BAMS Article

Setting the Standard: Recommended Practices for Data Preprocessing in Data-Driven Climate Prediction

Jason C. Furtado,[a] Maria J. Molina,[b] Marybeth C. Arcodia,[c] Weston Anderson,[b] Tom Beucler,[d] John A. Callahan,[e,f] Laura M. Ciasto,[g] Vittorio A. Gensini,[h] Michelle L'Heureux,[g] Kathleen Pegion,[a] Jhayron S. Pérez-Carrasquilla,[b] Maike Sonnewald,[i] Ken Takahashi,[j] Baoqiang Xiang,[k,l] and Brian G. Zimmerman[m]

KEYWORDS: Statistical techniques; Climate prediction; Artificial intelligence; Machine learning

ABSTRACT: Artificial intelligence (AI)—and specifically machine learning (ML)—applications for climate prediction across time scales are proliferating quickly. The emergence of these methods prompts a revisit to the impact of data preprocessing, a topic familiar to the climate community, as more traditional statistical models work with relatively small sample sizes. Indeed, the skill and confidence in the forecasts produced by data-driven models are directly influenced by the quality of the datasets and how they are treated during model development, yielding the colloquialism, "garbage in, garbage out." As such, this article establishes protocols for the proper preprocessing of input data for AI/ML models designed for climate prediction (i.e., subseasonal-to-decadal and longer time scales). The three aims are to 1) educate researchers, developers, and end users on the effects that data preprocessing has on climate prediction; 2) provide recommended practices for data preprocessing for such applications; and 3) empower end users to decipher whether the models they are using are properly designed for their objectives. Specific topics covered include creating (standardized) anomalies, dealing with nonstationarity and the spatiotemporally correlated nature of climate data, and handling of extreme values and variables with potentially complex distributions. Case studies will illustrate how using different preprocessing techniques can produce different predictions from the same model, which can create confusion and decrease confidence in the overall process. Ultimately, implementing the recommended practices set forth in this article will enhance the robustness and transparency of AI/ML in climate prediction studies.

DOI: 10.1175/BAMS-D-24-0292.1
Corresponding author: Jason C. Furtado, jfurtado@ou.edu
Arcodia's current affiliation: University of Miami Rosenstiel School of Marine, Atmospheric, and Earth Science Department of Atmospheric Sciences and the Frost Institute for Data Science and Computing, Miami, Florida.
Jason C. Furtado, Maria J. Molina, and Marybeth C. Arcodia contributed equally to this work.
Manuscript received 18 November 2024, in final form 18 March 2026, accepted 27 March 2026
© 2026 American Meteorological Society. This published article is licensed under the terms of the default AMS reuse license. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

SIGNIFICANCE STATEMENT: With the rapid expansion of artificial intelligence (AI) in atmospheric science, the need for high-quality, properly prepared data for input into AI/ML models is important. In this article, we offer several recommended steps to properly preprocess input data for AI models used for climate predictions (i.e., time scale ranging from few weeks to many years). Among other topics, we discuss appropriate ways to calculate departures (or anomalies) from data that vary in time and space, how to handle large trends, and what to do with extreme values. We then conduct two case studies to illustrate how using different techniques for preprocessing can produce different predictions from the same model. Ultimately, following these recommendations will help make such studies more transparent, reproducible, and trustworthy.

AFFILIATIONS: [a] University of Oklahoma, Norman, Oklahoma; [b] University of Maryland, College Park, College Park, Maryland; [c] Colorado State University, Fort Collins, Colorado; [d] University of Lausanne, Lausanne, Switzerland; [e] Ocean Associates, Inc., Arlington, Virginia; [f] NOAA/National Ocean Service, Silver Spring, Maryland; [g] NOAA/NWS/NCEP/Climate Prediction Center, College Park, Maryland; [h] Northern Illinois University, Dekalb, Illinois; [i] University of California, Davis, Davis, California; [j] Instituto Geofísico del Perú, Lima, Perú; [k] NOAA/GFDL, Princeton, New Jersey; [l] Cooperative Programs for the Advancement of Earth System Science, UCAR, Boulder, Colorado; [m] Macquarie, Houston, Texas

1. Introduction

The integration of artificial intelligence (AI) and machine learning (ML) in weather and climate science is rapidly revolutionizing predictions and our understanding of Earth climate system. Such techniques offer increased capabilities in handling large datasets, identifying complex patterns, and making accurate predictions and projections. Recent attention has heavily centered on model choice (i.e., selecting which type of ML model is most appropriate; Dueben and Bauer 2018; de Burgh-Day and Leeuwenburg 2023; Molina et al. 2023b) and explainability of the predictions (e.g., Mamalakis et al. 2022, 2023; Yik et al. 2023; Camps-Valls et al. 2025). However, the effectiveness of data-driven models strongly depends on data quality. This consideration is paramount and led to the popular adage "garbage in, garbage out," credited to computer programmers in the late 1950s (Lidwell et al. 2003). Simply put, if flawed or poor-quality data are fed into a model, the resulting predictions will likely also be flawed.

Data quality can be evaluated in multiple ways. One method is based on sample size and the fidelity of the data. These considerations are important for AI/ML applications and have been addressed in several works across disciplines (e.g., Dueben et al. 2022; de Burgh-Day and Leeuwenburg 2023; Zantvoort et al. 2024; Xie et al. 2025). For the atmospheric sciences, so-called benchmark platforms that provide datasets ready for AI/ML applications have been developed—e.g., WeatherBench (Rasp et al. 2020, 2024), AQ-Bench (Betancourt et al. 2021), ClimateBench (Watson-Parris et al. 2022), ClimSim (Yu et al. 2024), and ChaosBench (Nathaniel et al. 2024). However, even if one has high-quality data with sufficient samples, a data-driven model may still produce poor results, as "garbage" can result from other erroneous assumptions or treatments of the input data, e.g., wrong assumptions around data biases, incorrect assessment of the "true" distribution, incorrect classification labels, and inconsistent thresholds and definitions of phenomena.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

Climate data present unique challenges for use in data-driven models. Most climate datasets are inherently spatiotemporal and sparse and possess spatial and temporal autocorrelations. The data are often nonstationary, especially in recent decades, due to anthropogenic climate change. This effect changes the core statistics of the variables of interest (e.g., temperature, wind, cloud cover, geopotential height, sea level rise). Studies handle nonstationarity in different ways. For example, when removing trends from a time series, some studies may simply remove a linear or second-order polynomial trend (e.g., Long et al. 2025), while others may use more complex techniques, such as empirical mode decomposition (e.g., Huang et al. 1998), to detect changing trends over time. Along with changing background statistics, climate variables also have varying distributions, many of which are nonnormal (e.g., gamma, bimodal, lognormal, and skew normal), and exhibit nonlinear interactions among themselves. As such, many traditional methods in statistics and AI/ML cannot simply be used "out of the box" when working with climate variables. Additionally, climate data are collected from diverse sources, including satellite observations, weather stations, and climate models, and can be noisy, incomplete, and heterogeneous. The rapid development of AI/ML applications in climate prediction—i.e., forecasting the state of the climate system (in probabilities) on time scales ranging from several weeks to a couple of decades in the future—necessitates a guiding set of recommended preprocessing steps for climate data to secure some degree of harmony between studies and applications. As such, understanding the rationale for why certain data preprocessing steps are taken can improve trust in these methods for end users by demystifying the process and providing ways to evaluate and critique the models and their results.

