EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR ASSESSING DECADAL VARIABILITY IN EUROPEAN SUBSEASONAL FORECASTS

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

JING CHEN

14486679

MASTER INFORMATION
STUDIES DATA SCIENCE
FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

SUBMITTED ON FEBRUARY 2, 2025

Supervisor Title, Name Dr. Ali
Alsahag Affiliation UvA Supervisor
Email a.m.m.alsahag@uva.nl

ABSTRACT

Subseasonal forecasts (SSFs) are essential for managing climate risk. However, their accuracy is significantly influenced by decadal climatic changes. As climate shift continues, this challenge is expected to grow. To address this, we apply Explainable AI (XAI) to improve transparency and understanding of deep learning models in climate forecasting. XAI helps reveal how models make predictions, ensuring interpretability and scientific reliability. This study uses a ConvLSTM2D deep learning model to predict European SSFs and compares it to a baseline Random Forest model. Both models are trained on high-resolution ERAS5 reanalysis data. We ap- ply Gradient-weighted Class Activation Mapping (Grad-CAM) and Feature Importance Analysis (FIA) to explain the model's decision- making process and identify key meteorological drivers.

The model's overall prediction performance is modest (R? — = -0.0424), highlighting its current limitations in subseasonal forecasting. However, XAI provides useful insights by identifying meteorologically relevant patterns, such as synoptic systems and moisture gradients. FIA highlights important variables like sea level pressure, 2-meter temperature, and total column water vapor. Temporal Feature Importance also shows the model's reliance on historical data, with both a recency effect and a persistent influence from past time steps. These findings show the potential of XAI to improve the ability of deep learning models to interpret climate science, offering necessary diagnostic information to refine models and providing a path to developing more trustworthy, transparent SSFs systems for Europe and elsewhere, strengthening climate measures in response to impactful decadal climate changes.

KEYWORDS

Subseasonal forecasting, Deep learning, CNN-LSTM, Explainable AI

GITHUB REPOSITORY

https://github.com/Jingchen222/Thesis-Project

modern society. They support decision-making in areas such as energy, agriculture, public health, and disaster preparation. In Europe, where climate conditions vary and extreme weather events pose risks, the demand for accurate predictions is especially high [29, 32). Traditional weather forecasts focus on short-term timescales, while seasonal climate predictions provide long-range outlooks. However, a gap remains in the subseasonal range, covering forecasts from two weeks to two months [29]. Subseasonal forecasts (SSFs) bridge the gap between short-term weather predictions and seasonal climate outlooks, offering critical lead time for climate risks management [30, 32]. Accurate SSFs are becoming more important as climate change increases extreme weather events. In Europe, extreme weather such as heatwaves, droughts, and heavy rainfall have become frequent and severe. Southern Europe is experiencing more droughts, while northern regions face more flooding. These changes highlight the need to improve SSFs to enhance climate adaptation measures[2, 6]. However, improving the accuracy of SSFs remains a major challenge due to the complexity of atmospheric processes and interactions within the climate system [29, 30]. A of the key factor affecting the performance of SSFs is the decadal climate variability, which involves natural shifts in climate patterns over timescales of 10 to 30 years [14, 22]. Large-scale climate oscil- lations, such as the Atlantic Multidecadal Oscillation (AMO) and North Atlantic Oscillation (NAO), strongly affect regional weather patterns and impact seasonal and subseasonal forecasts [14, 22, 25). In Europe, the NAO affects winter temperatures and precipita- tion, while the AMO changes weather patterns over longer time scales [4, 25]. Additionally, Anthropogenic climate change driven by greenhouse gas emissions (GHG) adds further complexity. It changes the characteristics of these decadal oscillations and further impact the reliability of forecasting systems [5, 6]. Understand- ing these decadal changes is crucial to improve robustness and accuracy of European SSFs, ensuring a better climate adaptation in the rapidly changing world. Despite the clear impact of decadal changes on SSFs, research on its effect on European SSFs remains limited. However, recent progress in machine learning (ML) and explainable artificial intelligence (XAI) offers potential solutions to address this gap [8, 20]. ML models, specifically deep learning architectures like Convolutional Neural Networks (CNNs), Recur- rent Neural Networks (RNNSs) such as Long Short-Term Memory networks (LSTMSs), as well as Transformer networks, have shown outstanding abilities to capture complex spatial and temporal pat- terns large climate datasets [10, 12, 21).

However, ML models are often seen as "black boxes", meaning their internal decision-making processes are not easily understood [27]. For example, a deep learning model might predict an upcoming cold surge over Europe when forecasting, but is unable to provide explicit reason and explanation of which factors are responsible (e.g. atmospheric pressure, oceanic patterns, or humidity levels)—led to

XAI technology, models can better understand the decisions they make and their processes. This helps to determine which input fea- tures or patterns have the most influence on SSFs, thus providing a clearer link between model outputs and real-world atmospheric dynamics. It also improves trust, supports better model refinement, and enhances the practical use of ML in subseasonal forecasting [20, 34].

The goal of this study is to extend previous research on SSFs [1] by assessing the impact of decadal climate changes on European datasets. Using an innovative ML-XAI integration approach to en- hance understanding and improve prediction accuracy.

We aim to demonstrate the effectiveness of this ML and XAI intergration approach in improving the decadal variability of subseasonal forecasting. Hence, we address the following research question: To what extent does decadal climate variability impact subseasonal forecasting performance, and how can Explainable AI improve the interpretability and reliability of SSFs models? To answer this research question, we have formulated the following sub-questions:

e Which atmospheric variables play a key role in decadal vari- ability and how do they affect SSFs? e To what extent does the ConvLSTM2D deep learning model capture subseasonal patterns in Europe, and how does its forecasting performance compare to benchmark models. e To what extent can XAI methods help reveal the decision-making processes of deep learning model in SSFs, improving their interpretability?

This research aims to uncover how decadal variability affect SSFs performance and how XAI can help interpret these effects, provid- ing valuable insights for improving future climate modeling.

2 RELATED WORK

To provide context for the current study, this section examines the existing literature across four main areas: subseasonal, decadal climate changes and their influence on forecasting, applications of ML in SSFs, and the emerging field of XAI in climate science.

2.1 Subseasonal Forecasting

Subseasonal to seasonal (S2S) forecasting bridges the gap between short-term weather predictions and seasonal climate outlooks, pro- viding lead times of two weeks to two months [29]. SSFs are es- sential for water resource management, agriculture, and disaster preparedness, but remain challenging due to the complex interac- tions between atmospheric and oceanic processes [29, 32]. Unlike short-term forecasts, which rely mainly on initial conditions, and seasonal forecasts, which depend on slowly evolving boundary con- ditions, subseasonal predictability is influenced by both [30]. Key factors affecting SSFs include large-scale climate oscillations, such as the North Atlantic Oscillation, Quasi-Biennial Oscillation (QBO), Madden-Julian Oscillation (MJO), and mid-latitude atmospheric blocking [25, 29). Traditional forecasting methods use dynamical models, which simulate atmospheric processes, and statistical ap- proaches, which identify historical patterns [8]. However, dynami- cal models lose accuracy over longer lead times, requiring statistical post-processing and bias correction [8]. These limitations empha- size the need for alternative approaches, particularly machine learn- ing, which can capture nonlinear relationships in climate data more effectively than traditional methods. This study addresses these challenges by applying deep learning to SSFs, specifically using a ConvLSTM2D model, to analyze how subseasonal temperature predictions in Europe are influenced by decadal climate variability.

