engineering proceedings

MDPI

Proceeding Paper

Short-Term Forecasting of Non-Stationary Time Series †

Amir Aieb [1,2,*], Antonio Liotta [1], Alexander Jacob [2] and Muhammad Azfar Yaqub [1]

[1] Faculty of Engineering, Free University of Bozen-Bolzano, 39100 Bolzano, Italy; antonio.liotta@unibz.it (A.L.); muhammadazfar.yaqub@unibz.it (M.A.Y.)
[2] Institute for Earth Observation, Eurac Research, 39100 Bolzano, Italy; alexander.jacob@eurac.edu
* Correspondence: amir.aieb@stud-inf.unibz.it
† Presented at the 10th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 15–17 July 2024.

Abstract: Forecasting climate events is crucial for mitigating and managing risks related to climate change; however, the problem of non-stationarity in time series (NTS) arises, making it difficult to capture and model the underlying trends. This task requires a complex procedure to address the challenge of creating a strong model that can effectively handle the non-uniform variability in various climate datasets. In this work, we use a daily standardized precipitation index dataset as an example of NTS, whereby the heterogeneous variability of daily precipitation poses complexities for traditional machine-learning models in predicting future events. To address these challenges, we introduce a novel approach, aiming to adjust the non-uniform distribution and simplify the detection time lags using autocorrelation. Our study employs a range of statistical techniques, including sampling-based seasonality, mathematical transformation, and normalization, to preprocess the data to increase the time lag window. Through the exploration of linear and sinusoidal transformation, we aim to assess their impact on increasing the accuracy of forecasting models. A strong performance is effectively observed by using the proposed approach to capture more than one year of time delay across all the seasonal subsets. Furthermore, improved model accuracy is observed, notably with K-Nearest Neighbors (KNN) and Random Forest (RF). This study underscores RF's consistently strong performance across all the transformations, while KNN only demonstrates optimal results when the data have been linearized.

Keywords: short-term forecasting; data transformation; non-stationary time series; machine learning; time lag

check for updates

Citation: Aieb, A.; Liotta, A.; Jacob, A.; Yaqub, M.A. Short-Term Forecasting of Non-Stationary Time Series. *Eng. Proc.* **2024**, *68*, 34. https://doi.org/10.3390/engproc2024068034

Academic Editors: Olga Valenzuela, Fernando Rojas, Luis Javier Herrera, Hector Pomares and Ignacio Rojas

Published: 10 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Climate change is one of the major challenges of the 20th century, characterized by rising temperatures, variations in precipitation, and the intensification of extreme weather phenomena, which alter global and regional precipitation patterns. In some regions, this may result in more frequent and prolonged dry spells [1]. Droughts can lead to long-term environmental degradation, including soil erosion, loss of biodiversity, and depletion of water resources. Drought estimating and forecasting involve a combination of monitoring, data analysis, and robust predictive modeling techniques. Additionally, they involve a thorough analysis of detecting different patterns, periodicity, and the interaction of multiple environmental factors based on the historical observations [2,3].

Two main methodologies stand out to forecast climate parameters, i.e., physical, and data-driven models [4–6]. These represent distinct approaches for understanding and predicting the complexities of the Earth's climate system. Physical models are often employed for medium- and long-term forecasts, as they incorporate complex parameters such as oceanic circulation and large-scale interactions between the atmosphere and the oceans. Although these models are demanding in terms of computational resources, they offer an in-depth understanding of meteorological and climatic processes [7]. Conversely,

data-driven models are generally more suitable for short-term forecasts due to their ability to quickly capture trends and patterns from real-time data. Their algorithmic flexibility also enables them to adapt to rapid changes in weather patterns. However, the long-term reliability of these models may be affected by sudden and unforeseen variations in input data, posing challenges for their use in longer-term climate-forecasting contexts [8,9]. The first models developed to address this issue include statistical extrapolation techniques, which are extensively employed in various fields, such as meteorology, urban science, and energy [10,11]. Among these models, there are Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Fourier Forecast [12–14]. The effectiveness of ARIMA and its extension, SARIMA, lies in the fact that it adeptly captures seasonal fluctuations. In contrast, the Fourier model breaks down data into frequencies, uncovering periodic patterns that provide precise predictions and insights into evolving temporal forecasts. However, their limitation in modeling nonlinear data with non-stationary distributions has driven the advancement of Artificial Neural Networks (ANNs), which now stand as the primary research focus in developing various intelligent models [15].

Time series analysis is a crucial research domain aimed at tackling forecasting challenges across diverse applications. This extends to the examination of non-stationary data, which frequently showcases non-uniform trends, seasonal cycles, and other temporal structures undergoing irregular changes. This complexity poses challenges in capturing and modeling temporal data patterns, as traditional techniques suffer from having to effectively generalize models for nonlinear data distributions [16]. Therefore, the development of robust models for non-stationary data forecasting often requires the use of advanced machine- and deep-learning models, such as Recurrent Neural Networks (RNNs), Long Short-Term Memory models (LSTMs), or Long-Term Gaussian Processes (GP-LSTMs). These approaches enable us to capture the complex temporal dependencies and dynamically adjust the model parameters based on data evolution and preprocessing analysis, which improves models' accuracy in forecasting multiscale data series [17–19]. However, enhancing the accuracy of forecasting models also involves a foundational step known as data preprocessing, which demands considerable time and attention within any data analytics workflow. Data preprocessing ensures the quality of the data utilized as input for machine-learning techniques, thereby minimizing the risk of generating inaccurate or flawed results. This typically includes data cleaning to rectify inconsistencies, errors, and features, sample selection to pinpoint relevant data subsets, outlier removal to enhance data reliability, normalization to standardize data distributions, and data transformation to discern tendencies, seasonality, and periodicity. Data preprocessing makes it easier to capture the period of the similar motifs in time series and enhance the model fitting [20]. According to the literature, various transformation methods have been developed and applied to improve the non-stationary pattern of data time series and make them more homogeneous. These methods are mainly classified into two categories, which are data mapping and decomposition. A set of parametric and nonparametric models has been proposed for mapping data based on trend adjustment via detrending methods (linear and polynomial detrending, seasonal adjustment. . .etc.), or by ignoring tendency using mathematical transformations (linearization, nonlinear smoothing, and fractional differences). On the other side, the second category of transformed models focus on decomposing data into trends and seasonality using the time or time–frequency domains, such as seasonal and trend decomposition using locally estimated scatterplot smoothing (LOESS), empiric mode decomposition, and wavelet transformation [20,21].

