Abstract—In the past forty years, the high-performance computing (HPC) community has been developing powerful and rigorous tools for predicting the performance of supercomputers from log traces. In this paper, we transform one of these approaches previously used for predicting idle resources in high-end clusters into a method for capturing extreme climate events in geographical locations of interest. Our method uses an analysis based on empirical cumulative distribution functions (ECDFs) to benchmark and model occurrences of climate events including extreme temperature and precipitation. The method comprises two phases: a learning phase and a prediction phase. The learning phase applies the ECDF-based empirical analysis to historical climate data in order to identify suitable modeling and forecasting windows, both given in years. The prediction phase applies the modeling window to the most recent climate data in order to estimate the likelihood that given portions of the region of interest can experience extreme climate events in the forecasting window. The research is the first of its kind to extend HPC performance modeling techniques to study extreme climate events.

I. INTRODUCTION

This paper presents new computational methods for predicting the likelihood of extreme climate occurrences from big climate data. To this end, we transform existing frequency analysis approaches that are successfully used to model and predict event occurrences in high-end computer clusters (e.g., CPU idle times) into methods that allow us to model and predict the occurrence of extreme climate conditions such as extremely high or low temperature and high precipitation. This work is inspired by previous work of our group [1] in which we predicted the occurrence of idle CPUs and communications in high-end clusters by collecting and benchmarking data from cluster nodes.

Remarkable similarities exist between the predictions of idle times in a high-end cluster and climate events in a geographical region using a frequency analysis approach. Figure 1 depicts a step-by-step analogy between cluster and climate modeling. In Step 1, the historical data—that is, performance traces for the cluster and climate data for the region of interest—is collected. For cluster modeling, this trace data comprises time-stamped performance measurements for cluster nodes in terms of CPU computation time, communication time, and idle time. For climate modeling, the traces are composed of spatial-temporal climate event data (e.g., temperature, precipitation at geographical points) for the region of interest. In Step 2 of the frequency analysis approach, suitable windows for event sampling are determined through benchmarking. A “window” is defined as the period of time for which the “events” must be sampled for the frequency analysis. An example of an event for cluster modeling is idle resource time occurrences, whereas an event in the context of climate modeling is the number of geographical points (given by latitude and longitude) experiencing abnormally high temperatures. In Step 3, the events are sampled in the analysis windows. In Step 4, an empirical cumulative distribution function (ECDF)-based frequency analysis is applied to the events in the analysis window. From the ECDF models obtained in Step 4, several interesting questions pertaining to cluster and climate modeling are answered in Step 5. For instance, a typical question regarding the compute cluster is “What is the likelihood that more than x% of the CPUs will be idle in the next window of y seconds in the high-end cluster?” Similarly, a typical question with respect to climate modeling is “What is the expected proportion of the region that will experience extremely hot climate in the next y years?”

In this paper, we redesign and extend the frequency analysis approaches used to study performance traces by Pallipuram and co-workers in [1] to answer climate-related questions and we integrate the transformed approach into a general climate data analysis framework. From the broad range of questions that our framework can answer, we are interested in the following critical questions: (1) “What is the expected proportion of a given region of interest that will experience extremely hot climate in the next y years?”

Fig. 1: Step-by-step analogy between prediction of idle times in a compute cluster and climate events from historical data.
extreme climate events during a given season in the future years?” and (2) “What are the chances that this proportion of the region will experience the extreme climate event repeating multiple times in the next future years?” We answer these questions for the *matorral tropical* or *subtropical* subregion in Mexico. This region is particularly interesting for several reasons: it suffers from water limitations; it is susceptible to periodic drought; and it experiences rapid landscape changes as a result of agriculture-driven deforestation. Therefore, there is a significant need to understand how the climate variability influences this region. Specifically, the contributions of this paper are follows:

- We present a general climate data analysis framework that uses historical data to predict the likelihood of a given region undergoing an extreme climate event within a near-future time window of a given size.
- We use the framework to predict the proportion of the matorral tropical region in Mexico that likely will experience extremely high or low temperature as well as high precipitation multiple times and the recurrence period of these events.

The rest of this paper is structured as follows. Section II describes our climate data analysis framework based on the frequency analysis approach. Section III shows the framework put into action where we answer the previously mentioned critical questions. Section IV discusses related work. We provide conclusions in Section V.