2.2 Decadal Climate Variability

152 A major challenge in SSFs is the impact of decadal climate vari- 153 ability, which refers to natural shifts in climate patterns over 10-30 154 years [14, 22]. These long-term oscillations, such as the North At- 155 lantic Oscillation, Atlantic Multidecadal Oscillation, and Pacific Decadal Oscillation (PDO), introduce variability in atmospheric and oceanic conditions that can significantly affect SSFs accuracy [22]. In Europe, the NAO strongly influences winter temperatures and precipitation, while the AMO modulates sea surface temper- atures (SSTs)and alters atmospheric circulation and storm tracks [4, 25). These decadal fluctuations affect subseasonal predictability by changing the persistence and recurrence of key weather pat- terns, making it difficult for forecasting models to maintain skill over extended periods. Furthermore, anthropogenic climate change adds complexity to the climate system. For example, greenhouse gas emissions, artificial cooling, and other human activities alter baseline climate conditions and also modify the frequency, ampli- tude, and spatial patterns of these decadal oscillations [5, 6]. This evolving climate poses a challenge for SSFs models, which rely on historical relationships between climate drivers and atmospheric responses.

2.3 Machine Learning for Subseasonal Forecasting

Machine learning techniques have become powerful tools in climate modeling, including subseasonal forecasting [8, 31]. ML can learn complex, non-linear relationships from large datasets, potentially overcoming the limitations of traditional statistical and dynamical approaches in capturing climate system details [8].

One commonly used ML model in SSFs is Random Forest, which handles high-dimensional data and captures nonlinear patterns by aggregating multiple decision trees. RF has been successfully applied in predicting atmospheric blocking events and subseasonal precipitation extremes [35, 36]. However, it struggles with temporal dependencies, making it less effective for time-series forecasting. While RF is a strong baseline, its limitations in handling complex interactions and time-series dynamics highlight the need for more advanced SSFs methods.

More advanced ML approaches, such as Recurrent Neural Networks, are particularly useful for modeling temporal dependencies. Long Short-Term Memory networks, for example, have been used to pre- dict Madden-Julian Oscillation phases, capturing their intraseasonal evolution and improving forecasting[3]. Similarly, Gated Recurrent Units (GRUs) have been applied to seasonal temperature and precip- itation forecasting, demonstrating better generalization compared to traditional models [13]. Convolutional Neural Networks, orig- inally developed for image recognition, are effective in climate forecasting due to their ability to extract spatial patterns from grid- ded climate data. They have been applied to downscaling climate models and predicting tropical cyclone formation by identifying key spatial features in atmospheric variables [12, 21]. More re- cently, Transformer models, known for their attention mechanisms, have shown promise in long-term climate forecasting. For instance, self-attention-based architectures have improved in ENSO (El Nifo- Southern Oscillation) prediction by capturing ocean-atmosphere interactions across multiple time scales, outperforming traditional statistical models [10]

Hybrid architectures, such as CNN-LSTMs, combine the strengths

211 of both CNNs (spatial feature extraction) and LSTMs (temporal 212 modeling) and have been proposed for SSFs applications. These 213 models have been used for predicting atmospheric river events and 214 sub-seasonal temperature anomalies, demonstrating higher skill 215 than traditional dynamical models [13, 16]. Despite their advan- 216 tages, deep learning models require large, high-quality datasets for 217 effective training, emphasizing the importance of ERA5 reanalysis 218 data for SSFs applications [31].

This study builds on these advancements by evaluating the Con- vLSTM2D model's ability to capture both spatial and temporal patterns in SSFs, particularly under the influence of decadal climate variability. By integrating explainable AI techniques, we aim to gain deeper insights into the model's learning process and assess its suitability for subseasonal forecasting in a changing climate.

226 24 Explainable AI in Climate Science

As ML models become even more complex and widely applied in climate science, including SSFs, the need for explainability and interpretability has become essential [20, 34]. Although deep learning models can attain remarkable prediction performance, their "black box" nature often limits our awareness of their decision-making, preventing trust and scientific insight. XAI techniques present a wide range of approaches to deal with this difficulty, intending to open the "black box" and determine how ML models make predic- tions [20]. Among several XAI approaches, Grad-CAM has become valuable for visualizing the temporal features and spatial regions that are most relevant to modeling predictions, stressing aspects of attention and focus within complicated sets of climate data [34]. Furthermore, FIA methods, such as ones based on masking or per- turbation, can assess the importance of input variables or temporal procedures in determining model outputs, thereby providing a detailed understanding of the factors that influence SSFs [1]. By applying XAI techniques to SSFs models, important insights into model behavior can be gained, possible limitations or biases can be identified, trust in model predictions can be established, and guidance can be provided for the creation of more reasonable and scientific prediction systems. [20].

3 METHODOLOGY

3.1 Data Sources and Preprocessing

3.1.1 ERA5 Dataset. This study uses ERA5 reanalysis datasets from the European Centre for Medium-Range Weather Forecasts (ECMWE) [11] as the primary dataset for SSFs from 1980 to 2020. The ERAS5 reanalysis dataset was chosen for its high spatial and temporal resolution and its inclusion of atmospheric and oceanic variables. These features make it well-suited for analyzing SSFs dynamics in the European climate.

This study focuses on key atmospheric variables, including total precipitation (tp), mean sea level pressure (msl), 10 m wind compo- nents (u10, v10), 2-meter temperature (t2m), and total column water vapor (tcwv). These variables were selected due to their strong link to climate change in Europe and their influence on SSFs forecasting. For instance, msl plays a key role in large-scale circulation patterns, such as the NAO, which strongly affects European weather [4]. tp and tcwv help track moisture transport and precipitation extremes, both essential for subseasonal weather patterns [25]. t2m is a key indicator of surface climate conditions, making it essential for temperature forecasting. Lastly, u10 and v10 provide insights into storm dynamics and atmospheric flow, which influence regional weather variability. By incorporating these variables, we aims to capture key subseasonal climate signals and assess how decadal climate variability impacts SSFs performance in Europe.