Our study presents a new approach to address the challenge of forecasting non-stationary time series data on a small scale. To exemplify our method, the daily standardized precipitation index analyzed for the Adige basin (in the Alps, Italy) is chosen as a case study, where the dataset exhibits a non-stationary distribution due to the non-uniform variability of daily precipitation. This complexity makes it challenging to adapt machine-learning parameters for forecasting future events. Our approach aims to adjust the

non-stationary distribution and simplify the detection of time delays using a basic linear autocorrelation method. We employ a set of statistical techniques in the preprocessing phase, including sampling-based seasonality, mathematical transformation, and normalization, to standardize the values and facilitate comparative performance analysis among the different models. In this context, we apply two transformation methods to evaluate their impact on the length of the time lag window and the accuracy of the machine-learning forecasting models. This paper is structured into three sections, starting with this introduction, which provides an overview of the state-of-the-art techniques and models used for time series forecasting. In the subsequent section, we introduce the methods, datasets, and the proposed approach. Finally, the results, conclusions and future works are presented in the last section.

2. Material and Methods

In the literature, various techniques and approaches have emerged for examining the stochastic nature of data time series from historical observations and adapting diverse regression models for predicting future events. This section introduces the transformation methods applied to a non-stationary dataset, discusses machine-learning approaches based on using time lags, and provides a statistical overview of the Standardized Precipitation Index (SPI) estimated dataset.

2.1. Data Transformation

Having non-stationary distributions in data time series presents significant hurdles to ensuring the accuracy of the model's prediction and forecasting. To this end, a multitude of methodologies have been proposed to address this issue, including data transformation. One good method suggests transforming data to be within the same range as the output of the activation function that is used in the model training. In this study, we outline the transformation techniques used to tackle the challenge of non-stationary data distribution [20]. Our approach involves selecting a mathematical transformation added to normalization (ranging between 0.1 and 1). The choice of the minimum value aims to linearize the data through the Log function (see Figure 1). Additionally, another sinusoidal transformation was incorporated using the Sin function, which enhances the visibility of the periodicity compared to the Linear transformation. Following model processing, it is essential to revert the data to its original scale. Therefore, invertible methods like the Exp function are employed for this purpose.

Original data -> Transformed data

```
def Linear_N(data):
    X = scaler.fit_transform(data)
    Y = np.log (X)
    Data_N = scaler.fit_transform(Y)
    return Data_N

def Sinusoidal_N(data):
    X = np.sin (data)
    Data_S= scaler.fit_transform(X)
    return Data_S
```

Transformed data -> Original data

```
def Inv_Linear_N (data):
    X = scaler.inverse_transform(data)
    Y = np.exp (X)
    Inv_data_N = scaler.fit_transform(Y)
    return Inv_data_N

def Inv_Sinusoidal_N (data):
    X= scaler.inverse_transform(data)
    Inv_data_S = np.arcsin (X)
    return Inv_data_S
```

Figure 1. Algorithm for transforming non-stationary time series data.

2.2. Machine-Leaning Model Forecast Based on the Time Lag

Using the time lag (or time delay) of machine-learning forecasting involves leveraging historical data with lagged features to forecast future values. The time lag refers to the

manipulation of temporal data by shifting the timestamps (T) of historical observations to create lagged features. These lagged features represent past observations of the target variable or other relevant variables at different time intervals leading up to the prediction point [22]. By incorporating the time lag into machine-learning forecasting models, algorithms can capture temporal dependencies and patterns in the data, enabling the prediction of future values based on historical trends and relationships. This approach leverages the principle that past observations can provide valuable information for predicting future outcomes [22,23]. In this work, the maximum time lag should be shorter than half of the timespan of the target dataset. However, determining the maximum time window is a critical step in establishing robust parameters for forecasting models. The process of splitting data between input (X_train) and regressors (Y_train) for model training is conducted as follows:

$$X_train = data\ [:T] \quad (1)$$

$$Y_train = data\ [T:] \quad (2)$$

where T is the time lag, varying between 1 and len(data)/2.

Two supervised machine-learning methods are chosen in this study based on their ability to capture subsets of each X_train data that exhibited a strong fit with Y_train. K-Nearest Neighbors (KNN) is a simple, yet effective algorithm used for classification and regression tasks. It operates on the principle that similar data points tend to have similar labels or outcomes. KNN predicts the label of a new data point by considering the majority class among its K Nearest Neighbors based on a distance metric [24]. As the second forecasting method, Random Forest was selected in this analysis. It represents an ensemble-learning technique that integrates multiple decision trees in both the training and prediction phases. When making predictions, the outputs of individual trees are combined to generate the final prediction [25].

2.3. Metrics of Performance

Various techniques are used in this paper to analyze the performance and the bias that machine learning may exhibit when incorporating time lag and data transformations for forecasting future events. These parameters include the coefficient of determination (R^2), adjusted coefficient of determination (R^2_{Adj}), mean square error (MSE), mean absolute error (MAE), and *t*-test [26–29].

2.4. Data Collection and Analysis

The data used in this study pertain to daily precipitation obtained from a historical record of the Adige basin for the period between 2010 and 2021. This region is located in the northeastern part of Italy, between 45.8 and 46.6 degrees north latitude, and 10.8 to 12.6 degrees east longitude. It is characterized by a spatially non-uniform distribution due to its mountainous morphology [30]. In this study, our focus is on the temporal scale. Figure 2A provides the results of a statistical analysis applied on the daily precipitation of one site in the Adige catchment during the specified period. A non-stationary distribution is observed in the plot of the monthly rolling mean and median, where the values exhibit the lack of a consistent relationship between the two parameters. This information is also evident using monthly standard deviation parameters, where the values show significant variability, especially in 2014, 2020, and 2021. The use of the third-order Pearson method and the gamma distribution function is common for estimating the SPI. This approach allows consideration of the statical characteristics of the precipitation distribution on SPI processing. According to Figure 2B, a very good fit is observed by using both stochastic models, as reflected by *p*-values equal to 0.82 and 0.87, respectively.

In the end, the final SPI result is achieved by combining both series. Figure 2C,D show the SPI data subsets, as divided into training and testing. The data spanning from 2010 to 2017 are proposed to train the ML model, while the period between 2018 and 2021 is reserved for model validation. The choice of data for both training and testing emphasizes

the importance of capturing temporal dynamics and trends. This ensures that the training phase covers a wide range of temporal patterns and variations relevant to the problem being addressed. This is substantiated by the application of the Pettitt test to both sample series, which highlights the influence of non-uniform rainfall regimes on the distribution of SPI data. The findings indicate that both series exhibit non-homogeneity and support the existence of at least two mean values (mu1 and mu2).