The aim of this paper is to present recommended practices on proper data preprocessing steps aimed for climate prediction studies using AI/ML. This paper has three overarching goals:

1) educate researchers and end users alike on different effects that dataset preprocessing can have on climate prediction across time scales (i.e., subseasonal to decadal and longer);
2) provide researchers with recommended practices for dataset preparation in such studies; and
3) empower end users to determine whether the models they are using are properly designed for their objectives, which can enhance the trustworthiness of AI/ML.

2. Recommendations for initial steps in data preparation and minimizing data leakage

The initial step in climate prediction is to clearly identify what the researcher wants to predict, over which time scale(s) to make this prediction, and what AI/ML methods are most appropriate for addressing the prediction problem. This step is essential and should be carefully considered—the reader is referred to other works for discussion of problem setup and different AI/ML methods used in prediction studies (e.g., Molina et al. 2023b; Yang et al. 2024; Camps-Valls et al. 2025). Upon making these decisions, the next step is to select a set of potential input features (i.e., predictors) and outputs used to generate a prediction or classification with that AI/ML model during training or testing. Most prediction problems will use a supervised learning framework requiring inputs and outputs; unsupervised learning applications only need inputs. Inputs and outputs can take the form of a numerical value (e.g., the Niño-3.4 index), numerical data fields (e.g., sea surface temperature), categories (e.g., an El Niño or La Niña event), or a probability. More about the terminology above can be found in Chase et al. (2022).

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

We recommend a period of "data exploration" first to identify key first-order statistics of the input features and locate any missing or erroneous data. As mentioned, climate variables can possess autocorrelation and covariances between them in space and time. Quantifying these relationships allows us to identify the effective sample size (N_{eff}), which can often be smaller than the total number of samples (N; e.g., Bretherton et al. 1999). Recognizing this concept is important, as it will impact the choice for the number of training samples needed. Sparse data (e.g., weather stations) introduce additional challenges, such as potential sampling biases. Interpolation or imputation can be used to fill undersampled data, but care must be taken with the choice of method (e.g., bilinear vs piecewise constant). Finally, understanding the distributions and trends of the variables will also inform the preprocessing steps needed (see section 3).

After initial data exploration, one should establish a representative training dataset, which will be used to "fit" the model's parameters (e.g., weights or coefficients). In choosing the number of training samples needed, one should consider (i) model complexity, (ii) available computational resources, (iii) number of total input features and their ability to represent a range of possible outcomes, and (iv) required/or desired accuracy. Having many input features, however, often requires a larger training sample size for the given model. As such, input features can be reduced using filtering methods (i.e., eliminating features with little-to-no statistical relationship with the target) or feature extraction (i.e., dimensionality reduction using, for example, principal component analysis). We advise considering one or more of these methods for feature selection for one's AI/ML problem [see Cai et al. (2018) and references therein for a variety of techniques for feature selection].

Thereafter, data "splits" need to be decided. For this, the datasets are typically divided into three distinct subsets: 1) training (used for fitting the model), 2) validation (used for adjusting configurable components of the model, known as hyperparameters), and 3) testing (used for final model evaluation). Commonly used ratios for training, validation, and testing sets include 60:20:20 or 80:10:10. Since the validation dataset is used during model development, it should not be considered in assessing test results. Doing so contributes to data leakage (discussed below), potential data misuse, and even unethical use of AI/ML. We further recommend that results for the training, validation, and testing datasets be openly reported to assess, for example, the generalizability properties of the model (e.g., how well the model will perform with the same predictors at a future time). We encourage the use of cross-validation (CV) methods to mitigate learned bias as well as using the full training and validation sets for more robust hyperparameter tuning (e.g., Sweet et al. 2023). The k-fold cross validation is the most commonly used cross-validation method in practice; Table 1 shows other CV methods, along with their limitations.

As mentioned, improper splitting choices of the data can inadvertently lead to data leakage, resulting in artificially inflated performance metrics and poor generalization of the results. Leakage of training or testing data occurs when input features contain or are derived from information that can reveal the target variable and would not be (feasibly) available during real-time predictions (i.e., taking information from the future, which one would not have access to at the time of prediction). Data leakage is especially problematic in climate prediction, where variables often possess strong temporal autocorrelation (i.e., "memory"), quasi periodicity, and/or long-term trends. Figure 1a illustrates an example of data leakage during detrending of a global (latitude weighted) time series of the monthly average of daily maximum temperatures. The use of the full-time period for detrending (1940–2024; Fig. 1a, red line) versus the training period (1940–2010; Fig. 1a, black line) results in much cooler ground-truth (i.e., target) temperatures during the (hypothetical) test period (2011–24), which can potentially bias the AI/ML model. The choice of a linear (first-degree) or quadratic (second-degree) polynomial for detrending can also have notable effects, where the

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

Table 1. CV methods relevant to weather and climate applications, including example use cases, descriptions, and associated limitations. Types of CV can be combined when the use case has several data properties to consider. In the table, i.i.d. stands for "independent and identically distributed."

Type of CV	Use case(s) and description	Limitation(s)
Standard *k* fold	General data (i.i.d.). Equal-sized random splits.	Leakage is possible (e.g., if there is spatial or temporal autocorrelation).
Stratified *k* fold	Imbalanced data. Retains class proportions in each fold.	Leakage is possible (e.g., if there is spatial or temporal autocorrelation).
Time series	Sequential data with temporal dependencies. Respects temporal order.	Assumes long-term stationarity and no spatial autocorrelation.
Spatial *k* fold	Spatial autocorrelation in data. Ensures spatially disjoint folds (e.g., with clusters or distance)	Training data may become limited due to spatial constraints. Sensitive to disjoint definition.
Spatial block	Irregularly sampled or sparse spatial data. Data are divided into independent blocks.	Sensitive to block size and placement.
Leave-one-out	Suitable for small data. Each sample serves as validation once.	Computationally expensive for large data. Does not address temporal dependencies.
Monte Carlo	Flexible and randomly repeated splits for general data with specified fold ratios.	Does not address spatial or temporal dependencies. Allows overlapping between folds.

second-degree polynomial minimizes data leakage effects during the (hypothetical) test period and removes low-frequency artifacts in the training period (Figs. 1a,b). Data leakage can also occur during the scaling and normalizing of the input features. Considering temporal autocorrelation in climate data, we recommend block splitting; i.e., the dataset is split into temporally continuous training/validation and testing blocks or periods (e.g., Fig. 1a), with a gap between the two blocks adequately long relative to the time scales of interest to increase independence (e.g., Zhu et al. 2023).

To prevent leakage, we recommend splitting data before any preprocessing steps. The testing data should be set aside and not used until model development is finalized. For instance, feature selection and calculation of scaling factors or trends should be made with the training/validation data after the split and later applied to the testing data. For data with strong autocorrelation, quasi periodicity, and/or trends, data splitting by obeying time series dependencies or phenomenon properties (e.g., choosing El Niño "years" based on season instead of calendar year) allows the model to "learn" about the evolution of the phenomena, thereby reducing the impact of autocorrelation between the training and testing datasets. If data splits do not consider the properties of the phenomenon, then learning relies more on autocorrelation, which by definition inflates skill metrics from the model. Since subseasonal and longer temporal dependencies can result in an insufficient N_{eff} to represent the processes of interest (e.g., Mayer et al. 2024), other techniques should be considered. For example, using data from long simulations of climate models to develop the initial versions of the AI/ML models and overcome the limitations of short observational records is becoming more common (e.g., Ham et al. 2019; Rivera Tello et al. 2023). This practice assumes that climate models adequately represent the phenomena of interest and that fine-tuning procedures can correct existing biases. To address spatial autocorrelation, one can stratify geographically distinct regions and build buffer zones between them. If regions are climatologically distinct and such properties are important for model performance, ensuring that data splits contain samples from respective regions while limiting temporal autocorrelation may be appropriate.

