated. The LSTM+ model, constructed using past precipitation, temperature, and streamflow data, produced accurate predictions under low and medium flow conditions but showed limited capability in capturing extreme values. Within the Wavelet–LSTM framework, the time series were decomposed into different frequency components using five wavelet families (Sym5, Db4, Dmey, Haar, and Coif3), and separate LSTM models were trained for each component. Performance comparisons revealed that the Sym5 wavelet achieved the highest accuracy on the test dataset (NSE = 0.963; RMSE = 0.545). Although models based on other wavelet types also outperformed the baseline LSTM+ model, Sym5 consistently yielded the best results. The seasonal performance analysis showed that the model performed best during winter (DJF), maintained high accuracy during spring (MAM) and summer (JJA), but exhibited a notable decrease in accuracy during autumn (SON). This decline reflects the challenges of modeling transitional hydrological processes in the basin during the autumn months. Overall, the results demonstrate that the Wavelet–LSTM approach outperforms the conventional LSTM model, both in terms of overall accuracy and its ability to capture seasonal dynamics. In particular, the choice of wavelet type plays a critical role in determining model performance. The superior results obtained with the Sym5 wavelet indicate that this wavelet type offers a robust alternative for monthly streamflow forecasting in basins with pronounced seasonal variability, such as the Gediz Basin.

REFERENCES

Slater, L.J., and Villarini, G. (2018). Enhancing the predictability of seasonal streamflow with a statistical-dynamical approach. *Geophysical Research Letters*, *45*(13), 6504-6513.

Troin, M., Arsenault, R., Wood, A.W., Brissette, F., and Martel, J.L. (2021). Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years.

Lin, Y., Wang, D., Wang, G., Qiu, J., Long, K., Du, Y., Xie, H., Wei, Z., Shangguan, W., and Dai, Y. (2021). A hybrid deep learning algorithm and its application to streamflow prediction. *Journal of Hydrology*, *601*, 126636.

Xie, M., Wang, B., Zhu, S., Ma, G., Yang, Z., Liu, B., and Jia, Y. (2022). Daily streamflow forecasting using hybrid long short-term memory model. In *Journal of Physics: Conference Series* (Vol. 2271, No. 1, p. 012019). IOP Publishing.

Wang, K., Band, S.S., Ameri, R., Biyari, M., Hai, T., Hsu, C.C., Hadjouni, M., Elmannai, H., Chau, K.W., and Mosavi, A. (2022). Performance improvement of machine learning models via wavelet theory in estimating monthly river streamflow. *Engineering Applications of Computational Fluid Mechanics*, *16*(1), 1833-1848.

Shoaib, M., Shamseldin, A.Y., and Melville, B.W. (2014). Comparative study of different wavelet based neural network models for rainfall–runoff modeling. *Journal of hydrology*, *515*, 47-58.

Barzegar, R., Fijani, E., Moghaddam, A. A., and Tziritis, E. (2017). Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models. *Science of the Total Environment*, *599*, 20-31.

Barzegar, R., Aalami, M.T., and Adamowski, J. (2021). Coupling a hybrid CNN-LSTM deep learning model with a boundary corrected maximal overlap discrete wavelet transform for multiscale lake water level forecasting. *Journal of Hydrology*, *598*, 126196.

Wei, M., and You, X.Y. (2022). Monthly rainfall forecasting by a hybrid neural network of discrete wavelet transformation and deep learning. *Water Resources Management*, *36*(11), 4003-4018.

Muhammad, A.U., Li, X., and Feng, J. (2019). Using LSTM GRU and hybrid models for streamflow forecasting. In *International Conference on Machine Learning and Intelligent Communications* (pp. 510-524). Cham: Springer International Publishing.

Kumanlioglu, A.A. (2020). Characterizing meteorological and hydrological droughts: A case study of the Gediz River Basin, Turkey. *Meteorological Applications*, *27*(1), e1857.

Taylan, E.D. (2024). Enhanced hydrological drought prediction in the Gediz Basin: integrating meteorological drought via hybrid wavelet-machine learning-random oversampling models using. Journal of Water and Climate Change, 15(9), 4790-4816.

GBDMP (2019) Gediz Basin Drought Management Plan, Republic of Turkey Ministry of Agriculture and Forestry, General Directorate of Water Management. Republic of Türkiye, Ankara, p. 484.

Krichen, M., and Mihoub, A. (2025). Long Short-Term Memory Networks: A Comprehensive Survey. AI, 6(9), 215.

Jothi, L. S. (2025). Time Series Analysis Using LSTM Networks and Its Application to Financial Forecasting. In Artificial Intelligence for Financial Risk Management and Analysis (pp. 19-58). IGI Global Scientific Publishing.

Guo, J., Peng, Y., Zhou, Q., and Lv, Q. (2019). Enhanced lstm model for short-term load forecasting in smart grids. In International Conference on Cloud Computing (pp. 650-662). Cham: Springer International Publishing.

Guo, Z., Zhang, Q. Q., Li, N., Zhai, Y. Q., Teng, W. T., Liu, S. S., & Ying, G. G. (2023). Runoff time series prediction based on hybrid models of two-stage signal decomposition methods and LSTM for the Pearl River in China. Hydrology Research, 54(12), 1505-1521.

Eniola, V., Adeyemi, K., Adamu, M., Fasipe, O., Aikhuele, J., Zarmai, M., and Uthman, M. (2025). Daily Streamflow Forecasting Using an Enhanced LSTM Neural Network Model. In Modelling the Energy Transition: Cultures, Visions, Narratives (pp. 135-171). Cham: Springer Nature Switzerland.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Ohno, K., and Kumagai, A. (2021). Recurrent neural networks for learning long-term temporal dependencies with reanalysis of time scale representation. In 2021 IEEE International Conference on Big Knowledge (ICBK) (pp. 182-189). IEEE.

Daubechies, I. (2002). The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, *36*(5), 961-1005.

Torrence, C., and Compo, G.P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, *79*(1), 61-78.

Ogbonna, C.J., Nweke, C.J., Nwogu, E.C., and Iwueze, I.S. (2016). Wavelet Transform as an Alternative to Power Transformation in Time Series Analysis. *Bulletin of Mathematical Sciences and Applications*, *17*, 57-74.

Mallat, S.G. (2002). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, *11*(7), 674-693.

Percival, D.B., and Walden, A.T. (2000). *Wavelet methods for time series analysis* (Vol. 4). Cambridge university press.

Hong-fa, W. (2012). Clustering of hydrological time series based on discrete wavelet transform. *Physics procedia*, *25*, 1966-1972.

Bunrit, S., Kerdprasop, N., and Kerdprasop, K. (2018). Multiresolution analysis based on wavelet transform for commodity prices time series forecasting. *International Journal of Machine Learning and Computing*, *8*(2), 175-180.

Addison, P.S. (2017). *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press.

Kaddar, B., and Fizazi, H. (2017). Spatiotemporal analysis for NDVI time series using local binary pattern and Daubechies wavelet transform. *Int. Rev. Aerosp. Eng.(IREASE)*, *10*(2), 96-104.

Partal, T., and Kişi, Ö. (2007). Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. *Journal of Hydrology*, *342*(1-2), 199-212.

Nourani, V., Kisi, Ö., and Komasi, M. (2011). Two hybrid artificial intelligence approaches for modeling rainfall–runoff process. *Journal of Hydrology*, *402*(1-2), 41-59.

Deo, R.C., and Şahin, M. (2015). Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia. *Atmospheric research*, *161*, 65-81.