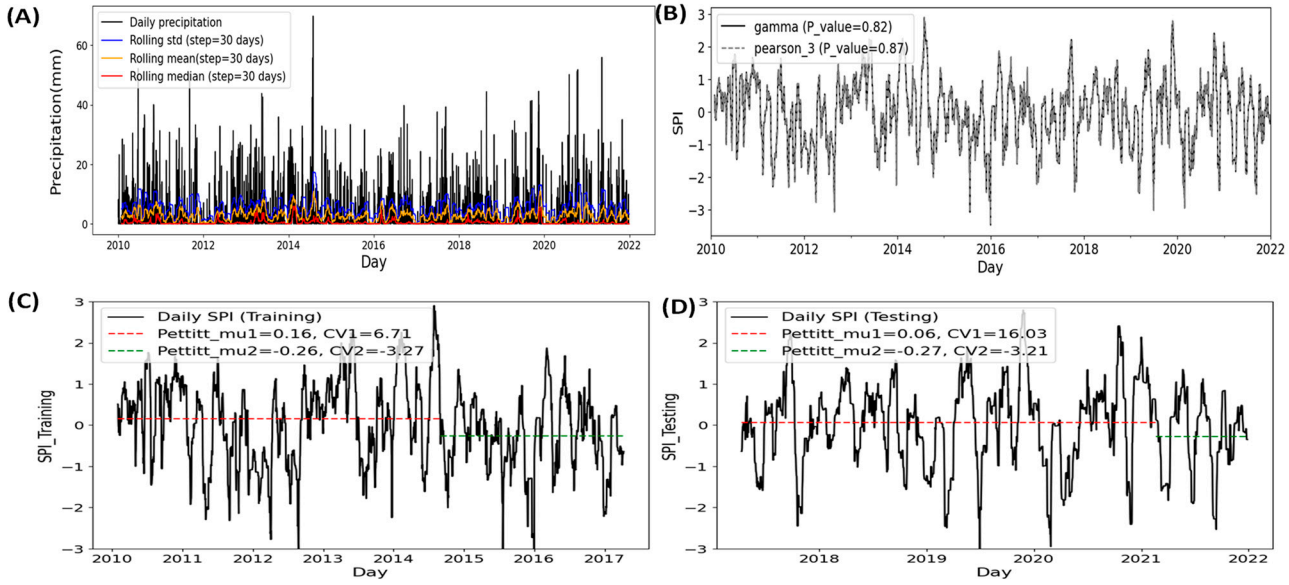


Figure 2. Statistical description of daily precipitation (**A**) and standardized precipitation index obtained from the gamma and Pearson_3 models (**B**), followed by the test of homogeneity for both the training and testing datasets (**C**,**D**).

3. Proposed Method

The complexity of non-stationary data in time series analysis becomes more apparent at finer scales, such as daily and infra-daily. This non-uniform variability poses additional challenges for modeling and forecasting, as the variability is due to the irregular fluctuation of the means and variances across various time intervals. This temporal heterogeneity complicates the adaptation of the ML model's parameters, especially for applications requiring continuous time forecast. In this section, we present a method to reduce the non-uniform trend in the data by employing sampling techniques, data transformation, and auto-correlation to identify time lags where the data series displays periodicity and similarity. From a statistical perspective, this is where the data show a strong correlation. The flowchart illustrated in Figure 3 outlines the various stages of the analysis, starting from the preprocessing step and moving through to the training and validation of the ML model. In this study, we focus on forecasting drought using SPI daily data as a case study; nevertheless, the model is adaptable to any time series data exhibiting a non-stationary distribution.

The model initiates by organizing and ranging the data according to seasonal sampling, followed by detrending using linearization (log) and sinusoidal transformation (Sin) methods. Subsequently, data normalization is applied to facilitate statistical analysis, such as modeling, performance analysis, and comparison between different results.

The dedicated model is primarily based on the use of time lags to predict the future daily events within time windows. These time frames are detected through autocorrelation analysis by dividing the data series into input and regressor segments using a daily shift (T). We chose simpler machine-learning methods over deep learning due to the stationary distribution achieved through preprocessing. Furthermore, this analysis is univariate and utilizes features as a sample vector.

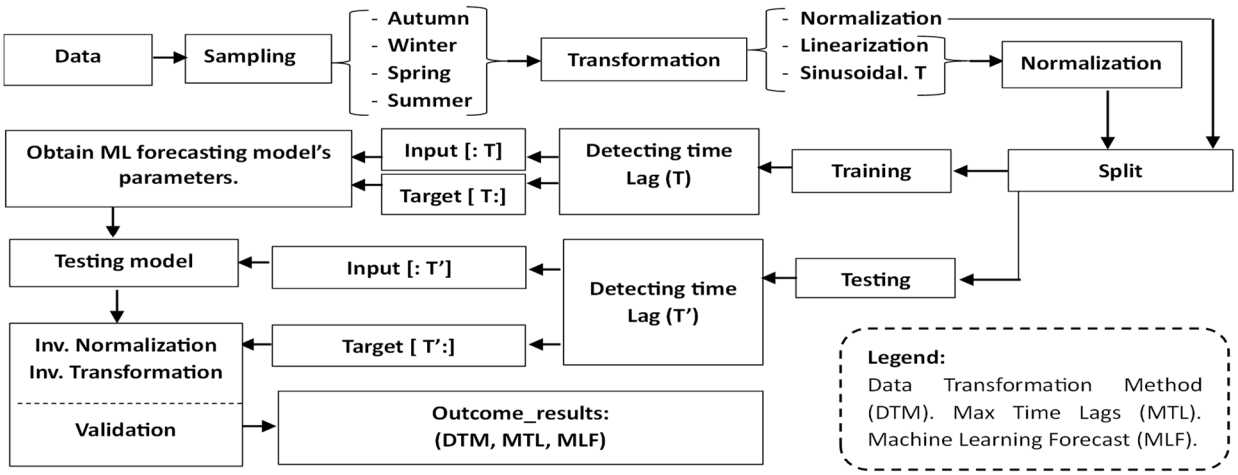


Figure 3. The flowchart summarizes the workflow of the proposed approach for time series data forecasting.

4. Results and Discussion

In this section, we outline the outcomes of the data preprocessing, which focus on the data transformation and the determination of the time lag using autocorrelation for both the training and testing datasets across various transformations. Following that, we present the results obtained from using K-Nearest Neighbors (KNN) and Random Forest (RF) for predicting future SPI events based on the maximum time lag. Table 1 provides a descriptive statistical analysis of the training and testing datasets used during the processing step. Various tests are conducted on the data series after transformation and normalization. It offers a comparative examination of each data sample before and after adjustment. The outcomes reveal that the mean and median values closely approximate the mid-range of the normalized data (between 0 and 1), particularly for linearized and sinusoidal-transformed data. Following the preprocessing phase, the SPI values achieved stationarity, where the variability of the linearized dataset decreases and shows an enhancement of the data distribution (uniform) compared to the other dataset. In Table 1, we have highlighted the best transformation results for each seasonal sample using the green color. The results selection is based on taking the minimum CV value and the degree of similarity between the mean and median values. From the results, the best transformation applied to the autumn data for both training and testing was observed with the sinusoidal transformation. The same level of performance was observed because the variability of the original data was consistent for both the training and testing datasets. On the other hand, the statistics applied to the winter and spring data showed good model performance with the sinusoidal transformation for the data used in training the ML models. However, for the testing data, the results indicated a good performance of the linear transformation. Similar results were also observed with the summer datasets due to the uniform distribution. In general, the best stationarity of the time series is achieved when using the sinusoidal and linear trans-formation methods. However, the choice of the best method depends on the nonlinear distribution of the original data.