Note that, for operational applications where models are periodically refit or recalibrated as new data become available, rolling origin evaluation or other related approaches (e.g., Hewamalage et al. 2023) can be used during model development to expose the model to evolving climate states while preventing data leakage of future trends into the historical training data. However, even in operational contexts, one should ensure that the models

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

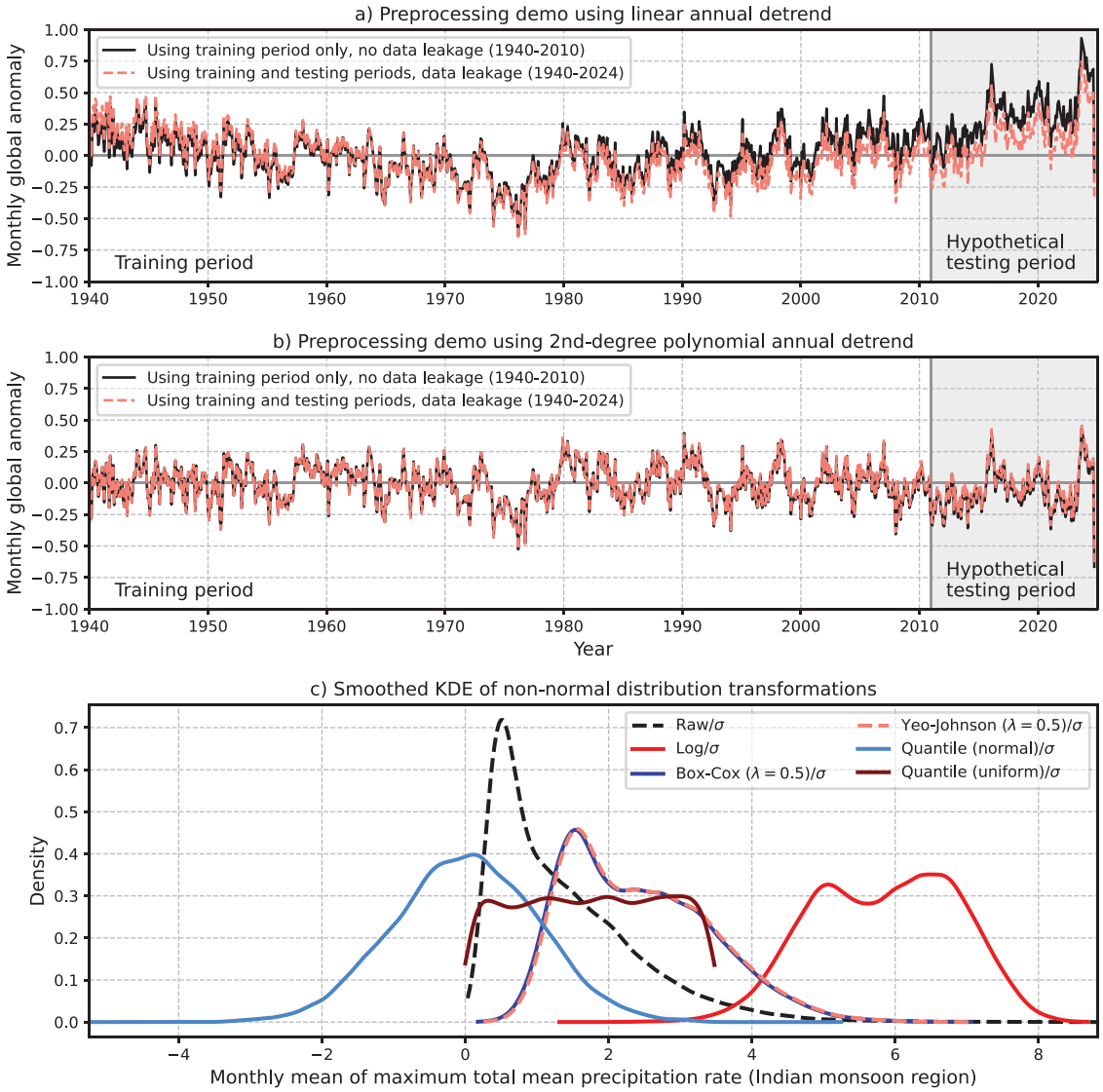

Fig. 1. Illustration of various time series preprocessing transformations using ERA5 (Hersbach et al. 2020). (a), (b) Preprocessing data leakage for global (latitude weighted) monthly mean of daily maximum temperature anomalies derived from a monthly climatology (1981–2010), where the black lines represent no data leakage due to detrending using only the training set period and the pink lines represent leakage due to detrending using the training and (hypothetical) testing periods. (c) Various kernel density estimations of data transformations for a nonnormal precipitation variable (black dashed line).

learn physical relationships in the climate system rather than simply fitting a long-term trend, which provides no skill for anomalies relative to the present climate.

3. Recommendations for preprocessing different types of input features

a. Working with numerical-valued data: Anomalies and standardization. Most problems in climate prediction involve predicting the anomaly (i.e., departure from an average) in a given field, allowing for skill evaluation beyond just the varying climatology (Hamill and Juras 2006). However, the computation of anomalies depends on the problem design, and the computation method can yield different interpretations. A baseline period may be fixed (e.g., 1991–2020) or centered rolling windows (e.g., ±15 years), the latter of which may address nonstationarity. Choosing a baseline period early in the record may result in artificial skill gains during evaluation due to the effects of climate change (Wulff et al. 2022). Choosing a base period later in the record, particularly one with a strong trend, will skew early values. Long-term trends should also be removed to avoid overinflating correlations between

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

input features or relationships derived therein. When calculating anomalies and trends, the spatial element of the climate variables should also be considered. These quantities can be computed using global, regional, or gridcell-specific climatology. A global or regional baseline climatology (latitudinally weighted if using gridded data) may be necessary when spatial anomaly patterns must be preserved (e.g., modes of variability). A gridcell-specific climatology accounts for local variations, which may be useful when interested in strong horizontal gradients (e.g., precipitation anomalies). Altogether, we recommend that there should not be a "one-size-fits-all" approach to anomaly calculation.

A special case exists when using hindcast simulations (i.e., forecasts made with a model for past events). Relatively short hindcast databases (e.g., 10–20 years long) restrict choices for climatologies, potentially leading to averages that use "future" periods and artificially inflating prediction skill (Risbey et al. 2021). Furthermore, hindcast climatologies are special because they are a function of the initialization date and lead time, so as to avoid model drift and bias. Therefore, when working with hindcast simulations, we suggest computing separate lead-time-based climatologies and applying smoothing windows to multiyear averages to address gaps in initialization dates (e.g., Pegion et al. 2019). For ensemble-based hindcasts, the above-suggested recommendations for anomaly and trend calculations work well for single-model analyses. When considering multimodel means, anomalies should be computed for each model separately.