### II. CLIMATE DATA ANALYSIS

This section discusses the climate modeling workflow along with its two phases: the learning phase and prediction phase.

#### A. Modeling Workflow

Figure 2 shows the high-level workflow of our climate data analysis. The process is partitioned into two key phases: a learning phase and a prediction phase. In the learning phase, we build window-size models for predictions; we use these models in the prediction phase to estimate the occurrence of extreme climate conditions.

In the *learning phase* in Figure 2, we first choose a geospatial region of interest from a larger geographical area and select the temporal climate data associated with the region. We refer to this data as the historical climate data. In this phase, we are interested in automatically identifying extreme climate conditions (e.g., unusually high or low temperature, high precipitation) within the region of interest. We annotate the climate data as either “normal” or “expected” (when associated to expected weather conditions for the considered season) or “extreme” (when associated to extreme weather conditions). The annotation procedure for climate data is similar to traces in the traditional HPC approach, where the traces are streams of compute node performance data over time [1]. We apply the ECDF-based frequency analysis to extreme event occurrences in the annotated data to benchmark the historical climate data. The analysis provides the optimal window-size models (i.e., optimal modeling and forecasting windows) for predicting future occurrences of the same events across the region of interest.

In the *prediction phase*, we consider the same region of interest, and we use the window-size models built in the previous phase to predict event occurrences over a different time interval. Specifically, the current climate data is input to a prediction module. This module uses the optimal modeling window to generate the ECDF model for predicting the likelihood that the region of interest experiences a given occurrence of the extreme event or *hot spot* of interest. In general, the likelihood information is used by the scientists to estimate with a certain confidence whether a sequence of extreme events will take place within an optimal window size. An interesting and highly informative representation of ECDF model is the *survival model*, which provides the likelihood for an event exceeding a particular value, namely, \( P(X > x) \), often called the probability of exceedance. It is simply obtained as \( 1 - ECDF \). In Section III, we show how the survival and window-size models enable us to answer the critical questions posed in Section I.

#### B. Learning Phase

Figure 3 shows the learning phase of the framework with its two modular stages consisting of the “feature extraction stage” and the “cumulative distribution function stage.” The learning phase has two inputs: a Shape file and a set of GeoTiff files. The Shape file contains the geospatial vector of points associated with the user-selected geographical region of interest. The GeoTiff files contain the georeferenced temporal climate event information sampled at a given time resolution (e.g., daily, monthly, yearly) over a given time period. In this paper, we use the GeoTiff files for temperature and precipitation events sampled at monthly resolution.

Figure 4 zooms into the components of the “feature extraction stage.” As shown in the figure, the Shape file and GeoTiff files are the inputs to the feature extraction stage. This stage first masks the GeoTiff files by using the Shape file to extract only temporal climate events of interest for the study (via the
The geo-referenced temporal climate data used for our study come in sets of two-dimensional matrices sorted based on the latitudes and longitudes of the geographical points in the GeoTiff files. We use the Shape file to mask those geographical points that are not part of the region of interest. Ultimately, the region of interest should consist only of those points in the 2D matrix associated with geographical areas with specific characteristics that are relevant to scientists (e.g., regions with tropical climate or with high population density). To this end, for each temporal snapshot of the spatial climate data, the feature extraction mask in Figure 4 first removes unmeasured points (e.g., locations over sea and/or lakes) and then marks off those remaining points that are not associated with the region of interest. The feature extraction mask then flattens the remaining sparse 2D matrix of selected points into an unordered set of climate measurements (e.g., temperature or precipitation).

The extracted events are then provided to the extreme event classifier that categorizes the climate events as either “normal” or “extreme.” To this end, the extreme event classifier first builds the empirical distribution of the data; Figure 5(a) shows the distribution of measured temperatures used in this paper for our example. Next, the classifier automatically defines the thresholds for extreme conditions (e.g., unusually high or low temperature or high precipitation) as a function of the month in which the measurements are taken. Specifically, it uses the top five percentile for upper extreme conditions and the bottom five percentile for extremely low conditions for the same month across the sampled years, as shown in Figure 5(b) for the same temperature dataset used in Figure 5(a). With some modification, the framework allows us to incorporate other methods of determining extreme event thresholds (e.g., [2]).