Autocorrelation tests, as a fundamental concept in time series analysis, play a crucial role in identifying temporal dependencies and patterns within sequential data by measuring the correlation between a time series and its lagged values at various time intervals. Figure 4,sets the stage for discussing the importance of autocorrelation tests in analyzing data time series and outlines their implications for understanding temporal dynamics within the transformed SPI datasets. It also underscores the broader importance of identifying the best transformation method to make data more stationary across different seasons. In this analysis, we shifted the data based on the time windows by generating time lags that range from one day to half the size of the observation period. It is important to note that a small sample size can impact the quality of the results provided in the modeling

step. Figure 4, illustrates scenarios where the time lag exhibits a robust correlation between historical and future observations.

Table 1. Statistical comparison of the stationarity analysis between the original and the transformed daily SPI during the four seasons.

Method	Samples	Training Data			Testing Data		
		Mean	**Median**	**CV**	**Mean**	**Median**	**CV**
Original_N	Autumn_SPI	0.74	0.61	0.31	0.76	0.67	0.29
	Winter_SPI	0.61	0.94	0.42	0.86	0.62	0.47
	Spring_SPI	0.54	0.41	0.40	0.58	0.44	0.26
	Summer_SPI	0.50	0.54	0.22	0.56	0.50	0.14
Linearization_N	Autumn_SPI	0.61	0.46	0.27	0.76	0.72	0.25
	Winter_SPI	0.65	0.82	0.39	**0.68**	**0.70**	**0.18**
	Spring_SPI	0.56	0.49	0.18	**0.52**	**0.49**	**0.25**
	Summer_SPI	**0.55 ***	**0.52**	**0.17**	0.46	0.59	0.17
Sinusoidal_N	Autumn_SPI	**0.65**	**0.68**	**0.22**	**0.56**	**0.54**	**0.23**
	Winter_SPI	**0.67**	**0.73**	**0.19**	0.58	0.67	0.41
	Spring_SPI	**0.52**	**0.47**	**0.16**	0.50	0.55	0.28
	Summer_SPI	0.56	0.50	0.29	**0.50**	**0.54**	**0.11**

Standardized precipitation index (SPI), normalization (N), coefficient of variation (CV). * Best results.

This analysis was conducted across all the seasonal samples of daily SPI data following linearization and sinusoidal transformation. The plots show the common time lag, providing insights into the performance of the transformation methods. Furthermore, the outcomes identify the maximum lag appropriate for forecasting data using historical observations. In this analysis, we used data normalization to facilitate comparison across different plots. The results from the training data show that data transformation contributes to stabilizing the distribution across the year. It is observed that the maximum time lag occurs by shifting the daily data using these samples compared to the original one. During the autumn season, linearized data provide the maximum time lag (285 days) compared to all the other data transformation techniques. Conversely, during the winter and spring periods, sinusoidal transformation demonstrates the most effective data shift compared to both the linearized and original scales, with observed lags of 330 and 341 days, respectively.

However, during the summer season, which is characterized by a weak rainfall regime, the original data show a maximum time lag just after application of a simple normalization. The stationary SPI datasets derived through mathematical transformations are used in this section for training and testing the accuracy and the bias of the KNN and RF models. Moreover, 70% of the data were selected for building and training both models used to forecast the SPI based on the historical and future regressor data vectors provided via autocorrelation analysis. The remaining 30% of the data were reserved for assessing the performance of the model's parameters. In this part, we experiment with all four seasonal subsets obtained from the sampling, which helps provide the annual variability of the SPI. Regarding the used time lag, we opt for the maximum value of time lag to visualize the residual trends over all the range of future events. In addition, the forecasting data help to understand the limits of the model application for each machine-learning method. Multiple statistical tests, such as R^2, R^2 Adjusted, MSE, MAE, and *t*-tests, were employed to control and compare the accuracy, bias, and variability between the forecasting and ground truth data.

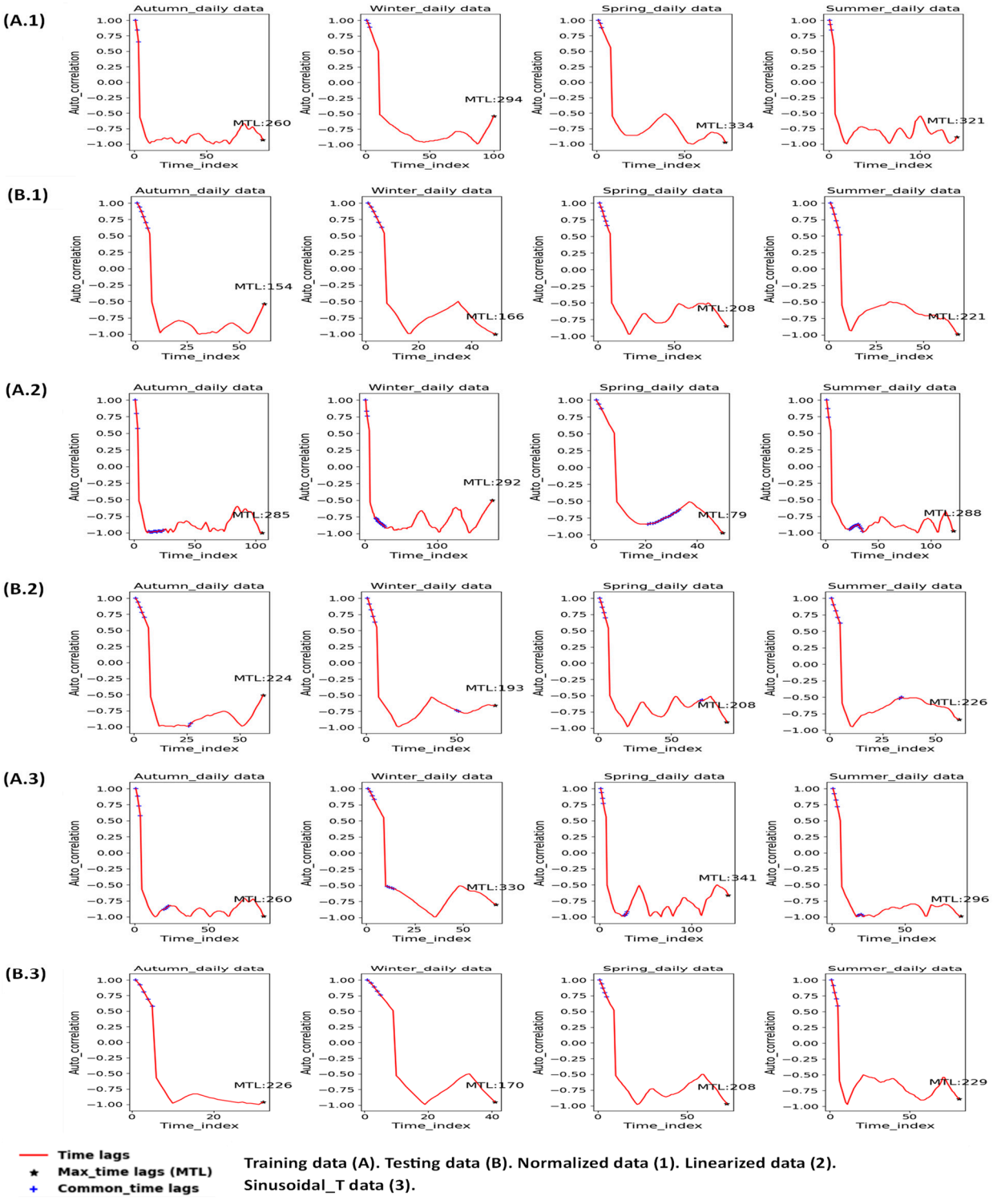


Figure 4. Autocorrelation analysis illustrating the time lag detection using various data transformation methods.

