Henri Funk, Alexander Sasse, Helmut Küchenhoff, Ralf Ludwig

Climate And Statistics

Contents

Preface

Author: Henri Funk

As the world faces the reality of climate change, natural hazards and extreme weather events have become a major concern, with devastating consequences for nature and humans. The quantification and definition of climate change, extreme events and its implications for life and health on our planet is one of the major concerns in climate science.

This book explains current statistical methods in climate science and their application. We do not aim to provide a comprehensive overview of all statistical methods in climate science, but rather to give an overview of the most important methods and their application. This book is the outcome of the seminar "Climate and Statistics" which took place in summer 2024 at the Department of Statistics, LMU Munich.

FIGURE 1: Creative Commons License

This book is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[1].

[1] http://creativecommons.org/licenses/by-nc-sa/4.0/

Technical Setup

The book chapters are written in the Markdown language. To combine R-code and Markdown, we used rmarkdown. The book was compiled with the bookdown package. We collaborated using git and github. For details, head over to the book's repository[2].

[2] https://github.com/henrifnk/Seminar_ClimateNStatistics

1

Introduction

Author:

Supervisor:

1.1 Intro About the Seminar Topic

1.2 Outline of the Booklet

2

DL Classification of Weather Patterns over Europe

Author: Ziyu Mu, Laura Schlueter

Supervisor: Henri Funk

Degree: Master

2.1 Abstract

Daily weather in the mid-latitudes is dominated by large-scale atmospheric circulation types (CTs), which influence regional climate variability and extremes. Accurate classification of these CTs is crucial for diagnosing long-term climate dynamics and analysing future changes. This study implements and extends a novel deep learning-based classification method for European CTs, building on the Hess & Brezowsky (HB) framework.

Using a convolutional neural network (CNN) architecture (inspired by applications in climate science and optimization techniques from), we trained the model on high-resolution ERA5 reanalysis data (1950–1980), with sea-level pressure (SLP) and geopotential height at 500 hPa (z500) fields as key predictors. To ensure robustness, a nested cross-validation approach was employed, achieving an overall accuracy of 53.17% and a macro F1 score of 55.14%, outperforming traditional classification methods.

Applied to the CanESM2 ensemble projections, which simulate data with assumptions that climate conditions were consistent with history, our results show significant future frequency shifts in CTs, notably an increase in Icelandic High, Cyclonic (HNZ) occurrences during both summer and winter half-years. This methodological advancement not only enhances classification accuracy but also offers robust ensemble analyses critical for climate-informed decision-making.

2.2 Introduction

In the mid latitudes, daily weather is dominated by large-scale atmospheric circulation patterns, defined by the position of high and low pressure centres (Häckel, 2021; Mittermeier et al., 2022). Due to steep temperature and pressure gradients along the frontal zone between polar and tropical air masses (around 45° latitude), the jet stream is formed in about 10 km altitude. It is characterized by mean wind speeds of 300 km/h and the long-term average wind direction is westerly (Häckel, 2021). Here, dynamic high and low pressure systems develop (Häckel, 2021; Mittermeier et al., 2022). However, the westerly jet stream is not stationary, but deflected with varying amplitude creating the so-called Rossby waves, which influence the position of high and low pressure systems (Häckel, 2021). While an infinite number of atmospheric conditions would be conceivable, recurrent weather patterns with similar meteorological features are observed in practice. Therefore, it is possible to classify weather patterns according to synoptic properties (Bissolli, 2001).

The first concept for a classification of different weather patterns was introduced by Baur et al. (1944). They used surface pressure charts from 1881 to 1939 and assigned the respective weather situation to a specific class for each day based on the spatial distribution of atmospheric pressure and the position of the frontal zones. This lead to a classification into 21 Grosswetterlagen for Europe and East Atlantic (Baur et al., 1944). In 1963 Baur defined Grosswetterlage as the mean spatial air pressure distribution of a large area, at least the size of Europe, over a period of several successive days. This classification has been revised and improved several times. In 1977, Hess and Brezowsky published their third edition of a revised catalogue of Grosswetterlagen. Here, the authors determined that a Grosswetterlage should only be classified if it can be recognised on at least three successive days. Furthermore, the number was extended to 29 different Grosswetterlagen. In addition to the surface pressure chart, the classification was also based on the geopotential height in 500hPa (Bissolli and Dittmann, 2001; Bissolli, 2001; Werner and Gerstengarbe, 2010). In the following, Werner and Gerstengarbe published seven editions of a catalogue of the Grosswetterlagen in Europe with updated data. The last edition (2010) covers the years 1881 to 2009 and provides daily information on the Grosswetterlage over Europe. Today, the German Weather Service (DWD) still constantly updates the catalogue of European Grosswetterlagen and publishes the results monthly. Since 1944, the classification of Grosswetterlagen is conducted by experts, hence it is subjectively biased (Hess, 2005). In the following, the classification according to Hess & Brezwosky will be referred to as HB CT.