Finally, feature scaling, either normalization (e.g., min–max scaling) or standardization (i.e., z scoring), of data inputs for AI/ML applications must be carefully done. Single fields with large magnitudes and/or multiple fields with varying magnitudes can cause instability and prevent convergence, requiring feature scaling. We recommend that users consistently scale all features while also considering model design when making such a choice. Feature scaling is unnecessary when using tree-based models (e.g., random forests) since they use feature thresholds in their structure but is especially important when using algorithms sensitive to variance, data containing outliers, and when using distance-based models (e.g., k-means clustering). Since outliers can disproportionately affect feature scaling, unless one is interested in predicting outlier events, we suggest either removing the outliers or winsorizing the distribution—i.e., outlier values are reassigned to a specified percentile of the data.

When dealing with variables that have non-Gaussian distributions, other preprocessing transformations can help align input features and should be performed before feature scaling. Figure 1c provides examples of different transformations done on the distribution of monthly mean total precipitation maxima in the Indian monsoon region, depending on user needs: (i) the log transformation (Fig. 1c, red line), (ii) the Box–Cox (Fig. 1c, solid blue line; Box and Cox 1964) or Yeo–Johnson (Fig. 1c, dashed red line; Yeo and Johnson 2000) transformations, and (iii) quantile transformations (Fig. 1c, light blue and solid red lines). Feature scaling should then be applied after these transformations.

b. Working with categorical inputs and outputs. Sometimes prediction problems require predicting a category or class (e.g., El Niño or La Niña). Several methods should be considered when working with categorical data. If the data have ordered ranks (e.g., labeling 2-m temperatures as below average, average, and above average), then integer encoding, where a unique integer value is assigned to the respective category (e.g., 0, 1, or 2), is appropriate. Classes representing unordered data (e.g., European weather regimes; Grams et al. 2017) should use one-hot encoding, whereby a single value in the vector is set to one (corresponding to the category), and the others are left as zero. For the seven European weather regimes, for example, a day labeled "Atlantic Ridge" may be encoded [0, 0, 1, 0, 0, 0, 0, 0] (the eighth category representing "no regime"; Grams et al. 2017). One-hot encoding prevents the AI/ML

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

model from interpreting the categories as having an ordinal relationship. Binary categorical problems (e.g., severe or nonsevere thunderstorms) can choose between integer or one-hot encoding. Missing values can be assigned as a separate category/class (e.g., "missing") before encoding, or if there are few missing values, they may be deleted.

Some climate prediction problems, particularly when working with extreme events, may have data with very few events and many "null" events—e.g., if classifying days with 2-m temperatures exceeds the 99th percentile, there will be 1 event for every 100 days, on average. In this example, the extreme heat days comprise a minority class compared to the nonextreme heat days. Hence, the user is faced with what is known as class imbalance, meaning there are very few "hits" for the AI/ML model on which to train and, thus, overall poor performance by the AL/ML model (Molina et al. 2023b). Since minority classes can be important in climate prediction studies, one should consider resampling techniques to address the imbalance. For example, the size of the minority class can be increased by randomly duplicating existing samples or using algorithms to create synthetic samples, known as data augmentation. One such data augmentation method is the synthetic minority oversampling technique (SMOTE; Chawla et al. 2002). Care should be taken when oversampling autocorrelated data; techniques for time series imputation such as time-sliced SMOTE can be used (Baumgartner et al. 2022). Undersampling of the majority class can also be employed to reduce the imbalance (e.g., Gensini et al. 2021; Rivera Tello et al. 2023). Class weights can also be applied, where larger magnitude weights can be assigned to the minority class for an added penalty in its erroneous classification during training. Importantly, the minority class should contain a diverse set of examples from which to learn. Thus, we recommend the input data to have a good N_{eff} as opposed to focusing on the total sample size.

4. Putting the recommendations into practice: Case studies

Figure 2 summarizes the overall workflow for an AI/ML problem in climate prediction, including our recommendations for initial data preparation. We have highlighted the ordering and importance of the preprocessing steps in this flowchart to serve as a template for scientists

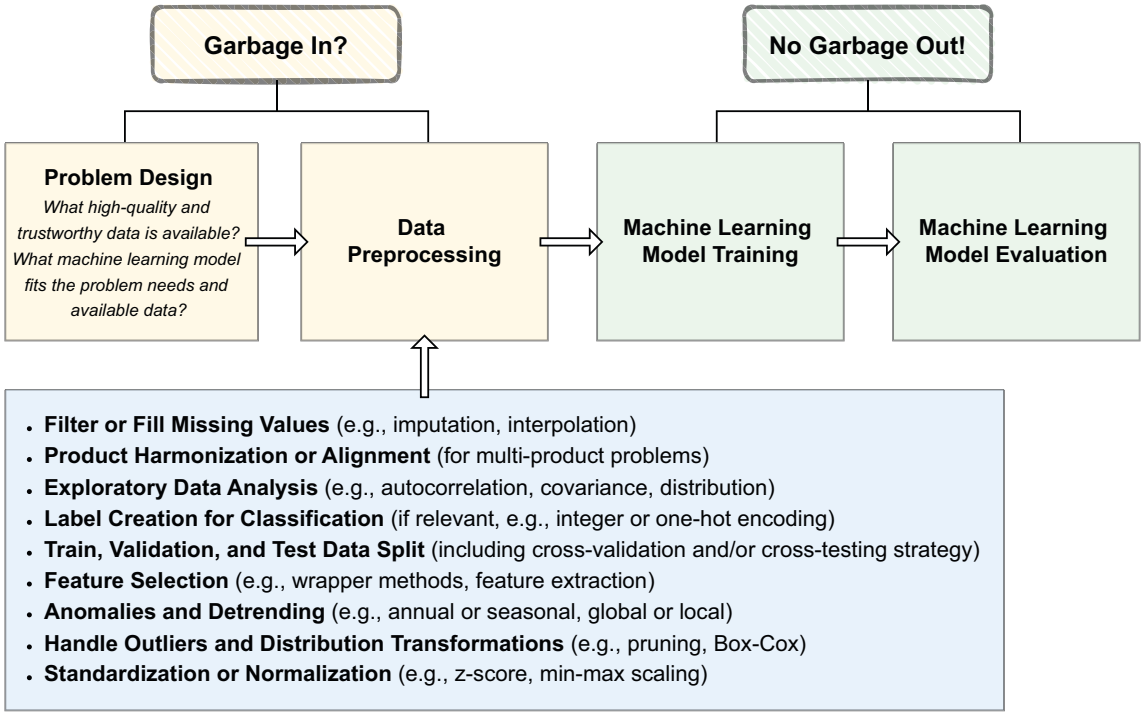

Fig. 2. Summary figure illustrating a typical AI/ML workflow, detailing the preprocessing steps for numerical and categorical data, presented in a recommended order. However, preprocessing steps (and their order) may be application specific.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

and practitioners in the field. To further emphasize the importance of these data preprocessing steps, we have designed two small case studies and offer the differences in interpretation and skill that would arise should these steps be followed or not.

a. Case study 1: Subseasonal weather regimes. Weather regimes are large-scale, persistent, and recurrent atmospheric patterns useful for subseasonal predictions due to their relationship with surface weather anomalies (e.g., Molina et al. 2023a). These regimes are often defined using 500-hPa geopotential height (Z500) anomalies over a specified domain (e.g., North America) and *k*-means clustering. Here, we demonstrate how seemingly minor differences in preprocessing choices can lead to a different number of preferred weather regimes with different spatial characteristics. To highlight this sensitivity, we will calculate the North American weather regimes in three different ways, detailed below. Daily mean Z500 comes from ERA5 (Hersbach et al. 2020) and is regridded to a 1° horizontal grid spacing using the nearest neighbor method, consistent with Molina et al. (2023a). Regridding methods and target grid spacing are additional methodological choices whose impact could be evaluated in future sensitivity tests. The analysis region is chosen as 20°–80°N, 180°–30°W, as in Lee et al. (2023). A daily climatology spanning 1940–2023 is established, utilizing a 60-day centered rolling window method. A 10-day low-pass filter is applied to the anomalies, and they are also detrended by subtracting a latitudinal-weighted third-degree polynomial fit for each day of the year. The data are then standardized using the 60-day centered running mean of the area-averaged Z500 anomaly standard deviation for each day of the year. Subsequent steps are detailed in Pérez-Carrasquilla and Molina (2025).