Figures 5 and 6, show the graphical results of training and testing both KNN (K = 3) and RF, as applied on transformed data during the four seasonal periods. The choice of a K equal to 3 is based on the optimization analysis illustrated in Figure S1, where we trained and tested the model for K values ranging from 1 to 50, incorporating all the time delays provided by the cross-correlation. The optimal K is determined by its ability to maintain a strong correlation across all cases. In the testing step, we present the results by

using the original scale to assess the impact of data transformation on the accuracy of the forecasting model during the training phase. A very good performance is observed with the RF method compared to KNN, as demonstrated by an R^2 consistently exceeding 0.86. This performance is also evident in the test phase when using the same model parameters, which were obtained during training. Conversely, KNN exhibits some bias when utilizing normalized data, particularly during the spring and summer seasons (R^2 equals 0.52 and 0.54, respectively). When predicting future events through linear transformation, both models provide a satisfactory fit with the ground truth data. However, KNN demonstrates the highest accuracy, with R^2 values ranging between 0.87 and 0.92.

Table 2 shows the statistical analysis regarding testing both the KNN and RF models. It offers different insights to determine if the observed bias significantly produces a change in forecasting data distribution by applying the *t*-test hypothesis to compare these results with ground truth data. We highlight the best results using the green color, which are determined based on the maximum time forecast and the highest accuracy of the ML models across different metrics.

Table 2. Performance analysis of machine-learning forecasting models applied to each sample of data after transformation.

Method	Sample	Time_lag	KNN			RF		
		Day	R^2_Adj	MAE	*t*-test (P)	R^2_Adj	MAE	*t*_test (P)
Original_N	Autumn_SPI	154	0.65	0.47	0.79	0.87	0.21	0.71
	Winter_SPI	166	0.66	0.35	0.87	0.86	0.18	0.91
	Spring_SPI	208	**0.53 ***	**0.51**	**0.04**	0.81	0.25	0.63
	Summer_SPI	221	**0.51**	**0.33**	**0.01**	0.81	0.15	0.76
Linearization_N	Autumn_SPI	224	0.86	0.11	0.88	0.71	0.38	0.53
	Winter_SPI	**193 ***	**0.86**	**0.15**	**0.89**	0.68	0.28	0.28
	Spring_SPI	**208**	**0.81**	**0.25**	**0.74**	0.53	0.42	0.11
	Summer_SPI	224	0.78	0.17	0.61	0.62	0.30	0.26
Sinusoidal_N	Autumn_SPI	**226**	0.61	0.30	0.60	**0.80**	**0.18**	**0.81**
	Winter_SPI	170	0.63	0.28	0.53	0.81	0.17	0.24
	Spring_SPI	208	**0.56**	**0.39**	**0.04**	0.77	0.27	0.11
	Summer_SPI	**229**	**0.53**	**0.26**	**0.02**	**0.78**	**0.16**	**0.93**

Standardized precipitation index (SPI), normalization (N), *p* value (P), k-nearest neighbors (KNN), random forest (RF), adjusted coefficient of determination (R^2_Adj), mean absolute error (MAE), *p* value (P). * Worst results. * Best results.

In addition, the red color is used to highlight the significant biases provided by both ML models based on the *t*-test scores (see Table 2). Generally, the best choice of forecasting method also has a strong correlation with the stationarity statistical analysis (refer to Table 1 for the case of the testing data). Since the CV provides a good result, the autocorrelation helps in detecting the maximum time lag. From the results, the linear transformation method enhances the performance of the KNN method.

Moreover, the *p*-values reveal notable shifts in the mean values of the outcome data obtained using this method. Contrariwise, in the instances of normalized data and data derived from sinusoidal transformation, significant bias is evident when employing the KNN model to forecast the daily SPI for spring and summer, as indicated by *p*-values below 0.05. RF, as an unsupervised method, is adapted to learn the nonlinear behaviors of time series. It can only show good results by using simple normalization. However, the best selection is performed by using sinusoidal transformation to train the RF model for both the autumn and summer periods based on the maximum time forecasting, where an increase in the time window is shown from 154 to 226 and from 221 to 229, respectively.