3.1.2 Data Preprocessing. Since the ERA5 reanalysis data contain huge amounts of European climate data and the DL models used in this study (ConvLSTM2D) require high computational power, we conducted our study on GPUs. Before the model was trained, the ERAS5 data were preprocessed, and exploratory data analysis (EDA) was conducted. The EDA results are presented in the following section. During preprocessing, the raw data was converted into input sequences designed for the ConvLSTM2D model, including a four-dimensional reshaping to reflect latitude, longitude, chan- nels, and time steps. As shown in the CNNTEST. ipynb, in order to optimize computational efficiency and enhance the signal-to-noise ratio, GPU-accelerated PCA (via CuPy) was used. The GPUPCAPro- cessor class in modeltest.ipynb handled this process. PCA was im- plemented across spatial dimensions (latitude, longitude) to obtain dominant spatial patterns while maintaining 95% data variance,as shown in Figure 1. This approach reduces computational overhead and ensure the model focuses on the most relevant spatial features. As shown in Figure 2, there are correlations present between the

Figure 1: PCA Explained Variance Analysis.

climate variables in the ERA5 data. These variables were chosen due to their known impact on European subseasonal variability. For instance, tcwv and t2m show a strong correlation, suggesting that capturing water vapor dynamics can help predict temperature. This highlights the importance of preprocessing techniques like PCA to address multicollinearity and improve model efficiency. After reducing the dimensionality, the data was normalized using the StandardScaler from scikit-learn. This process was applied to both features and target variables in CNNTEST.ipynb (within the DataPreprocessor class) and in modeltest.ipynb (within the GPUP- CAProcessor class). Normalization ensures that all input features are on a comparable scale by removing the mean as well as scaling to unit variance.

In preprocessing the data for the Random Forest model, dimen- sionality reduction was not applied. Instead, retaining the original feature space allows the model to perform feature selection inter- nally. All input features were normalized using StandardScaler to ensure comparability between variables. This approach aligns with

310 the preprocessing used for the ConvLSTM2D model. The differences 311 in preprocessing reflect the distinct requirements of tree-based methods and neural network architectures.

Table 1: ConvLSTM2D Model Summary. Placeholder for Figure: Model Summary table from CNNTEST. ipynb

Figure 2: Variable Correlation Matrix, displaying pairwise correlations between key climate variables (t2m, msl, tp, tcwv) and illustrating their inter-dependencies within the ERA5 dataset.

313 3.2 ConvLSTM2D Model Architecture

The main element of the forecast framework used in this study is a Convolutional Long Short-Term Memory (ConvLSTM2D) neural network, which is designed to reproduce the spatial and tempo- ral complexities of subseasonal climate data. ConvLSTM2D was chosen for its ability to process sequences of data while also obtain- ing spatial features during convolutional operations. This makes it well-suited for climate prediction applications [13, 16). The model architecture, described in the WeatherForecastModel class inside CNNTEST. ipynb, consists of a deep sequenced network. The corre- sponding formulas are shown in the following table 1. As shown in Table 1, the model architecture includes an Input Layer defining the shape of the input sequence. It consists of two stacked ConvL- STM2D layers with batch normalization (32 and 64 filters), followed by a Conv2D layer with batch normalization and MaxPooling2D (64 filters). A GlobalAveragePooling2D layer is then applied, followed by two dense layers (128 and 64 units) with Dropout and batch normalization. Finally, a dense output layer completes the architec- ture. The model is compiled using the Adam optimizer (learning rate = 0.001), with mean squared error (MSE) as the loss function. Evaluation metrics include mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (Rô) [28] In this study, RF is used as a baseline model, implemented through the Scikit-learn library. It consists of 200 estimators with a max- imum depth of 6, optimized using a grid search method. Unlike the ConvLSTM2D, which extracts sequential and spatial feature extraction, RF accessed variable using it built-in feature importance mechanism to evaluate the contribution of individual variables to subseasonal forecasts. This approach highlights the utility of RF in understanding variable importance while serving as a benchmark for the hybrid deep learning framework.

3.3 Experimental Setup and Training

The experimental design involved a chronological partitioning of the ERA5 dataset into training (70%), validation (15%), and test- ing (15%) sets, consistent with CNNTEST. ipynb. A rolling-window method was implemented for training, using historical data sequences to predict SSFs. This technique ensures that historical data sequences are used to predict future states while preserving temporal consistency. The training process was shaped with a sequence length of 30 time steps and a batch size of 16, as per CNNTEST. ipynb. The model was trained for 50 epochs, with early stopping imple- mented to prevent overfitting and to optimize the training time.

For the Random Forest model, the same partitioning of the ERA5 dataset (70% training, 15% validation, and 15% testing) was adopted to ensure comparability with the ConvLSTM2D model. Hyperpa- rameter tuning, including the number of estimators, maximum depth, and minimum child weight, was performed using a grid search method to optimize model performance. Unlike ConvL- STM2D, which required sequential input data, RF uses a flat feature matrix, maintaining temporal resolution through explicit feature engineering.

To optimize training, callbacks implemented via the WeatherTrainer class in CNNTEST. ipynb. For example, ModelCheckpoint was used to save the optimal validation loss weights. EarlyStopping (patience

369 = 10) was applied to prevent overfitting. ReduceLROnPlateau (factor 410 370 = 0.5, patience = 5) was used to dynamically adjust the learning rate. 411 371 Finally, TensorBoard was used to visualize the training metrics.

34 Model Evaluations

Different evaluation methods are used to assess model performance.

3.4.1 MSE. MSE calculates the mean squared difference between predicted and actual values [28]. In SSFs, the model predicts meteo- rological variables (e.g., temperature, precipitation) for the next 7 to 28 days.

378 However, MSE has limitations. Since it squares the errors, large 379 errors can disproportionately increase the final loss value. This 380 makes MSE highly sensitive to outliers, which can distort model 381 training [9]. 382 34.2 MAE. MAE is implemented to measure the average magni- 383 tude of prediction errors. It is calculated as:

384 which emphasizes subseasonal predictions throughout Europe, MAE 385 shows how close these predictions are to true values while demon- 386 strating less sensitivity to extreme outliers.[33] However, since 387 MAE minimizes absolute differences, the model tends to predict 388 the median instead of the mean of the target distribution, which 389 can introduce prediction bias. 390 3.4.3 _RMSE. RMSE measures the difference between predicted and 391 actual values in regression tasks, providing a quantitative evalua- 392 tion of a models accuracy [28]. It is particularly useful for assessing 393 subseasonal forecasts across Europe. The dataset consists of numer- 394 ical temperature values, requiring precise predictions to capture 395 seasonal anomalies and variations effectively.