We conduct three experiments (Table 2) for this case study, each with 500 random centroid initializations to ensure robustness of resultant clusters. The "control" experiment aligns with the definitions and methods used by Lee et al. (2023) except for using the 1940–2023 climatological base period. "Experiment 1" involves standardizing the detrended anomalies using gridcell daily averaged standard deviation instead of using the domain-averaged standard deviation, as done in Lee et al. (2023). "Experiment 2" involves using the climatological period 1979–2023. All other preprocessing steps are consistent among the experiments. The optimal number of clusters for each experiment is found using the intercluster correlation, which is the correlation between the centroid coordinates. An intercluster correlation at zero or just below zero is preferred. The optimal number can vary depending on the chosen metric, and more than one metric can be used for robustness; however, here we focus on sensitivity to preprocessing choices, not the chosen metric(s). Regime names were assigned subjectively based on similarities in Z500 anomalies, though distance or similarity metrics (e.g., Pearson correlation) can be used for more robust cluster alignment across experiments.

Figure 3 summarizes the results of the experiments. The control experiment results in the six "preferred" North American weather regimes due to a relative "best" intercluster correlation value at $k = 6$ compared to other k values (Fig. 3a). In contrast, experiment 1 yields four regimes, whereas experiment 2 results in six preferred regimes (Fig. 3a). In experiment 1, local standardization alters the spatial patterns of the Greenland high (GH), Pacific Ridge (PR),

Table 2. Description of the North American weather regime experiments for case study 1.

Experiment name	Methodology	Number of regimes	Regime names
Control	Lee et al. (2023), but with 1940–2023 climatology and third-degree polynomial detrending.	6	AL, Alaskan Ridge (AR), North Atlantic high (NAH), PT, PR, and GH
Experiment 1	As in control, but with local standardization.	4	PT, CUSH, PR, and GH
Experiment 2	As in control, but with 1979–2023 climatology.	6	GH, CUSH, AR, PR, PT, and NAH

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

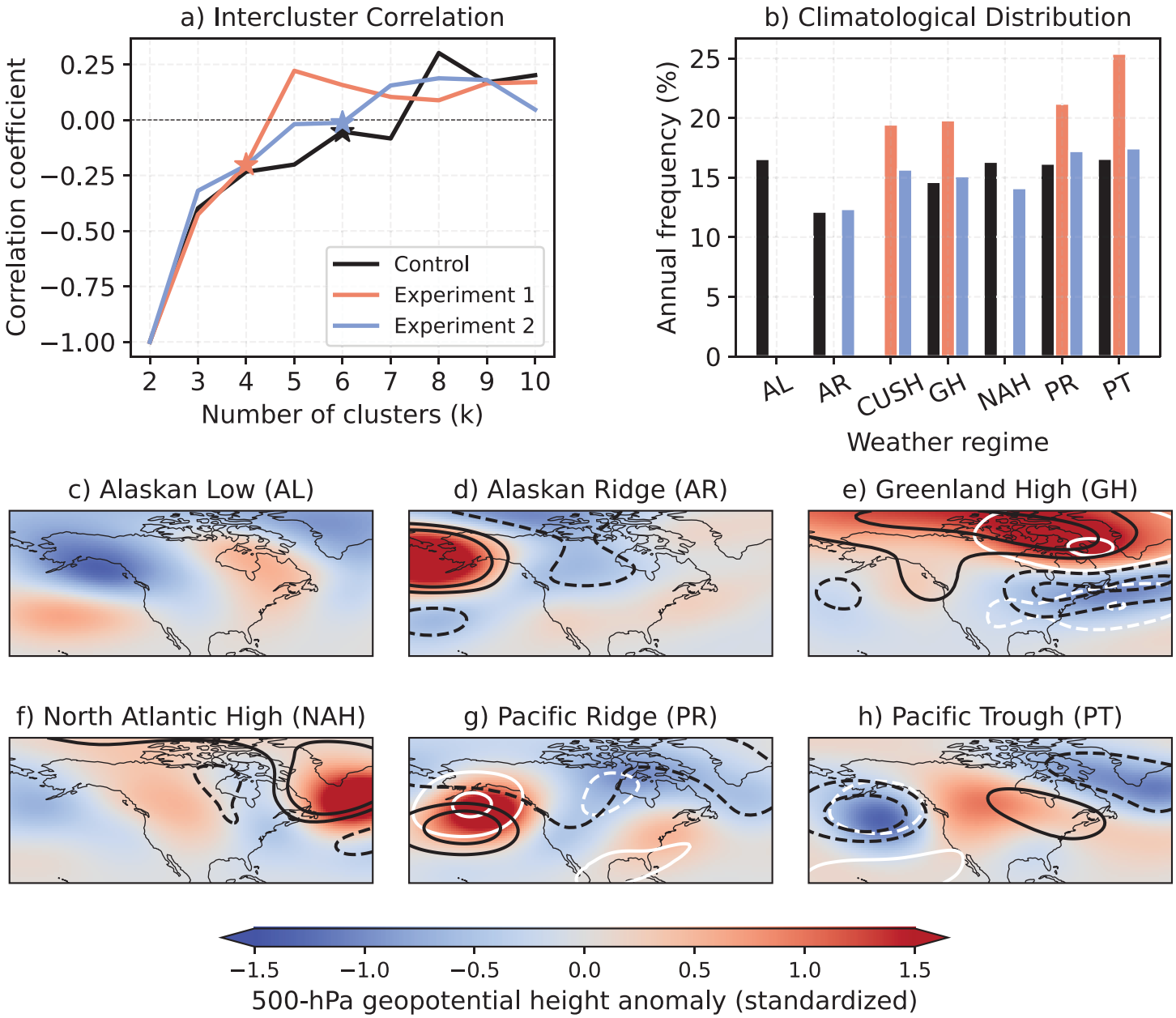

Fig. 3. (a) Intercluster correlation, with star markers indicating the "preferred" number of clusters, which keep the correlation between centroids at or just below zero. (b) The annual climatological frequency of weather regimes. The control and experiments in (a) and (b) are specified in the legend of (a). (c)–(h) Standardized Z500 anomalies for the control. When the respective regimes were identified in the experiments, contour lines were overlaid. White contour lines represent experiment 1, while black contour lines denote experiment 2. Solid lines indicate positive anomalies (+0.5 and +1.0 contour), and dashed lines represent negative anomalies (–0.5 and –1.0).

and Pacific Trough (PT) regimes (Figs. 3e,g,h). Local standardization also leads to the emergence of the central U.S. high regime (not shown), which was absent in the control experiment. In experiment 2, using a shorter climatology yields a clustering solution in which a distinct Alaskan low (AL) regime is no longer identified among the preferred regimes (Fig. 3b). The decision to use grid cell rather than regional or global standardization stems from the necessity to capture localized signals, potentially useful for subseasonal high-impact weather events. A shorter climatology may be preferred if polynomial detrending is ineffective in removing regional trends. A challenge with unsupervised learning is determining the "correct" final groupings, but the methodology used in the control experiment is preferred due to the need to capture large-scale patterns (regional standardization) and the intention to investigate trends in weather regimes (longer climatology).