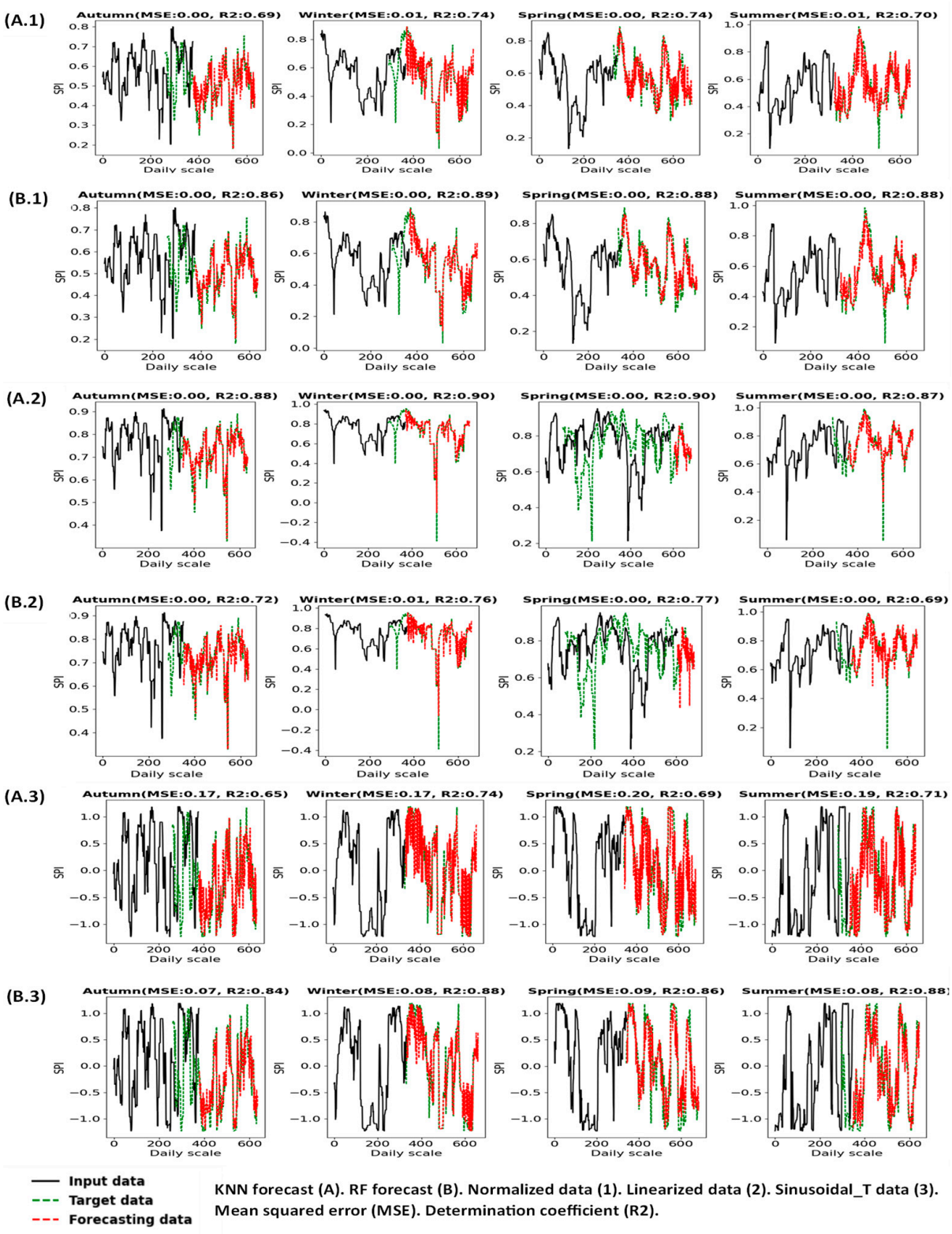


Figure 5. Results of training K-Nearest Neighbors (KNN) and Random Forest (RF) models to forecast the Standardized Precipitation Index (SPI), under different data transformation methods.

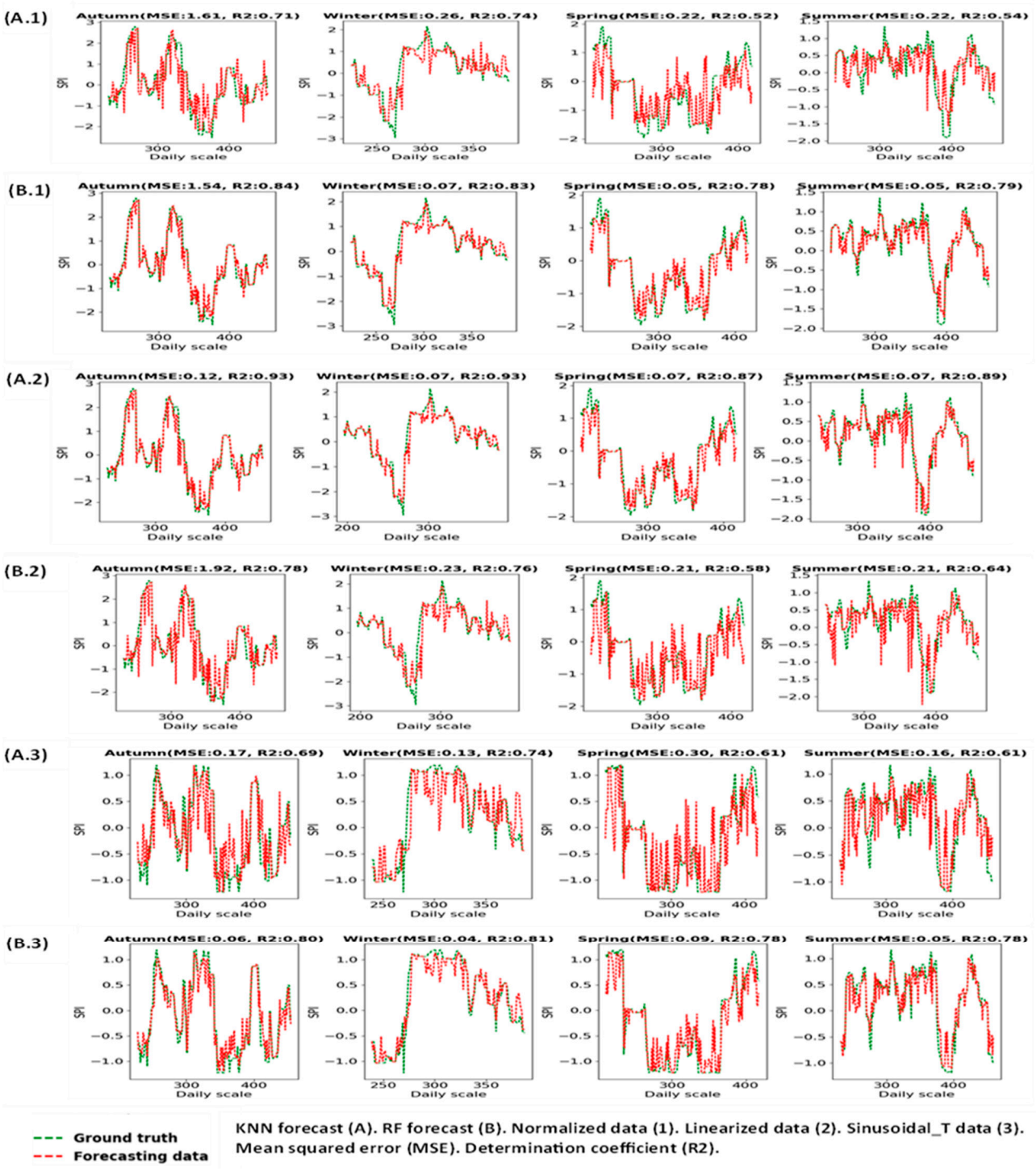


Figure 6. Results of testing the K-Nearest Neighbors (KNN) and Random Forest (RF) models for the standardized precipitation index (SPI) forecasting on the original scale.

Furthermore, Figure 7 illustrates the accuracy assessment of each proposed model using cross-validation for a 10-day time step, where the R^2 and MAE results obtained for each season using data transformation are presented. The scores highlight the robustness of RF throughout all the periods, particularly when addressing nonlinear data. The combined application of sinusoidal transformation with both ML forecasting models demonstrates consistent and accurate outcomes, with R^2 values surpassing 0.5 in each subcase. Conversely, in scenarios involving normalization, RF displays notable performance compared to KNN, as evidenced by the MAE curves (refer to Figure 7).