396 Similar to MSE, RMSE squares the errors before averaging, making 397 it sensitive to outliers. Additionally, RMSE assumes that errors are 398 normally distributed. If the error distribution is skewed or has heavy 399 tails, RMSE may fail to accurately reflect model performance. 400 3.4.4 _R-Squared. The R?score is used to evaluate how well the 401 models in a task and explains the variance in the dependent variable 402 [9]. It is calculated as:

403 TheR? core ranges from —oo to 1, with a value close to 1 indicat- 404 ing that the model can explain most of the variance in the target 405 variable.R?is particularly valuable for assessing the goodness-of-fit 406 of ML models.However, R-squared has limitations. For example, a 407 high R-squared does not guarantee model accuracy, as the model 408 may still have large errors, which MSE or RMSE can better reflect. 409 Additionally, R-squared always increases or remains unchanged

3.5 Explainable AI Techniques

interpret the decisions of the ConvLSTM2D model and under- the drivers of SSFs, we applied XAI techniques, Grad-CAM, feature importance analysis. e Grad-CAM (Gradient-weighted Class Activation Map- ping): Grad-CAM was used to visualize the spatial and temporal regions that is influential to the model's predictions and highlight specific convolutional layers [7]. Heatmaps were created by calculating the gradients of the predicted output in relation to the feature maps of this layer. These gradients are aggregated and weighted to identify key focus areas. Finally, the heatmaps is overlaid on the input climate data to enhance interpretability. Feature Importance Analysis: An analytical approach quantifies the importance of input variables and time steps 15]. — Temporal Importance: Temporal significance was assessed by systematically masking each time step and quan- tifying the resulting changes in predictions against a base- line. — Feature Channel Importance: The importance of each feature channel was evaluated by masking individual chan- nels and measuring the consequent prediction changes relative to a baseline [30].

In addition to Grad-CAM and Feature Importance analysis of the ConvLSTM2D model, RF is used to assess the relative contribution of input variables to forecasts. By computing the Gini impurity de- crease for each variable,the key drivers of subseasonal forecasting can be identified. Variable importance complements Grad-CAM results, providing a comparative evaluation of both model architec- tures.

4 RESULTS

This section presents the experimental results, including model performance and the impact of the XAI technique. The goal is to assess the performance of ConvLSTM2D and identify key influenc- ing factors, particularly in the context of decadal climate variability. ConvLSTM2D is compared with the RF model to highlights ConvLSTM2D's strengths and provides a deeper understanding of its capabilities.

4.1 Model Performance Evaluation

The performance of the ConvLSTM2D model was evaluated based on a held-out validation dataset with standard metrics, including MSE, RMSE, MAE, and R?. Table 2 presents the performance metrics for RF and ConvLSTM2D, providing a brief summary of their pre- diction techniques. This comparison helps assess the strengths and limitations of ConvLSTM2D compared to traditional machine learn- ing. Table 2 presents the performance metrics of the ConvLSTM2D and RF models on the validation dataset. The evaluation includes MSE, RMSE, MAE, and R° scores, providing a comparison of model

Table 2: Model Performance Metrics on Validation Dataset

accuracy. The results show that ConvLSTM2D underperforms com- pared to RF across all metrics. ConvLSTM2D has a higher MSE (14.4969 vs. 7.1534) and RMSE (3.8075 vs. 2.6746), indicating lower predictive accuracy. Additionally, the MAE of RF (2.0707) is lower than that of ConvLSTM2D (3.0845), suggesting that RF produces smaller average errors. The R' scores highlight the performance gap between the models. RF achieves a positive R? (0.5624), meaning it explains a significant portion of the variance, capturing linear or simpler nonlinear patterns effectively. In contrast, ConvLSTM2D has a negative R? (-0.0424), indicating it fails to outperform a simple mean predictor. However, it is important to note that these perfor- mance metrics alone do not fully capture the insights provided by ConvLSTM2D and its interpretability techniques. In the following sections, we explore how xai methods help to provide a valuable perspectives on the learning mechanisms and potential benefits of the model beyond the evaluation.

To visually represent the model's learning process, Figure 3 shows the loss and MAE curves for training and validation, respectively. These plots track the progression of metrics during training and offer visual insights into the model's learning behavior and general- ization ability. As shown in Figure 3, the training and validation loss

Figure 3: ConvLSTM2D Model Training and Validation Loss Curves

and MAE curves are converging, indicating that the ConvLSTM2D model is learning from the data. Although some fluctuations appear in the later stages, the overall downward trend suggests that overfit- ting is not severe and the model has a certain level of generalization ability. This convergence is a key sign of the model's effectiveness, showing that it can extract relevant features from training data and apply them to validation data. This provides a foundation for future predictions. However, there is still room for improvement, particu- larly in refining the model to enhance prediction accuracy, espe- cially for climate-related deep learning models. To further analyze predictive performance, Figure 4 presents a time-series comparison of the model's predictions and actual observations for the first 100

Figure 4: Comparison of Predictions and Ground Truth Time Series (First 100 Samples)

samples in the test dataset. As shown in Figure 4, the model's pre- dictions (orange dashed line) tend to cluster around the mean. This suggests underdispersion and a limited ability to capture the full range of variations and extreme fluctuations in the ground truth data (blue solid line). While the model follows the general trend of the time series and captures overall temporal patterns, it under- estimates peaks and overestimates lower values. This indicates a systematic bias toward the mean forecast. Such behavior is com- mon in climate prediction, especially at subseasonal scales, where models often struggle to predict the intensity of extreme weather events accurately. This indicates a systematic bias toward the mean forecast. Such behavior is common in climate prediction, especially at subseasonal scales, where models often struggle to predict the intensity of extreme weather events accurately. To further examine this pattern, Figure 5 presents a scatter plot comparing predictions with ground truth, providing a clearer view of their relationship.

Figure 5: Scatter Plot: Predictions vs. Ground Truth Observa- tions

In Figure 5, the majority of data points are tightly clustered along the zero-prediction line, deviating significantly from the ideal

515 1:1 relationship (red dashed line). This confirms the model's mean 516 reversion tendency and its difficulty in capturing extreme values. 517 Ideally, if the model were to predict perfectly, all data points should 518 be closely aligned along the red dashed line. However, the current 519 distribution reveals the model's predictive shortcomings, suggest- 520 ing areas for improvement. Additionally, the residual distribution 521 in Figure 6 provides further insights into the nature of the model's errors.

Figure 6: Distribution of Prediction Residuals (Prediction - Ground Truth)

The histogram in Figure 6 reveals that the prediction residuals exhibit a near-normal distribution centered around zero, suggesting that the errors are largely unbiased in their mean. This is a positive aspect, indicating that on average, the model is not systematically over- or under-predicting. However, the distribution also shows a relatively wide spread and a pronounced peak, indicating a substan- tial variance in prediction errors and a systematic tendency towards under-prediction of extreme events, consistent with the time series and scatter plot analyses. Despite the near-zero mean error, the large variance in errors implies that the reliability of the model predictions still needs improvement, especially for extreme weather forecasting. Finally, Figure 7 presents the temporal evolution of the prediction residuals across the test dataset.

The time series plot of residuals in Figure 7 shows that the residuals fluctuate around the zero line, with no apparent trend of increasing or decreasing bias over time. It indicates that the model's systematic underestimation is relatively consistent throughout the time domain, rather than worsening or improving at specific periods. This temporal stability in the bias is critical for understanding the error characteristics of the model, suggesting that the bias may originate from the model architecture or the training data itself, rather than time-dependent factors.