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

b. Case study 2: Predicting temperature anomalies in the Southwest United States. Time series prediction is commonly performed in climate science, particularly for understanding and predicting climate patterns. Furthermore, many climate modes and large-scale processes are captured through time series indices, such as El Niño–Southern Oscillation (ENSO). Forecasting time series evolutions can aid in weather forecasting and prediction of extreme events, such as heat waves. To illustrate the impact on prediction skill from common preprocessing missteps, we evaluate a simple neural network regression model to predict the monthly temperature anomaly in the Southwest United States. Note that the neural network built for this case study is intended to illuminate various preprocessing effects on outcomes and not for actual prediction. For many applications, the observational record is often too short for developing such neural network models, as time series can exhibit strong temporal autocorrelation. Many weather and climate forecasting applications are developed using multiglobal climate model data (e.g., Ham et al. 2019; Rivera Tello et al. 2023), which increases N_{eff} by at least an order of magnitude. However, we have ensured that our model is not overfit to the training data via implementing early stopping (e.g., patience = 50, which indicates how many epochs to wait for a model's performance to improve before stopping training) and is sufficient in demonstrating preprocessing differences.

The case study is set up as follows. Monthly mean average temperatures are taken from the Berkeley Earth Surface Temperatures dataset (Rohde and Hausfather 2020) from 1900 to 2025 and averaged across the Southwest United States (29°–39°N; 104°–117°W). We input the current temperature anomaly and three lagged time steps of the time series to the neural network (3 layers, 100 nodes in the first layer, 50 nodes each in the second and third layers) to predict the Southwest U.S. temperature anomaly 1 month later. We focus on three specific preprocessing components: 1) the data split, 2) the climatological period, and 3) the detrending period. We construct several variants of the model (i.e., experiments), highlighting the use and misuse of each of these three preprocessing steps (Table 3). The `clean` preprocessing experiment represents the case in which the recommended preprocessing steps highlighted in this article are fully followed. The data split between the training and validation and validation and testing periods is an 18-month gap to reduce data leakage. Additionally, the climatology

Table 3. The time periods for each of the data splits and computations with a description for the five experiments used in case study 2 and shown in Fig. 4.

Experiment	Training	Validation	Testing	Climo	Trend	Description
`clean`	1 Jan 1900 to 1 Dec 1979	1 Jul 1981 to 1 Feb 2001	1 Sep 2002 to 1 Dec 2024	1 Jan 1941 to 1 Dec 1970	1 Jan 1900 to 1 Dec 1979	No data leakage; 18-month gap between splits
`trend`	1 Jan 1900 to 1 Dec 1979	1 Jul 1981 to 1 Feb 2001	1 Sep 2002 to 1 Dec 2024	1 Jan 1941 to 1 Dec 1970	1 Jan 1900 to 1 Dec 2024	Trend computed during the entire period
`climo`	1 Jan 1900 to 1 Dec 1979	1 Jul 1981 to 1 Feb 2001	1 Sep 2002 to 1 Dec 2024	1 Jan 1991 to 1 Dec 2020	1 Jan 1900 to 1 Dec 1979	Climatology computed during test period
`split`	1 Jan 1900 to 1 Dec 1999	1 Jan 2000 to 1 Aug 2002	1 Sep 2002 to 1 Dec 2024	1 Jan 1941 to 1 Dec 1970	1 Jan 1900 to 1 Dec 1979	Training/validation/test split on sequential months and during ongoing ENSO events
`split_trend_climo`	1 Jan 1900 to 1 Dec 1979	1 Jan 1980 to 1 Aug 2002	1 Sep 2002 to 1 Dec 2024	1 Jan 1991 to 1 Dec 2020	1 Jan 1900 to 1 Dec 2024	Trend and climatology computed during test period; split is sequential months

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

is computed over a 30-yr period from the middle of the training period from January 1941 to December 1970. The linear trend (0.0289°C decade^{-1}) is also computed over the full training period of January 1900–December 1979. Skill for all experiments is computed via mean absolute error (MAE) with lower MAE indicating higher skill. Although the neural network was trained to minimize mean-square error, we used mean absolute error to evaluate skill because it provides an interpretable measure of typical prediction error in physical units. We first compute the MAE of the predictions using the test labels as they were originally (incorrectly) computed for each experiment, which results in inflated, "apparent" skill estimates. We then compute the adjusted MAE using the correctly cleaned test labels, which represents the true prediction error that would be observed in settings with no prior knowledge of the test data, such as real-time forecasting.

The model was trained using the Adam optimizer to minimize mean-square error loss. Additional parameters include a batch size of 64 and a learning rate of 0.000 01. Parameters were selected to minimize loss for the `clean` experiment, and the same parameters were used for all experiments.

The predictions from the `clean` experiment and its corresponding skill are shown in blue in Fig. 4. Comparisons between experiments are summarized below.

- For the `trend` experiment (Fig. 4, orange line), the linear trend was computed over the full dataset (0.1048°C decade^{-1}; January 1900–December 2024). The trend over the full dataset is nearly 3 times higher than the trend over the training period. Thus, the error is lower for the trend predictions, due to knowledge of the increased trend during model training. However, this trend from the testing period would not be known in a true testing sense where the testing data are completely unseen. Thus, the model's performance is inflated,