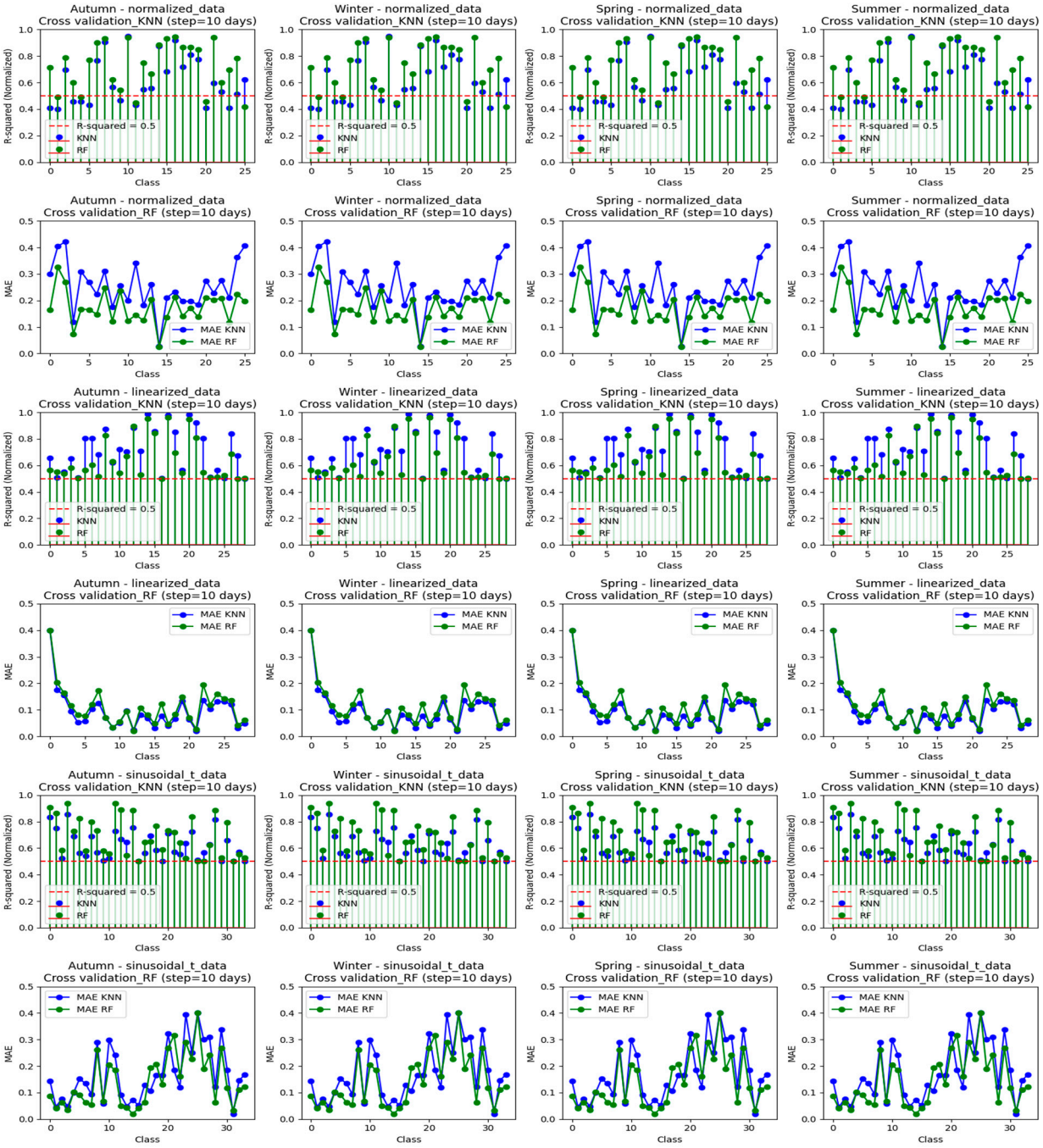


Figure 7. Cross-validation analysis for SPI forecasting via K-Nearest Neighbors (KNN) and Random Forest (RF) using the maximum time delay. Coefficient of determination (R-squared), mean absolute error (MAE).

5. Conclusions

This study underscores the critical need for innovative methodologies for forecasting non-stationary time series. By focusing on the daily standardized precipitation index (SPI) as a prime example, the research addresses the complexities arising from heterogeneous precipitation variability and its impact on the SPI temporal distribution, which poses significant challenges to several machine-learning models in fixing the model parameters to forecast data during different periods. Through the introduction of a novel approach centered on adjusting non-stationary distributions and simplifying time lag detection via autocorrelation, this study achieves substantial improvements in the forecasting accuracy. Moreover,

this study highlights the efficacy of mathematical transformations such as linearization, sinusoidal transformation, and normalization in stabilizing the data distributions over the whole observation, facilitating the capture of the maximum time lags, extending up to one year. The key findings highlight the pivotal role of data transformation in minimizing tendencies and standardizing both periodicity and seasonality, thereby enhancing the performance of machine-learning algorithms such as K-Nearest Neighbors (KNN) and Random Forest (RF). The results indicate RF's remarkable accuracy in forecasting the SPI, particularly when the data have a nonlinear distribution. Contrariwise, KNN faces a limitation, especially during the dryness period, where the mean values of the forecast data have a significance difference compared with the ground truth. However, it demonstrates the high performance of the KNN model when the data have a linear distribution. These findings emphasize the importance of adaptive methodologies in analyzing and forecasting climate data, offering valuable insights for climate change mitigation and risk management efforts. From another perspective, the model faces a limitation in generating a continuous data series when using different time lag values obtained from each seasonal sampling. In our forthcoming research, we aspire to explore additional transformation techniques, including data decomposition and smoothing employing parametric models. Our objective is to evaluate the efficacy of these methods in combination with deep-learning models for spatiotemporal forecasting.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/engproc2024068034/s1, Figure S1: Heatmaps of the R^2 scores for the KNN parameter optimization, tested across K values from 1 to 50, incorporating all the time delays from the cross-correlations.

Author Contributions: A.A., A.L., A.J. and M.A.Y. have directly participated in this study. A.A. worked on the coding, modeling, statistical analysis, validation, and co-editing. A.L. and A.J. worked on the research methodology, supervision, validation, co-editing, and reviewing. M.A.Y. worked on the validation and reviewing. All authors have read and agreed to the published version of the manuscript.

Funding: This project has been supported by European Union PNRR Funding under Italian DM 352/2022.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the Python code presented in this study can be requested from the corresponding author for reasonable purposes.