The classifier uses the thresholds to annotate the climate measurements as extreme or normal in order to generate the climate traces. Figure 6 shows an example of the annotated traces for a set of temperatures where the x-axis is time in months and the y-axis consists of unordered geographical points (in terms of latitudes and longitudes). In the figure, extremely low temperatures are in green, extremely high temperatures are in yellow, and normal temperatures are in gray. An interesting observation that supports our approach is the presence of visual signs of periodicity. The annotated climate traces are then input to the climate event occurrence generator that counts the number of geographical points of the given region observing extreme climate events, thereby transforming the climate events from the time domain to the frequency domain.

The cumulative distribution function (CDF) stage shown in Figure 7 has two main components: the cumulative distribution function generator and the Kolmogorov-Smirnov (KS) test [3]. The stage inputs the frequency-domain climate events from the feature extraction stage and outputs the optimal size of modeling and forecasting windows. The modeling window is defined as the number of past years for which the climate events should be sampled in order to construct an ECDF model. The forecasting window is defined as the number of future years that can be studied by using the ECDF model given by the modeling window. As shown in the figure, the frequency-domain climate events first enter the CDF generator. This generator samples the frequency-domain climate events in two separate windows: the modeling window and forecasting window. The sampling interval is one year for modeling
Fig. 5: Temperature data: (a) distributions across 12 months and (b) thresholds computed by the extreme event classifier.

Fig. 7: Cumulative distribution function stage with its inputs (i.e., frequency-domain climate events) to yield optimal modeling and forecasting windows.

each month in a calendar year. Since climate events for a given month can statistically depend on adjacent months, each sample for a given month also includes climate event values from adjacent months (e.g., the January sample also includes the December and February samples). After the completion of frequency-domain climate event sampling in the modeling and forecasting windows, the CDF generator builds two ECDF models: $ECDF_M$ using the modeling window and $ECDF_F$ using the forecasting window. To determine the suitability of these window models for predictive analysis, the CDF stage compares the two ECDF models for statistical similarity using the KS test. If the two ECDF models are similar, the KS test results in a hit, meaning the modeling window is statistically suitable for modeling the forecasting window. For each month in a calendar year, this process is repeated on the entire historical data with a sliding factor equal to one year, and KS test hit-rates are recorded for a given pair of modeling and forecasting windows. The hit rate is defined as the percentage of KS test hits observed for the complete traversal of the historical data. The CDF stage performs the historical data traversal for several empirically defined pairs of modeling and prediction windows to obtain a three-dimensional scatter space (i.e., hit rate versus modeling window-size versus forecasting window-size). The modeling and forecasting window pair that yields the highest hit rate is selected as the optimal window-size models for the ECDF-based frequency analysis in the prediction phase.

C. Prediction Phase

The prediction phase uses the optimal modeling and forecasting windows from the learning phase to perform the ECDF-based frequency analysis on the most recent climate data. This phase is operated by a prediction module, which is structurally similar to that of the learning phase (see Figure 3). The prediction module estimates the likelihood of certain climate event occurrences in a given geographical region. The inputs to the prediction module are the most recent climate data and a user query. The feature extraction stage converts the time-domain climate events to frequency-domain climate events. The CDF generator collects the frequency-domain samples in the optimal modeling window to generate the ECDF model. Based on the user query, the ECDF model in conjunction with the window-size models provides the likelihood prediction for the climate event occurrences in the geographical region for the next forecasting window.

III. FRAMEWORK IN ACTION

In this section, we put the climate data analysis framework discussed in Section II into action. We first discuss the critical questions and experiment set-up; then we describe three climate case studies in detail.

A. Critical Questions and Experiment Set-Up

We use our climate data analysis framework for examining three case studies, each addressing a critical question related to three types of extreme climate events: unusually high temperature, unusually low temperature, and unusually high precipitation. Specifically we want to answer the following questions: (1) What is the expected proportion of a region of interest that may experience unusually high-temperature events above a predetermined threshold? What is likelihood that this expected proportion will be exceeded multiple times
in the next years (henceforth referred to as exceeding high-temperature occurrences), and what is its recurrence period? (Case Study I); (2) What is the expected proportion of the region that may experience unusually low-temperature events? What is the likelihood of multiple exceeding low-temperature occurrences and what is its recurrence period? (Case Study II); and (3) What is the expected proportion of the region experiencing unusually high precipitation events (above a predetermined threshold)? What is the likelihood of multiple exceeding high-precipitation occurrences, and what is the recurrence period? (Case Study III).