4.2 Grad-CAM Visualization of Model Attention

546 In order to gain insight into the decision-making process of the 547 ConvLSTM2D model and identify the spatial-temporal patterns

Figure 7: Time Series of Prediction Residuals

that drive its predictions. Grad-CAM visualization was used as shown in Figure 8 shows an example of a Grad-CAM heatmap, on the input climate data, highlighting the regions and timesteps that the model considers most important for its predictions. The

Figure 8: Examples of Grad-CAM Heatmaps, overlaid on in- put climate data, visualizing model attention regions for selected forecast instances

results of the Grad-CAM heat map shows some interesting patterns in the model's attention mechanism. For example, in the case of accurate temperature predictions, the heat map typically highlights regions associated with weather-scale weather systems, such as mid-latitude cyclones and frontal boundaries. This suggests that the model is indeed learning to focus on weather-related atmospheric features. This strongly suggests that the model is identifying and uti- lizing physically meaningful atmospheric patterns for forecasting. In addition, in some cases, the Grad-CAM visualization suggests that the model is sensitive to regions with strong humidity gradi- ents or enhanced atmospheric instability, which is consistent with the understanding that moisture fluxes and convective processes play a key role in sub-seasonal temperature changes in Europe. This consistency with known meteorological processes further enhances the credibility of the model's learning representation and suggests that the model is capturing physically plausible drivers of climate change. However, it can also be observed that in cases of poor model performance, particularly when predicting extreme temperature anomalies, the Grad-CAM heat maps sometimes exhibit less coher- ent or spatially dispersed patterns of attention, suggesting that the model may have difficulty focusing on the most critical predictive features under extreme conditions.

4.3 Feature Importance Analysis

Additionally, we performed feature importance analysis to assess the relative contribution of different input variables and time steps

577 to the model performance. Figure 9 shows the temporal feature im- 578 portance scores, quantifying the impact of masking each time step in the input sequence on the model's predictive performance. As

Figure 9: Temporal Feature Importance Analysis: Impact of Time Step Masking on Model Predictions

shown in Figure 9, the temporal importance scores exhibit a clea pat- tern, with time steps closer to the forecast initialization time gener- ally exhibiting higher importance scores. This finding is consistent with intuitive expectations, where recent atmospheric conditions are anticipated to exert a stronger influence on near-term subsea- sonal forecasts. This aligns with meteorological common sense and validates the consistency of the model's learning with physical principles. However, it is worth noting that even time steps further removed from the initialization time, particularly those in the earlier part of the 30-day input sequence, still retain a non-negligible level of importance. This robustly demonstrates that the ConvLSTM2D model is indeed leveraging the temporal memory capabilities of LSTM layers to integrate information from the longer historical context, capturing the influence of slower-evolving climate signals on subseasonal timescales. This highlights the model's ability to capture climate information across different time scales, rather than solely relying on the most recent atmospheric state, which is crucial for enhancing the accuracy and reliability of subseasonal forecasts. After multiple runs of the model, the results consistently show that

Figure 10: Feature Channel Importance Analysis: Ranking of Input Variable Importance

msl and t2m are consistently the most influential variables of the ConvLSTM2D model for subseasonal forecasts in Europe. This find- ing is highly aligned with the meteorological understanding that sea level pressure is the primary indicator of large-scale general circulation models, which are the main drivers of climate change in Europe.Similarly, the 2-meter temperature, which is the target variable itself, naturally exhibits a high degree of autocorrelation and has a strong influence on its own sub-seasonal evolution. The model's understanding of the significance of t2m also reflects its ability to effectively capture the inherent time dependence of the target variable. Notably, total column water vapor also consistently ranks among the more important variables, highlighting the key role of atmospheric moisture content and transport in regulating temperature and precipitation patterns on subseasonal time scales over Europe. The importance of water vapor as a key component of the atmospheric energy and moisture cycle further confirms that the model is learning key physical processes within the cli- mate system. These feature significance results, combined with the Grad-CAM visualization, provide a multifaceted perspective on the learned representation of the ConvLSTM2D model and high- light the potential of XAI technology to extract meteorologically meaningful insights from complex deep learning models in climate science.

Figure 11: Random Forest Prediction Uncertainty

Figure 12: ConvLSTM2D Prediction Uncertainty (MC Dropout)

To further assess the reliability of the forecasting models, we conducted an uncertainty analysis using bootstrap and MC Dropout methods. For RF, we applied the streaming t-digest method to a subsample of the validation data. Figure 11 displays the predicted means of selected samples with their 95% confidence intervals. For example, the predicted mean for sample 0 is 0.75, with a 95% CI of [0.18, 0.96], while sample 1 has a mean of -1.77 and a 95% CI of [-2.66, -0.99]. These results highlight the presence of considerable uncertainty in some predictions, even when point estimates are provided.

Similarly, We quantified the uncertainty of the ConvLSTM2D model by activating MC Dropout during inference on the validation batch. By performing multiple stochastic forward passes, a distri- bution of predictions was obtained for each sample. As illustrated in Figure 12, most predictions have relatively narrow confidence intervals; however, in some cases (e.g., sample 9) exhibit much wider intervals. These findings highlight the inherent stochasticity of deep learning outputs and the need to consider uncertainty in performance evaluation.

A paired statistical test on the prediction errors of both mod- els showed a significant difference ($p < 0.01$), confirming that the performance gap is unlikely due to random chance.

5 DISCUSSION

This section reflects on the obtained results, and provides limita- tions for this study. The study highlights the model's strengths and limitations in subseasonal weather forecasting in Europe, con- sidering decadal climate change. The XAI technique explains the model's learning patterns and decision-making. These insights help diagnose model issues and guide future research.

5.1 model performance

The poor performance of the ConvLSTM2D model in subseasonal forecasting raises important questions about its suitability for this application. While deep learning has the potential to exploit spatial and temporal dependencies, the results suggest that feature selection, hyperparameter tuning, and model architecture may have contributed to its weak predictive ability. The current feature set, though meteorologically relevant, may not be fully optimized for deep learning. Key climate indices such as NAO, AMO, and MJO, which significantly impact subseasonal variability, are not explicitly incorporated. Additionally, minimal hyperparameter tuning may have limited the model's ability to generalize, as default settings for learning rate, batch size, and filter size may not be optimal for capturing long-term climate dependencies. Moreover, the ConvLSTM2D architecture struggles with vanishing gradients, making it difficult to effectively capture slow-moving climate patterns and decadal variability. A major limitation is the poor representation of extreme tempera- ture events in the training data. As shown in Figure 4, the model underestimates peak values and overestimates moderate ones, lead- ing to unreliable predictions for high-impact anomalies. This sug- gests that class imbalance and data sparsity may hinder its learning process. Addressing this issue could involve synthetic oversam- pling or weighted loss functions to improve sensitivity to extreme events. Additionally, integrating physics-informed constraints into principles, ensuring greater physical consistency in predictions. Interestingly, RF significantly outperforms ConvLSTM2D, achieving a much higher R? score (0.5624 vs. -0.0424). This suggests that RF may be a stronger baseline than initially expected. Unlike deep learning models, RF effectively captures nonlinear relationships without suffering from error propagation over time, a common issue in recurrent architectures. Given these results, it is worth reconsidering whether deep learning is the best approach for sub- seasonal forecasting. While ConvLSTM2D was expected to offer advantages in modeling sequential dependencies, its performance suggests that simpler models, such as gradient boosting (e.g., XG- Boost, LightGBM), may be more effective. Future research should explore hybrid architectures such as CNN-LSTM or Transformer- LSTM, which may provide a better balance between spatial pat- tern recognition and temporal learning. Additionally, integrating physics-informed machine learning, where deep learning models incorporate fundamental climate principles, could lead to more reliable subseasonal forecasts.