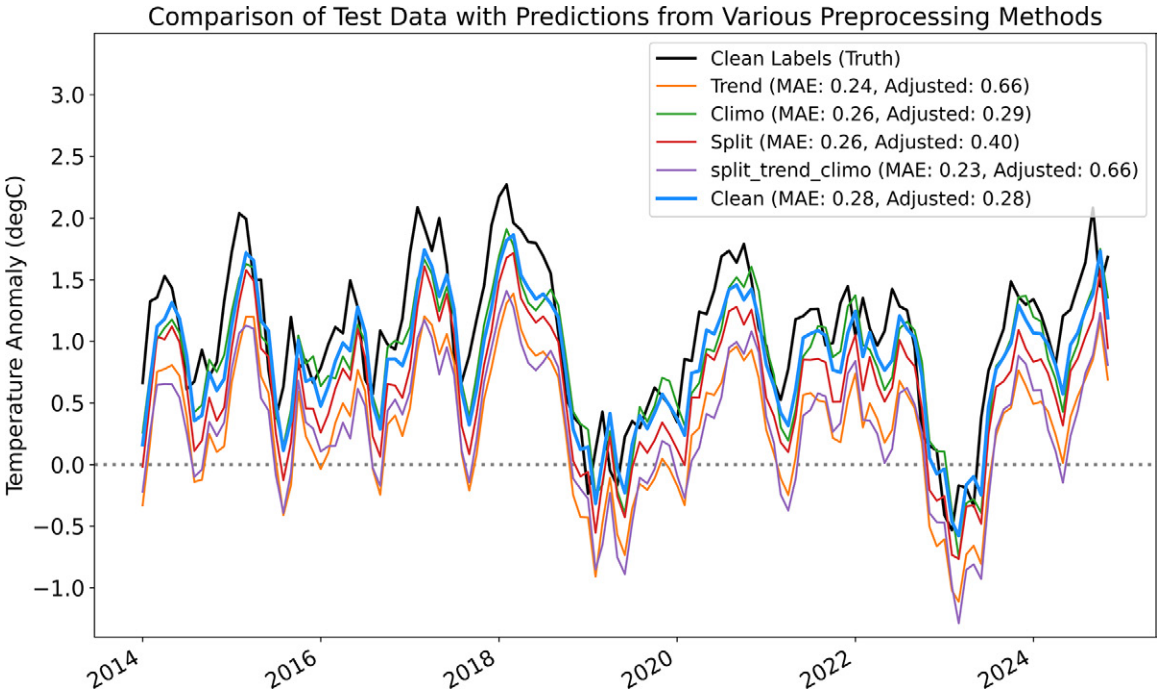

Fig. 4. Temperature anomaly (black; the truth) and neural network predictions using test datasets preprocessed in different ways: 1) trend computed during the test period (`trend`; orange); 2) climatology computed during the validation and test periods (`climo`; green); 3) data splits with potential leakage (`split`; red); 4) a combination of the trend, climo, and split preprocessing steps (`split_trend_climo`; purple); and 5) no data leakage (`clean`; blue). Corresponding skill scores are shown in parentheses. The MAE (°C) is first calculated using the incorrectly computed test labels for each experiment, producing inflated skill estimates. The "adjusted" MAE is calculated using the properly cleaned test labels to obtain an accurate measure of prediction error.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

as shown by the adjusted MAE being much higher; the trend has the largest impact on the overall skill in this case study.

- For the `climo` experiment (Fig. 4, green line), we use climatology calculated from January 1991 to December 2020 (i.e., spanning both the validation and testing periods). We find the error is slightly lower than the `clean` experiment, but again, this skill is inflated due to knowledge of information from the test data.
- For the `split` experiment (Fig. 4, red line), we split the training, validation, and testing sets by only a 1 month separation instead of splitting the data with a sizable gap between the next dataset to avoid data leakage from low-frequency variability. The resulting error is lower than the `clean` experiment, but the skill is again inflated due to data leakage.
- Finally, for the `split_trend_climo` experiment (Fig. 4, purple line), we combine the preprocessing missteps from the three previous experiments. The result is a slightly lower error than those three experiments, resulting in artificially inflated skill when compared with the adjusted MAE.

5. Conclusions

The use of AI/ML in climate prediction is rapidly expanding, introducing challenges with model design, skill assessment, and ultimately trust. A useful step toward building that trust is transparency in the process, including the initial problem design and data preprocessing. This work presents recommended steps for proper dataset preprocessing for different climate prediction problems. Such steps are recommended across applications and will serve as a way to make conscious choices when framing a prediction problem for climate time scales. The two case studies presented illustrate the importance of our recommended preprocessing steps and show how they can affect the interpretation of the predictions. Understanding the importance of these preprocessing steps will likely lessen the frequent critique of the "black box" nature of AI/ML (McGovern et al. 2019). Hence, following the recommendations laid out in this article will move the community toward AI/ML applications for climate prediction that are transparent and fairly evaluated.

While important, these recommended practices are not a replacement for coproduction of knowledge in AI/ML. Uncertainties, biases, and other unknowns in climate prediction studies require further work. Moreover, the data preprocessing steps presented here are not a complete substitute for the need to engage with domain experts and stakeholders alike in selecting the appropriate datasets, methods, and verification metrics. Thus, we encourage further collaboration between academics, operational forecasters, and industry scientists to ensure the model predictions are transparent, reproducible, and actionable. This collaboration and transparency includes ensuring any code for the model and the preprocessing steps be openly available. We also advocate for similar recommendations for benchmarking and evaluating AI/ML predictions used in subseasonal-to-seasonal and seasonal-to-decadal time scales. Open sourcing of all recommendations and associated software, from preprocessing to evaluation, would provide the community with an end-to-end roadmap to using AI/ML for a variety of climate prediction problems across scales and applications.

Acknowledgments. The authors thank the U.S. Climate Variability and Predictability Program (CLIVAR) for its support for this paper. M. J. M. was supported by the National Science Foundation (NSF) Grant 2425735. M. C. A. was funded, in part, by the National Oceanic and Atmospheric Administration (NOAA) Grant NA24OARX431C0065-T1-01 and by the U.S. Department of Energy's (DOE) Office of Biological and Environmental Research (BER) Regional and Global Model Analysis (RGMA) program, as part of the Program for Climate Model Diagnosis and Intercomparison (PMCDI) Project. J. S. P. C. was supported by a University of Maryland Grand Challenges Seed Grant. T. B. received support from the Horizon Europe project "Artificial Intelligence for enhanced representation

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

of processes and extremes in Earth System Models (AI4PEX)" (Grant Agreement ID: 101137682), funded by the Swiss State Secretariat for Education, Research and Innovation (SERI, Grant 23.00546). The scientific results and conclusions, as well as any view or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NWS, NOAA, or the Department of Commerce.

Data availability statement. ERA5 used in case study 1 can be obtained from the U.S. National Science Foundation (NSF) National Center for Atmospheric Research (NCAR) Research Data Archive (https://doi.org/10.5065/D6X34W69). Case study 1 and Fig. 1 software are publicly available at https://github.com/jhayron-perez/WRs_Preprocessing_BAMSPaper. Case study 2 software is publicly available at https://github.com/mbarcodia/bams-preprocess.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

References

Baumgartner, A., S. Molani, Q. Wei, and J. Hadlock, 2022: Imputing missing observations with Time Sliced Synthetic Minority Oversampling Technique. arXiv, 2201.05634v1, https://doi.org/10.48550/arXiv.2201.05634.

Betancourt, C., T. Stomberg, R. Roscher, M. G. Schultz, and S. Stadtler, 2021: AQ-Bench: A benchmark dataset for machine learning on global air quality metrics. *Earth Syst. Sci. Data*, **13**, 3013–3033, https://doi.org/10.5194/essd-13-3013-2021.

Box, G. E. P., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **26B**, 211–243, https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Bretherton, C. S., M. Widmann, V. P. Dymnikov, J. M. Wallace, and I. Bladé, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12**, 1990–2009, https://doi.org/10.1175/1520-0442(1999)012<1990:TENOSD>2.0.CO;2.

Cai, J., J. Luo, S. Wang, and S. Yang, 2018: Feature selection in machine learning: A new perspective. *Neurocomputing*, **300**, 70–79, https://doi.org/10.1016/j.neucom.2017.11.077.

Camps-Valls, G., and Coauthors, 2025: Artificial intelligence for modeling and understanding extreme weather and climate events. *Nat. Commun.*, **16**, 1919, https://doi.org/10.1038/s41467-025-56573-8.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: A machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Wea. Forecasting*, **37**, 1509–1529, https://doi.org/10.1175/WAF-D-22-0070.1.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002: SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357, https://doi.org/10.1613/jair.953.

de Burgh-Day, C. O., and T. Leeuwenburg, 2023: Machine learning for numerical weather and climate modelling: A review. *Geosci. Model Dev.*, **16**, 6433–6477, https://doi.org/10.5194/gmd-16-6433-2023.