Our region of interest is the matorral tropical or subtropical region in Mexico, as shown by the dark blue patches in Figure 8. As mentioned in Section I, this region is highly interesting for climate study because of its striking characteristics including water limitations and frequent landscape changes owing to interest in agriculture. The corresponding data is collected over 24 years, from 1980 to 2013. As described in Section II-B, this data is obtained by supplying the Shape file and GeoTiff files to the feature extraction mask of the feature extraction stage (see Figure 4).

![Fig. 8: Region of interest selected by a scientist for our study (i.e., dark blue patches in the map).](image)

For all three case studies, the learning phase uses the climate data from 1980 to 2000 as the benchmarking data to build the modeling and forecasting windows. The resulting modeling and forecasting windows are validated over a 13-year window, from 2001 to 2013, with a validation sliding factor equal to one year. Specifically, we perform the KS test counting the hits and computing the hit rate observed on the validation data from 2001 to 2013 using the modeling and forecasting windows generated by the learning phase. The KS test results in a hit if the ECDF models built using modeling and forecasting windows are statistically similar, and a miss otherwise. The hit rate is measured as the percentage of hits observed for the entire validation traversal over the 13-year window. We use the windows with higher confidence for future predictions.

### B. Case Study I: High Temperature

Case Study I deals with unusually high-temperature events. In this case study, we want to predict the expected proportion of the region observing unusually high temperatures during the four seasons and estimate the chances that future years may observe greater proportions of the region (more than the expected) experiencing unusually high temperatures multiple times (i.e., exceeding high-temperature occurrences). Figure 9 shows the three dimensional scatter space (i.e., hit rate versus modeling window versus forecasting window) obtained in the learning phase for the training temperature data. The blue circles refer to the hit rates for modeling-forecasting window pairs for which the KS test is confident, meaning the samples in the window pairs are statistically significant for the KS test. The red stars refer to the hit rates for the modeling-forecasting pairs for which the KS test is less confident. The learning phase estimates that the modeling window achieving the highest hit rate is equal to 4 years and the forecasting window is 6 years. In other words, this pair achieves the highest hit rate of 85% among the other window pairs, as shown in Figure 9. In our validation, the modeling window of 4 years and forecasting window of 6 years are applied to model all months in years 2001 through 2013 with a sliding factor of one year. The KS test hits are recorded. Specifically for Case Study I, the years 2001–2004, 2002–2005, 2003–2006, and so on constitute the modeling windows, whereas the corresponding forecasting windows are 2005–2010, 2006–2011, 2007–2012, and so on. Figure 10 shows the prediction of 4 years and forecasting window of 6 years are applied to model all months in years 2001 through 2013 with a sliding factor of one year. The KS test hits are recorded. Specifically for Case Study I, the years 2001–2004, 2002–2005, 2003–2006, and so on constitute the modeling windows, whereas the corresponding forecasting windows are 2005–2010, 2006–2011, 2007–2012, and so on. Figure 10 shows the prediction...
Fig. 11: Case Study I: Survival models for Spring (a) Summer (b) Fall (c) Winter (d).

hits for all the months in the forecasting window within 2001–2013. As seen in this figure, the modeling and forecasting windows achieve a 94% hit rate, meaning that the window pair is optimal for predictive analysis.

To obtain the expected proportion of the region that may experience unusually high temperature, the prediction phase employs the event samples collected in the modeling window. Recall that an event sample in this context is the proportion of the region experiencing unusually high temperature. Since all the event samples in the modeling window are equally likely, the prediction phase obtains the expected proportion as:

\[ E[X] = \frac{1}{N} \sum_{i=1}^{N} x_i \]  

where \( x_1, x_2, \ldots, x_N \) are the event samples in the modeling window. Using Equation 1 and modeling window of 4 years, the prediction phase obtains the expected proportion for spring (April), summer (July), fall (October), and winter (January) as 12.3%, 23.1%, 18.2%, and 6.8%, respectively.