Overall, these findings highlight the need for a more targeted approach when applying deep learning to subseasonal climate forecasting. While ConvLSTM2D shows some ability to capture relevant patterns, its predictive power remains weak, especially when compared to traditional machine learning methods. Future research should focus on optimizing feature selection, improving data preprocessing, refining hyperparameter tuning, and testing alternative architectures to determine whether deep learning can provide significant advantages over traditional methods in this field.

In the other hand, uncertainty analysis highlights the differences in reliability between the two models in SSFs. While most of the predictions have narrow confidence intervals, some samples (e.g., Sample 9 in the ConvLSTM2D model) show much wider inter- vals, indicating higher uncertainty. This suggests that although the models capture general patterns, their reliability for certain cases remains low. In particular, the ConvLSTM2D model, shows greater uncertainty in extreme cases, which also suggests model learning may be difficult in complex climate patterns.Statistical tests confirm a significant difference in prediction errors between the models ($p < 0.01$), indicating that one model is consistently more reliable. However, the presence of uncertainty in the predictions highlights the need to integrate uncertainty in model design. Future research could explore methods such as ensemble learning, Bayesian deep learning, or physics-informed constraints to improve reliability 19).

5.2 Deconstructing Model Intelligence: Interpretation of XAI Insights

A key finding is that the ConvLSTM2D model does not simply follow a complex algorithm—it focuses on meteorologically rele- vant patterns. The Grad-CAM visualization in Figure 8 shows that the model consistently highlights areas linked to weather systems, such as North Atlantic cyclone formation and frontal zones, both of which influence European weather.

732 dynamics, as some limitations remain. The XAI techniques used
733 in this study—Grad-CAM and feature importance analysis—serve
734 as diagnostic tools to help reveal how the model processes infor-
735 mation. Additionally, the model consistently focuses on moisture
736 gradients, as seen in Grad-CAM activations over regions with high
737 total column water vapor content. This suggests that it detects 738
patterns related to moisture transport and non-adiabatic processes,
739 which influence subseasonal temperature variability in Europe.

This observation aligns with established meteorological principles. The interaction between weather patterns and moisture transport plays a key role in subseasonal weather evolution in the European climate system [24]. Although the model can identify moisture-rich areas without direct instruction, these findings should be seen as a step toward improving deep learning-based SSF models, rather than proof that the model fully captures climate dynamics.

5.3 Feature Importance Analysis

Another finding is that feature importance analysis ranks input variables and time steps based on their impact on model predictions. The temporal feature importance scores (Figure 9) confirm the expected recency effect. Time steps closer to the forecast initialization have the greatest influence on predictions. This aligns with physical principles, as recent atmospheric conditions strongly shape short-term weather patterns [23]. However, Earlier time steps also play a role, though with less impact. This suggests that the ConvLSTM2D model effectively uses the LSTM memory to retain and integrate long-term historical information. The model appears to capture slow climate oscillations and decadal variability, both of which affect subseasonal predictability. This ability to process information across different time scales is crucial. It enhances the models potential for subseasonal forecasting, especially in a chang- ing climate where decadal variability can shift forecast baselines.

Furthermore, a finding in the feature channel importance analysis, shown in Fig. 10, highlights sea level pressure and 2 m temperature as the key input variables for the subseasonal temperature forecasts in Europe. The strong influence of sea level pressure underscores the model's reliance on large-scale atmospheric circulation patterns. This aligns with established science, reinforcing the understand- ing that the NAO and other pressure-driven teleconnections are dominant drivers of European weather and climate [17]. The high importance of 2-meter temperature is partly due to the natural auto- correlation within the temperature field. However, it also suggests that the model effectively learns to infer and refine temperature trends directly from the input data. This demonstrates its ability to capture the temporal dynamics and persistence of temperature patterns. Additionally, the consistent prominence of total column water vapor highlights the crucial role of moisture-related processes in SSFs. This finding supports the insights from Grad-CAM and aligns with the known climatology of moisture transport, la- tent heat release, and cloud radiation effects in shaping European weather systems [26].

785 Overall, XAI insights from Grad-CAM visualization and feature 786 significance analysis help explain how the ConvLSTM2D model extracts and uses meteorologically important information from complex climate data. While the model's performance is still modest, its focus on weather-scale systems, humidity gradients, and key climate variables—such as sea level pressure and near-surface temperature—adds scientific credibility to this approach. These findings highlight the potential of deep learning, combined with interpretable AI, to improve climate forecasting and deepen our understanding of climate system behavior. XAI is more than just a tool for post-hoc interpretation. It plays an active role in AI-driven scientific discovery, making it a valuable asset in climate research.

5.4 Limitations and Challenges

While ERA5 reanalysis data provide a high-quality, spatially, and temporally consistent dataset, they also have limitations. Despite its relatively high spatial resolution compared to historical datasets, ERAS5 cannot fully capture fine-scale meteorological features that may influence subseasonal forecasts. This is especially evident in areas with complex topography, such as mountains, or regions with intricate land-sea interactions, like coastal areas and islands. Even though it is valuable for training and model evaluation, it still contains biases and uncertainties. These arise from the data assimilation process and the observational network, potentially introducing subtle errors into learned representations and predic- tions in machine learning models. Another challenge is the size and time span of training datasets, which may be limiting for deep learning models, especially those designed to capture complex spa- tiotemporal dynamics. Models with millions of parameters, such as ConvLSTM2D, often require vast amounts of data to distinguish variations in the underlying distribution and generalize beyond the training domains.

The ConvLSTM2D model struggles to accurately predict extreme temperature events, as shown in Figure 4. It tends to underestimate peak values and overestimate moderate ones, leading to poor esti- mates of rare but impactful weather anomalies. This reduces the model's reliability in predicting extreme conditions. To improve performance, future research could explore specialized loss func- tions that give more weight to extreme values, such as quantile regression loss or weighted MSE for extreme events. Additionally, synthetic oversampling techniques like SMOTE could help increase the representation of extreme events in training data, allowing the model to learn from rare occurrences. Finally, incorporating physical constraints into the model could ensure it aligns with known meteorological principles, potentially improving its ability to capture extreme conditions. Addressing these limitations directly would create opportunities to enhance the model's predictive power for high-impact temperature anomalies.