Dueben, P. D., and P. Bauer, 2018: Challenges and design choices for global weather and climate models based on machine learning. *Geosci. Model Dev.*, **11**, 3999–4009, https://doi.org/10.5194/gmd-11-3999-2018.

——, M. G. Schultz, M. Chantry, D. J. Gagne, D. M. Hall, and A. McGovern, 2022: Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. *Artif. Intell. Earth Syst.*, **1**, e210002, https://doi.org/10.1175/AIES-D-21-0002.1.

Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the United States using ERA5 proximity soundings. *Wea. Forecasting*, **36**, 2143–2160, https://doi.org/10.1175/WAF-D-21-0056.1.

Grams, C. M., R. Beerli, S. Pfenninger, I. Staffell, and H. Wernli, 2017: Balancing Europe's wind power output through spatial deployment informed by weather regimes. *Nat. Climate Change*, **7**, 557–562, https://doi.org/10.1038/nclimate3338.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, https://doi.org/10.1038/s41586-019-1559-7.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hewamalage, H., K. Ackermann, and C. Bergmeir, 2023: Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Min. Knowl. Discovery*, **37**, 788–832, https://doi.org/10.1007/s10618-022-00894-5.

Huang, N. E., and Coauthors, 1998: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. Roy. Soc. London*, **454A**, 903–995, https://doi.org/10.1098/rspa.1998.0193.

Lee, S. H., M. K. Tippett, and L. M. Polvani, 2023: A new year-round weather regime classification for North America. *J. Climate*, **36**, 7091–7108, https://doi.org/10.1175/JCLI-D-23-0214.1.

Lidwell, W., K. Holden, and J. Butler, 2003: *Universal Principles of Design: 100 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach through Design.* Rockport Publishers, 272 pp.

Long, X., and Coauthors, 2025: Evaluating current statistical and dynamical forecasting techniques for seasonal coastal sea level prediction. *J. Climate*, **38**, 1477–1503, https://doi.org/10.1175/JCLI-D-24-0214.1.

Mamalakis, A., I. Ebert-Uphoff, and E. A. Barnes, 2022: Explainable artificial intelligence in meteorology and climate science: Model fine-tuning, calibrating trust and learning new science. *xxAI—Beyond Explainable AI: xxAI 2020*, A. Holzinger et al., Eds., Springer International Publishing, 315–339, https://doi.org/10.1007/978-3-031-04083-2_16.

——, E. A. Barnes, and J. W. Hurrell, 2023: Using explainable artificial intelligence to quantify "climate distinguishability" after stratospheric aerosol injection. *Geophys. Res. Lett.*, **50**, e2023GL106137, https://doi.org/10.1029/2023GL106137.

Mayer, K. J., K. Dagon, and M. J. Molina, 2024: Can transfer learning be used to identify tropical state-dependent bias relevant to midlatitude subseasonal predictability? arXiv, 2409.10755v1, https://doi.org/10.48550/arXiv.2409.10755.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Molina, M. J., J. H. Richter, A. A. Glanville, K. Dagon, J. Berner, A. Hu, and G. A. Meehl, 2023a: Subseasonal representation and predictability of North American weather regimes using cluster analysis. *Artif. Intell. Earth Syst.*, **2**, e220051, https://doi.org/10.1175/AIES-D-22-0051.1.

——, and Coauthors, 2023b: A review of recent and emerging machine learning applications for climate variability and weather phenomena. *Artif. Intell. Earth Syst.*, **2**, e220086, https://doi.org/10.1175/AIES-D-22-0086.1.

Nathaniel, J., Y. Qu, T. Nguyen, S. Yu, J. Busecke, A. Grover, and P. Gentine, 2024: ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. arXiv, 2402.00712v5, https://doi.org/10.48550/arXiv.2402.00712.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Pérez-Carrasquilla, J. S., and M. J. Molina, 2025: An Earth-system-oriented view of the S2S predictability of North American weather regimes. *Artif. Intell. Earth Syst.*, **4**, 240075, https://doi.org/10.1175/AIES-D-24-0075.1.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002203, https://doi.org/10.1029/2020MS002203.

——, and Coauthors, 2024: WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *J. Adv. Model. Earth Syst.*, **16**, e2023MS004019, https://doi.org/10.1029/2023MS004019.

Risbey, J. S., and Coauthors, 2021: Standard assessments of climate forecast skill can be misleading. *Nat. Commun.*, **12**, 4346, https://doi.org/10.1038/s41467-021-23771-z.

Rivera Tello, G. A., K. Takahashi, and C. Karamperidou, 2023: Explained predictions of strong eastern Pacific El Niño events using deep learning. *Sci. Rep.*, **13**, 21150, https://doi.org/10.1038/s41598-023-45739-3.

Rohde, R. A., and Z. Hausfather, 2020: The Berkeley Earth Land/Ocean Temperature Record. *Earth Syst. Sci. Data*, **12**, 3469–3479, https://doi.org/10.5194/essd-12-3469-2020.

Sweet, L.-B., C. Müller, M. Anand, and J. Zscheischler, 2023: Cross-validation strategy impacts the performance and interpretation of machine learning models. *Artif. Intell. Earth Syst.*, **2**, e230026, https://doi.org/10.1175/AIES-D-23-0026.1.

Watson-Parris, D., and Coauthors, 2022: ClimateBench v1.0: A benchmark for data-driven climate projections. *J. Adv. Model. Earth Syst.*, **14**, e2021MS002954, https://doi.org/10.1029/2021MS002954.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC

Wulff, C. O., F. Vitart, and D. I. V. Domeisen, 2022: Influence of trends on subseasonal temperature prediction skill. *Quart. J. Roy. Meteor. Soc.*, **148**, 1280–1299, https://doi.org/10.1002/qj.4259.

Xie, J., L. Sun, and Y. F. Zhao, 2025: On the data quality and imbalance in machine learning-based design and manufacturing—A systematic review. *Engineering*, **45**, 105–141, https://doi.org/10.1016/j.eng.2024.04.024.

Yang, R., and Coauthors, 2024: Interpretable machine learning for weather and climate prediction: A review. *Atmos. Environ.*, **338**, 120797, https://doi.org/10.1016/j.atmosenv.2024.120797.

Yeo, I.-K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959, https://doi.org/10.1093/biomet/87.4.954.

Yik, W., M. Sonnewald, M. C. A. Clare, and R. Lguensat, 2023: Southern Ocean dynamics under climate change: New knowledge through physics-guided machine learning. arXiv, 2310.13916v2, https://doi.org/10.48550/arXiv.2310.13916.

Yu, S., and Coauthors, 2024: ClimSim-Online: A large multi-scale dataset and framework for hybrid ML-physics climate emulation. arXiv, 2306.08754v6, https://doi.org/10.48550/arXiv.2306.08754.

Zantvoort, K., B. Nacke, D. Görlich, S. Hornstein, C. Jacobi, and B. Funk, 2024: Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digital Med.*, **7**, 361, https://doi.org/10.1038/s41746-024-01360-w.

Zhu, J.-J., M. Yang, and Z. J. Ren, 2023: Machine learning in environmental research: Common pitfalls and best practices. *Environ. Sci. Technol.*, **57**, 17 671–17 689, https://doi.org/10.1021/acs.est.3c00026.

Brought to you by UNIVERSITY OF OKLAHOMA LIBRARY | Unauthenticated | Downloaded 09/27/26 07:09 PM UTC