We also want to determine (1) the likelihood of multiple exceeding high-temperature occurrences and (2) the recurrence period. To this end, the prediction phase utilizes the survival models given in Figures 11(a) through 11(d). For each season, the prediction phase queries the corresponding survival model to obtain the probability \( p \) of exceeding the expected proportion of the region. In general for a forecasting window of size \( y \) years, the exceeding climate occurrences can be anywhere from zero up to \( y \) times. The probability that \( i \) number of exceeding climate occurrences happen within the forecasting window of \( y \) years is given by

\[ P = \binom{y}{i} \cdot p^i \cdot (1 - p)^{y-i}. \]  

For Case Study I, the size of the forecasting window is \( y = 6 \) years; therefore the exceeding high-temperature occurrences can happen anywhere from 0 to 6 times in the next 6 years. Using Equation 2, we obtain the likelihood of multiple exceeding high-temperature occurrences for each of the four seasons in Figures 12(a) through 12(d). The recurrence period of exceeding occurrences is simply the inverse of the probability \( p \) of exceeding the expected proportion of the region. We obtain the recurrence period for spring, summer, fall, and winter as 1.2 years, 1.5 years, 1.4 years, and 1.5 years, respectively.

Our framework provides interesting insights into the high-temperature patterns of the matorral desertico tropical region over the next 6 years. Our approach tells us that both spring and fall are more likely to observe a proportion of the region greater than 12.3% and 18.2%, respectively, experiencing unusually high temperature only once in the next 6 years (see Figure 12(a)), with recurrence periods of 1.2 and 1.4 years each. On the other hand, both summer and winter are more likely to observe a proportion of the region greater than
23.1% and 6.8%, respectively, experiencing unusually high temperature twice in the next 6 years, with recurrence periods of 1.5 years each.

C. Case Study II: Low Temperature

Case Study II deals with unusually low-temperature events. Specifically we want to predict the expected proportion of the region observing extremely low temperatures during the four seasons and the likelihood of exceeding the expected proportion of the region multiple times. Figure 13 shows the three-dimensional scatter space (hit rate versus modeling window versus forecasting window) obtained from the learning phase for the training temperature data. Similar to Case Study I, the learning phase uses the monthly temperature values for years 1980 through 2000 as training data in order to obtain the optimal modeling and forecasting windows. In this case study, the learning phase identifies a modeling window of size 5 years and a forecasting window of size 6 years. This window pair achieves the highest hit rate (equal to 92%).

The validation process is similar to that for Case Study I. The modeling window of 5 years and forecasting window of 6 years are applied to temperature data from 2001 through 2013, and KS test hits are recorded. Figure 14 shows the corresponding hits and hit rate observed for the low-temperature climate event modeling for each calendar month. As seen in the figure, the modeling and forecasting windows achieve a 95% hit rate, meaning that the window pair is optimal for predictive analysis.

To determine the expected proportion of the region observing unusually low-temperature events in the four seasons, the prediction phase utilizes the event samples collected in the modeling window of size 5 years and applies Equation 1, as described in the previous case study. The expected proportions of the region that may experience lower temperature than usual in spring, summer, fall, and winter are 5.5%, 1.9%, 8.3%, and 15.6%, respectively.

Using the survival models given in Figure 15, we cal-
calculate the likelihood of multiple exceeding low-temperature occurrences and the recurrence period. Since we obtained a forecasting window of 6 years for this case study, the region can observe exceeding low-temperature occurrences anywhere from 0 to 6 times in the next 6 years. Using Equation 2, we obtain the likelihood of multiple exceeding low-temperature occurrences for the four seasons in Figure 16 as described in Case Study I. Similar to Case Study I, the recurrence period for the four months is the inverse of the probability \( p \) of exceeding the expected proportion of the region; the values are equal to 1.4 years each for spring and summer and 1.5 and 1.6 years for winter and fall, respectively.

For Case Study II, our framework outlines how both Spring and Summer have approximately equal chances of observing a proportion of the region greater than 5.5% and 1.9% respectively, experiencing unusually low temperatures once or twice in the next 6 years, with a recurrence period of 1.4 years each (see Figures 16(a) and 16(b)). On the other hand, the fall and winter are more likely to observe such exceeding low-temperature occurrences twice in the next 6 years, with a recurrence period of 1.5 and 1.6 years each (see Figure 16(c) and 16(d)).