The XAI methods used in this study provide valuable insights into model behavior but also have methodological limitations. Grad-CAM visualizations, while visually informative, are qualitative. They offer heuristics rather than mathematically rigorous explanations of model behavior. Heat maps highlight areas of attention, but the exact quantitative contribution of each region to the final prediction remains unclear. Feature importance analyses, though providing quantitative rankings, are sensitive to the masking or

842 perturbation methods used. Different perturbation strategies can 843 lead to slight variations in rankings, and the analysis may not fully 844 capture complex, nonlinear interactions between variables or time 845 steps. Interpreting XAI results in deep learning models remains 846 challenging, especially within the highly complex and nonlinear 847 dynamics of the climate system. A careful and cautious approach is 848 necessary, considering both the limitations of the models and the 849 interpretability techniques used to analyze their behavior.

5.5 Implications and Future Directions

Despite these limitations, this study presents a strong methodological framework and a promising foundation for future research on improving European subseasonal forecasting. This application of Grad-CAM and feature importance analysis to the ConvLSTM2D model highlights the potential of XAI to reveal the inner work- ings of deep learning models in climate science. The integration of deep learning and XAI offers valuable opportunities for model diagnostics, targeted improvements, and, ultimately, advancements in climate forecasting and understanding of the climate system [27].

Future research could focus on improving forecasting of the ConvLSTM2D model and other advanced deep learning architectures. More powerful models, such as Transformer networks, which excel in sequence-to-sequence tasks with built-in attention mechanisms, or attention-enhanced LSTM variants, could better capture remote dependencies and complex spatiotemporal interactions in climate data [18]. Optimizing model hyperparameters through automated techniques like Bayesian optimization or evolutionary algorithms could further improve prediction accuracy and model robustness. Additionally, integrating multiple forecasting models could help reduce uncertainty and enhance the reliability of subseasonal predictions.

Improvements in data preprocessing and feature engineering could also enhance model performance. Incorporating established cli- mate indices, such as the NAO and AMO, as explicit input features may help the model better understand large-scale climate patterns and decadal variability affecting European weather. Exploring al- ternative dimensionality reduction techniques beyond PCA or in- tegrating higher resolution climate datasets, could further refine predictions and improve regional forecasts. Additionally, leveraging XAI techniques not just for post-hoc model interpretation but also for proactive model refinement offers great potential. XAI insights could guide feature selection, inform the design of physically con- strained loss functions, and even support the development of hybrid modeling approaches that better capture the underlying physics of the climate system.

Future research could focus on enhancing the model's ability to 890 predict extreme weather events while optimizing its performance across different lead times. This includes assessing its capability 892 to capture extreme temperature and precipitation events (eg, top 893 5% values) and comparing its accuracy under extreme and average 894 conditions to evaluate its usefulness for disaster preparedness and risk management. Additionally, while this study primarily exam- 896 ines a 28-day lead time, future work should assess performance at different lead times (e.g., 7, 14, and 21 days) to understand how SSFs changes over time and enable targeted optimization. Integrating these aspects could improve the model's reliability, making it more effective for both routine and high-impact weather forecasting.

6 CONCLUSION

This study explore the impact of a decade of climate change on subseasonal weather forecasts in Europe by applying XAI to a deep learning model. This approach improved transparency and pro- vided deeper insights into the model's internal mechanisms.

This study combines the ConvLSTM2D architecture with Grad-CAM visualization and feature importance analysis to assess ma- chine learning in weather forecasting. It highlights both the poten- tial and challenges of using advanced models in this field. Using XAI techniques, the study evaluates the predictive performance and interpretability of the ConvLSTM2D model. While the model does not yet achieve state-of-the-art accuracy, the methodological progress remains significant. Beyond basic input-output relation- ships, the study provides insights into the model's internal pro- cesses, revealing signs of emerging "climate intelligence" in deep learning systems.

Grad-CAM visualizations reveal that the model can identify spatially coherent and meteorologically relevant patterns in climate data. These patterns align with known atmospheric dynamics, such as cyclone structures, frontal systems, and moisture gradients. This improves the physical interpretability of the model, making its pre- dictions more scientifically meaningful. Feature importance analy- sis further quantifies the influence of key meteorological variables like sea level pressure and 2-meter temperature on forecasts. It also highlights the model's ability to use long-term temporal dependen- cies, suggesting potential for capturing extended climate signals in subseasonal forecasting.

These findings have important implications for subseasonal weather forecasting in Europe and AI applications in climate science. This study demonstrates that integrating XAI techniques into deep learn- ing models enhances scientific transparency and credibility. It pro- vides useful diagnostic insights into the ConvLSTM2D model's limitations, such as insufficient forecast dispersion and challenges in predicting extreme weather. These insights can guide improve- ments in model architecture, training strategies, and feature selec- tion. Future research should build on these findings to develop more advanced models. Incorporating established climate indices and teleconnection patterns could further improve forecast accuracy.

REFERENCES

[1] Marybeth C Arcodia, Elizabeth A Barnes, Kirsten J Mayer, Jiwoo Lee, Ana Ordonez, and Min-Seop Ahn. 2023. Assessing decadal variability of subseasonal forecasts of opportunity using explainable AI. Environmental Research: Climate 2, 4 (2023), 045002.

[2] Martin Beniston, David B Stephenson, Ole B Christensen, et al. 2007. Future extreme events in European climate: an exploration of regional climate model projections. Climatic change 81, 1 (2007), 71-95. https://doi.org/10.1007/s10584-006-9226-z

[3] Jasmin Praful Bharadiya. 2023. Exploring the use of recurrent neural networks for time series forecasting. International Journal of Innovative Science and Research Technology 8, 5 (2023), 2023-2027.