Acknowledgments: We wish to thank the Eurac Earth Observation Institute (Italy), for helping with data curation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hegerl, G.C.; Brönnimann, S.; Schurer, A.; Cowan, T. The early 20th century warming: Anomalies, causes, and consequences. *Wiley Interdiscip. Rev. Clim. Chang.* **2018**, *9*, e522. [CrossRef]
2. Mishra, A.K.; Singh, V.P. Drought modeling—A review. *J. Hydrol.* **2011**, *403*, 157–175. [CrossRef]
3. Fung, K.F.; Huang, Y.F.; Koo, C.H.; Soh, Y.W. Drought forecasting: A review of modelling approaches 2007–2017. *J. Water Clim. Chang.* **2020**, *11*, 771–799. [CrossRef]
4. AghaKouchak, A.; Pan, B.; Mazdiyasni, O.; Sadegh, M.; Jiwa, S.; Zhang, W.; Love, C.A.; Madadgar, S.; Papalexiou, S.M.; Davis, S.J.; et al. Status and prospects for drought forecasting: Opportunities in artificial intelligence and hybrid physical–statistical forecasting. *Philos. Trans. R. Soc. A.* **2022**, *380*, 20210288. [CrossRef]
5. Li, J.L.; Xu, K.M.; Jiang, J.H.; Lee, W.L.; Wang, L.C.; Yu, J.Y.; Stephens, G.; Fetzer, E.; Wang, Y.H. An overview of CMIP5 and CMIP6 simulated cloud ice, radiation fields, surface wind stress, sea surface temperatures, and precipitation over tropical and subtropical oceans. *J. Geophys. Res. Atmos.* **2020**, *125*, e2020JD032848. [CrossRef]
6. Ben-Bouallegue, Z.; Clare, M.C.; Magnusson, L.; Gascon, E.; Maier-Gerber, M.; Janoušek, M.; Rodwell, M.; Pinault, F.; Dramsch, J.S.; Lang, S.T.; et al. The rise of data-driven weather forecasting. *arXiv* **2023**, arXiv:2307.10128.

7. Kumar, S.; Merwade, V.; Kinter, J.L., III; Niyogi, D. Evaluation of temperature and precipitation trends and long-term persistence in CMIP5 twentieth-century climate simulations. *J. Clim.* **2013**, *26*, 4168–4185. [CrossRef]
8. Alomar, M.K.; Khaleel, F.; Aljumaily, M.M.; Masood, A.; Razali, S.F.M.; AlSaadi, M.A.; Al-Ansari, N.; Hameed, M.M. Data-driven models for atmospheric air temperature forecasting at a continental climate region. *PLoS ONE* **2022**, *17*, e0277079. [CrossRef]
9. Dalla Torre, D.; Lombardi, A.; Menapace, A.; Zanfei, A.; Righetti, M. Exploring the Feasibility of Data-Driven Models for Short-Term Hydrological Forecasting in South Tyrol: Challenges and Prospects. *SN Appl. Sci.* 2023, *under review*.
10. Vlahović, V.; Vujošević, I. Long-term forecasting: A critical review of direct-trend extrapolation methods. *Int. J. Electr. Power Energy Syst.* **1987**, *9*, 2–8. [CrossRef]
11. франко, П. Uncertainty in case of lack of information: Extrapolating data over time, with examples of climate forecast models. *Український Метрологічний Журнал/Ukr. Metrol. J.* **2022**, *3*, 3–8.
12. Mehrmolaei, S.; Keyvanpour, M.R. Time series forecasting using improved ARIMA. In Proceedings of the 2016 Artificial Intelligence and Robotics (IRANOPEN), Qazvin, Iran, 9 April 2016; pp. 92–97.
13. Samal, K.K.R.; Babu, K.S.; Das, S.K.; Acharaya, A. Time series based air pollution forecasting using SARIMA and prophet model. In Proceedings of the 2019 International Conference on Information Technology and Computer Communications, Singapore, 16–18 August 2019; pp. 80–85.
14. Afshar, N.R.; Fahmi, H. Rainfall forecasting using Fourier series. *J. Civ. Eng. Archit.* **2012**, *6*, 1258.
15. Teala, A. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Comput. Inform. J.* **2018**, *3*, 334–340. [CrossRef]
16. Cheng, C.; Sa-Ngasoongsong, A.; Beyca, O.; Le, T.; Yang, H.; Kong, Z.; Bukkapatnam, S.T. Time series forecasting for nonlinear and non-stationary processes: A review and comparative study. *Ice Trans.* **2015**, *47*, 1053–1071. [CrossRef]
17. Hewamalage, H.; Bergmeir, C.; Bandara, K. Recurrent neural networks for time series forecasting: Current status and future directions. *Int. J. Forecast.* **2021**, *37*, 388–427. [CrossRef]
18. Abbasimehr, H.; Paki, R. Improving time series forecasting using LSTM and attention models. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 673–691. [CrossRef]
19. Zhu, S.; Xu, Z.; Luo, X.; Liu, X.; Wang, R.; Zhang, M.; Huo, Z. Internal and external coupling of Gaussian mixture model and deep recurrent network for probabilistic drought forecasting. *Int. J. Environ. Sci. Technol.* **2021**, *18*, 1221–1236. [CrossRef]
20. Salles, R.; Belloze, K.; Porto, F.; Gonzalez, P.H.; Ogasawara, E. Nonstationary time series transformation methods: An experimental review. *Knowl.-Based Syst.* **2019**, *164*, 274–291. [CrossRef]
21. Mentaschi, L.; Vousdoukas, M.; Voukouvalas, E.; Sartini, L.; Feyen, L.; Besio, G.; Alfieri, L. The transformed-stationary approach: A generic and simplified methodology for non-stationary extreme value analysis. *Hydrol. Earth Syst. Sci.* **2016**, *20*, 3527–3547. [CrossRef]
22. Surakhi, O.; Zaidan, M.A.; Fung, P.L.; Hossein Motlagh, N.; Serhan, S.; AlKhanafseh, M.; Ghoniem, R.M.; Hussein, T. Time-lag selection for time-series forecasting using neural network and heuristic algorithm. *Electronics* **2021**, *10*, 2518. [CrossRef]
23. Grigonytė, E.; Butkevičiūtė, E. Short-term wind speed forecasting using ARIMA model. *Energetika* **2016**, *62*, 45–55. [CrossRef]
24. Tajmouati, S.; Wahbi, B.E.; Bedoui, A.; Abarda, A.; Dakkon, M. Applying k-nearest neighbors to time series forecasting: Two new approaches. *arXiv* **2021**, arXiv:2103.14200. [CrossRef]
25. Hu, J.; Liu, B.; Peng, S. Forecasting salinity time series using RF and ELM approaches coupled with decomposition techniques. *Stoch. Environ. Res. Risk Assess.* **2019**, *33*, 1117–1135. [CrossRef]
26. Renaud, O.; Victoria-Feser, M.-P. A robust coefficient of determination for regression. *J. Stat. Plan. Inference* **2010**, *140*, 1852–1862. [CrossRef]
27. Liao, J.; McGee, D. Adjusted coefficients of determination for logistic regression. *Am. Stat.* **2003**, *57*, 161–165. [CrossRef]
28. Hodson, T.O. Root mean square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev. Discuss.* **2022**, *15*, 5481–5487. [CrossRef]
29. Kim, T.K. T test as a parametric statistic. *Korean J. Anesthesiol.* **2015**, *68*, 540–546. [CrossRef]
30. Majone, B.; Villa, F.; Deidda, R.; Bellin, A. Impact of climate change and water use policies on hydropower potential in the south-eastern Alpine region. *Sci. Total Environ.* **2016**, *543*, 965–980. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.