### D. Case Study III: High Precipitation

Case Study III deals with a different type of climate data: the level of precipitation. We seek to predict the expected proportion of the region observing unusually high precipitations during the four seasons and to determine the likelihood of exceeding the expected proportion of the region multiple times. Figure 17 shows the three-dimensional scatter space (i.e., hit rate versus modeling window versus forecasting window) obtained in the learning phase for the training precipitation data. The highest hit rate of 93.5% is in this case associated with a modeling window of 4 years and a forecasting window of 6 years. For validation, Figure 18 provides the KS test hits using the modeling-forecasting window pair obtained from the learning phase, when tested on years 2001–2013 with a sliding factor equal to one year for all calendar months. As seen in the figure, the selected window pair achieves a hit rate of up to 95%, implying the pair’s efficacy to model future extreme climate events.

Using the modeling window of 4 years, the prediction phase
yields the expected proportions of region (see Equation 1) observing unusually high precipitation in the four seasons; the values are 7.5%, 9.3%, 6.8%, and 10.2%, respectively.

Using the survival models in Figure 19, we determine the likelihood of multiple exceeding high-precipitation occurrences and identify the recurrence period (1.3 years each for spring and summer; and 1.4 and 1.2 years for fall and winter). The results are given in Figure 20. As shown in the figure, our framework predicts that spring, summer, fall, and winter are all likely to observe one exceeding high-precipitation occurrence where more than the expected proportion of the matorral desertico tropical region may observe unusually high precipitation in the next 6 years.

IV. RELATED WORK

Researchers increasingly are recognizing the importance of the spatial and temporal dependency of spatial science data (see, e.g., [4]; [5]; [6]). At the same time, an unexplored opportunity exists to apply computational resources and computer science approaches to identify patterns in big data [7], [8], [9], and [10].

Much attention has been drawn to the study of extreme events related to terrestrial carbon dynamics [11], but these events are determined by statistical extremity with respect to a historical reference period (e.g., extraordinary deviation from the median of probability distributions [12]). Therefore, our understanding of the “extremeness” of these events is limited by the length of the time series (usually about a decade) and the scatter information from independent studies on the effect of “extreme climate events” such as hurricanes, fires, and droughts [13], [14], [15]. Currently, extreme statistical events can be determined by assessing extraordinary deviations from the median of a probability distribution [12] by calculating the upper 95th percentile using extreme-value distribution theory [17]. Therefore, spatio-temporally connected features that exceed a certain threshold of extremeness can be defined by using basic statistical principles. This approach has been used to analyze gross primary production data [18].

Algorithms developed by the climate science community identify hot spots (i.e., areas that show disproportionately high or low magnitude, changes, or rates in a defined variable) in a multidimensional climate space between present and future periods [20] and [19]. A multidimensional ensemble using climate- and carbon-related variables can be used to identify hot spots within a region of interest. Specifically, the standard Euclidean distance can be used to quantify the total change in this multidimensional space between two periods. Scaling to a maximum change in the response period provides a metric of aggregate change that can be compared between geographic areas. We studied an example of this methodology for carbon-related variables using unpublished information from Mexico (Vargas et al. unpublished). Unfortunately, although these
methods have proved effective for small data, their application results in poor scalability as the data increases.

V. CONCLUSIONS

This paper presents one of the first adaptations of an HPC performance modeling technique to benchmark and predict extreme climate events using big climate data. The paper presents a frequency analysis-based framework that comprises two phases: learning and prediction. The learning phase benchmarks the historical climate data in order to identify suitable modeling and forecasting windows for analyzing a specific geographical region of interest. The prediction phase uses the modeling and forecasting windows to predict future occurrences of extreme events in the same region of interest. Our framework enables us to answer several interesting questions: What is the expected proportion of the region that will experience a particular extreme climate event? What is the likelihood that this proportion of region is exceeded multiple times, and what is its recurrence period? The window-size models given by the learning phase of our tool are validated by using several test cases including unusually high- and low-temperature events and unusually high-precipitation events for four seasons. Our validation efforts yield more than 90% prediction hits. The end goal of this research is to present a framework that enables climatologists to study big climate data in a scalable manner.

ACKNOWLEDGMENTS

The authors are supported by NSF grants CCF-1318445/1318417. Code samples and timing data are available on github: https://github.com/TauferLab.

REFERENCES


Fig. 20: Likelihood of multiple exceeding high-precipitation occurrences in the forecasting window of 6 years for Case Study III.