Ileana Bladé, Brant Liebmann, Didac Fortuny, and Geert Jan van Oldenborgh. 2012. 1031 Observed and simulated impacts of the summer NAO in Europe: implications 1032 for projected drying in the Mediterranean region. Climate dynamics 39, 3 (2012), 1033 709-727. https://doi.org/10.1007/500382-011-1182-8 1034 Keith M Brander, Ute Daewel, Ken F Drinkwater, et al. 2010. Cod and future 1035 climate change. ICES Cooperative Research Reports (CRR) (2010). 1036 H. Lee Core Writing Team and J. (eds.) Romero. 2023. Climate Change 2023: 1037 Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assess- 1038 ment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, 1039 Switzerland (2023), 35-115. https://doi.org/10.59327/IPCC/AR6-9789291691647 1040 Pratyusha Das and Antonio Ortega. 2022. Gradient-weighted class activation 1041 mapping for spatio temporal graph convolutional network. In ICASSP 2022-2022 1042 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 1043 IEEE, 4043-4047. 1044 Catherine O de Burgh-Day and Tennessee Leeuwenburg. 2023. Machine learning 1045 for numerical weather and climate modelling: a review. Geoscientific Model 1046 Development 16, 22 (2023), 6433-6477. https://doi.org/10.5194/gmd-16-6433-2023 1047 Bradley J Erickson and Felipe Kitamura. 2021. Magician's corner: 9. Performance 1048 metrics for machine learning models. 200126 pages. 1049 , 10 Tobias Sebastian Finn. 2021. Self-attentive ensemble transformer: Representing 1050 ensemble interactions in neural networks for earth system models. arXiv preprint 1051 arXiv:2106. 13924 (2021). 1052 11 European Centre for Medium-Range Weather Forecasts. 2024. ERA5 Reanalysis 1053 Dataset. European Centre for Medium-Range Weather Forecasts (ECMWF) (2024). 1054 12 Fatemeh Ghobadi and Doosun Kang. 2022. Improving long-term streamflow 1055 prediction in a poorly gauged basin using geo-spatiotemporal mesoscale data 1056 and attention-based deep learning: A comparative study. Journal of Hydrology 1057 615 (2022), 128608. https://doi.org/10.1016/j.jhydrol.2022.128608 1058 13 Yoo-Geun Ham, Jin-Ho Kim, Jae-Heung Park, et al. 2019. Deep learning for 1059 multi-year ENSO forecasts. Nature communications 10, 1 (2019), 1-12. https: 1060 //doi.org/10.1038/s41467-019-10202-1 1061 14 Tristan Hauser, Entcho Demirov, Jieshun Zhu, and Igor Yashayaev. 2015. North Atlantic atmospheric and ocean inter-annual variability over the past fifty years- Dominant patterns and decadal shifts. Progress in Oceanography 132 (2015), 197-219. https://doi.org/10.1016/j.pocean.2015.01.001 15 Sijie He, Xinyan Li, Timothy DelSole, Pradeep Ravikumar, and Arindam Banerjee. L 2021. Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 169-177. 16 Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural computation 9, 8 (1997), 1735-1780. https://doi.org/10.1162/neco.1997.9.8. 1735 17 Thomas Huld and Irene Pinedo Pascua. 2015. of 2-meter air L Spatial downscaling temperature using operational forecast data. Energies 8, 4 (2015), 2381-2411. 18 Wei Jin, Wei Zhang, Jie Hu, Jiazhen Chen, Bin Weng, Jianyun Gao, and Tiangiang Huang. 2023. Transformer for sub-seasonal extreme high temperature probabilis- tic forecasting over eastern China. Theoretical and Applied Climatology 151, 1 (2023), 65-80. 19 Arun Kumar. 2009. Finite samples and uncertainty estimates for skill measures L for seasonal prediction. Monthly Weather Review 137, 8 (2009), 2622-2631. 20 Antonios Mamalakis, Imme and Elizabeth A Barnes. 2020. L Ebert-Uphoff, Explain- able artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. (2020), 315-339. 21 Alexander Marusov, Vsevolod Grabar, Yury Maximov, et al. 2024. Long-term L drought prediction using deep neural networks based on geospatial weather data. Environmental Modelling & Software 179 (2024), 106127. https://doi.org/10.1016/j.envsoft.2024.106127 22 Mateusz Norel, Michal Katczyúski, Iwona Pifskwar, Krzysztof Krawiec, and Zbig- niew W Kundzewicz. 2021. Climate variability indices—a guided tour. Geosciences 11, 3 (2021), 128. https://doi.org/10.3390/geosciences11030128 23 National Academies of Sciences, Division on Earth, Life Studies, Board on Atmo- L spheric Sciences, Committee on Extreme Weather Events, and Climate Change Attribution. 2016. Attribution of extreme weather events in the context of climate change. National Academies Press. 24 René Orth and Sonia I Seneviratne. 2014. Using soil moisture forecasts for sub-seasonal summer temperature predictions in Europe. Climate dynamics 43 (2014), 3403-3418. 25 Emmanuel Rouges, Laura Ferranti, Holger Kantz, and Florian Pappenberger. 2024. Pattern-based forecasting enhances the prediction skill of European heatwaves into the sub-seasonal range. Climate Dynamics (2024), 1-17. https://doi.org/10. 1007/s00382-024-07174-4 26 Yalu Ru and Xuejuan Ren. 2024. Subseasonal variability of sea level pressure and L its influence on snowpack over mid-high-latitude Eurasia during boreal winter. Climate Dynamics 62, 8 (2024), 8299-8318. 27 Emrullah SAHIN, Naciye Nur Arslan, and Durmus Özdemir. 2024. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. Neural Computing and Applications (2024), 1-107. 28 Abhishek V Tatachar. 2021. Comparative assessment of regression models based on model evaluation metrics. International Journal of Innovative Technology and Exploring Engineering 8, 9 (2021), 853-860. 29 F Vitart, AW Robertson, S2S Steering Group, et al. 2015. Sub-seasonal to seasonal prediction: linking weather and climate. Seamless prediction of the earth system: From minutes to months (2015), 385-401. 30 Frédéric Vitart and Andrew W Robertson. 2018. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. npj climate and atmospheric science 1, 1 (2018), 3. https://doi.org/10.1038/s41612-018-0008-3 31 Jonathan A Weyn, Dale R Durran, Rich Caruana, and Nathaniel Cresswell-Clay. 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. Journal of Advances in Modeling Earth Systems 13, 7 (2021), e2021MS002502. https://doi.org/10.1029/2 32 Christopher J White, Daniela I V Domeisen, Nachiketa Acharya, et al. 2022 Advances in the application and utility of subseasonal-to-seasonal predictions Bulletin of the American Meteorological Society 103, 6 (2022), E1448-E1472. https: //doi.org/10.1175/BAMS-D-21-0117.1 33 Cort J Willmott and Kenji Matsuura. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30, 1 (2005), 79-82. 34 Ruyi Yang, Jingyu Hu, Zihao Li, et al. 2024. Interpretable machine learning for weather and climate prediction: A review. Atmospheric Environment (2024), 120797. https://doi.org/10.1016/j.atmosenv.2024.120797 35 Gaohong Yin, Takao Yoshikane, Ryo Kaneko, and Kei Yoshimura. 2023. Improving Global Subseasonal to Seasonal Precipitation Forecasts Using a Support Vector Machine-Based Method. Journal of Geophysical Research: Atmospheres 128, 17 (2023), e2023JD038929. 36 Lujun Zhang, Tiantian Yang, Shang Gao, Yang Hong, Qin Zhang, Xin Wen, and Chuntian Cheng. 2023. Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous US using machine learning models. Atmospheric Research 281 (2023), 106502.

va Appendix A MORE EDA 1063
A1 Extremes

Figure 13: Frequency of High Temperature Extremes. This plot shows the time series of the frequency of high temperature extremes over the years, derived from the ERA5 dataset. It helps visualize the temporal trends in the occurrence of unusually high temperatures.

Figure 14: Temperature - DJF. This spatial map represents the average temperature pattern during Winter (December-January-February) across the region of interest, based on the ERAS dataset.

Figure 15: Mean Temperature. This spatial map displays the average temperature distribution across the region of interest calculated from the ERAS dataset over the considered time period

Figure 16: Temperature Time Series. This time series plot shows the variation of 2-meter temperature over the years, providing a temporal overview of temperature changes in the ERAS5 dataset.