HB CTs can be attributed to certain weather situation at certain locations in Europe (Werner and Gerstengarbe, 2010). Hence in the past, the classification of HB CTs has been used to identify regularities in frequency and duration of occurrences of certain weather conditions. Bissolli and Dittmann (2001) describes a relationship between frequency and duration of the HB CTs and mean annual air temperature and precipitation. Moreover, extreme weather events—such as heavy rainfall, floods, and heat waves—can often be linked to specific HB CTs (Mittermeier et al., 2022). Such kinds of application could not only be useful to analyse past climate. Future climate projections could also be analysed using the HB CTs, allowing conclusions to be drawn about future developments. However, analyzing future CTs requires many model ensembles to account for internal variability (Wyser et al., 2021). Manual classification of HB CTs for many ensemble members is no longer possible. It is therefore necessary to consider the possibility of automatization (Mittermeier et al., 2022). Mittermeier et al. (2022) have found a way to classify HB CTs automatically using a deep learning classifier.

2.3 Data

2.3.1 Data Sources

This section describes in detail the datasets utilized for model training, validation, and testing, providing comprehensive explanations of their characteristics, purposes, and roles within the study. Due to data availability constraints, we replaced the original study's ERA-20C (Poli et al., 2016) and SMHI-LENS (Wyser et al., 2021) data with ERA5 (European Centre for Medium-Range Weather Forecasts, 2017) (higher resolution, 1950–1980) for training and CanESM2 (Hua et al., 2015; Government of Canada, 2025) (use only 12 of 50 ensemble members) for projections.

2.3.1.1 Historical Data - Training and Validation

We use reanalysis dataset, which is the historical atmospheric data reconstructed by combining observational data and model simulations, for training and validation. Data is provided by European Centre for Medium-Range Weather Forecasts (ECMWF).

We only use two variables mentioned above:

1. SLP: Atmospheric pressure measured at Earth's surface, vital for identifying weather systems such as cyclones and anticyclones.
2. z500: Height of the 500 hPa pressure level in the atmosphere, indicative of atmospheric wave patterns and mid-tropospheric dynamics.

Due to reanalysis data's high-quality, accurate, and detailed historical atmospheric representation, it is enabled to get a robust and precise training of the deep learning model.

The difference of ERA5 (Figure 2.1) and ERA-20C (Appendix Figure 2.7) lies on:

TABLE 2.1: Difference between two reanalysis dataset

Reanalysis Dataset	Temporal Coverage	Spatial Resolution
ERA-20C (paper)	1900–1980	Approximately 125 km
ERA5 (implementation)	1950–1980	High-resolution, about 31 km

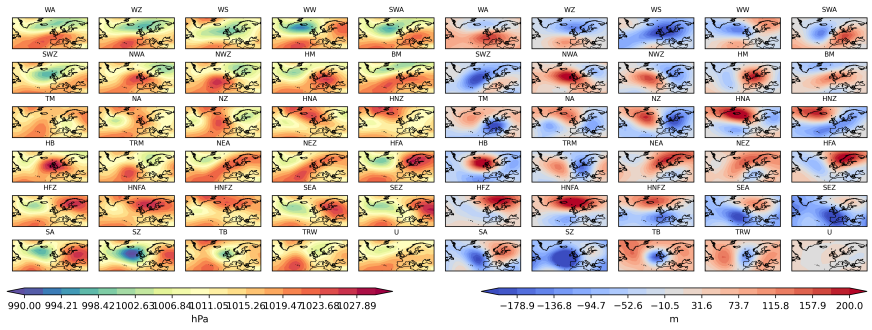


FIGURE 2.1: Synoptic patterns of the 29 HB CTs created using ERA5 data averaged over the period 1950 - 1980, showing SLP on the left and z500 on the right side.

2.3.1.2 Future Climate Projections - Change Analysis

We use ensemble datasets which projected future climate conditions under specific scenarios, to evaluate future changes in atmospheric circulation, considering internal variability, essential for accurate and reliable climate change projections. The data should be preprocessed with the same variables, scale, domain, resolution as training data.

1. CanESM2 (Government of Canada, 2025)
   - Name: CanESM2 (Canadian Earth System Model version 2)
   - Provider: Canadian Centre for Climate Modelling and Analysis (CCCma)
   - Model Used: Climate Model Intercomparison Project phase 5 (CMIP5)
   - Ensemble Members: 50 (due to downloading limitation, we only used 12 ensemble)
   - Scenario: historical, simulate the condition as observed past climate from 1850 to at least 2005 (Friedlingstein et al., 2008).
   - Temporal Coverage:
     - Historical Reference: 1971–2000
     - Future Projections: 2031–2060
   - Spatial Resolution: Around 2.8°
2. SMHI-LENS (Wyser et al., 2021)

- Name: SMHI-LENS (Swedish Meteorological and Hydrological Institute Large Ensemble)
- Provider: Swedish Meteorological and Hydrological Institute
- Model Used: EC-Earth3 (Climate Model Intercomparison Project phase 6, CMIP6)
- Ensemble Members: 50 (to effectively capture internal variability)
- Scenario: SSP3-RCP7.0, representing a high GHG emission pathway
- Temporal Coverage:
  - Historical Reference: 1991–2020
  - Future Projections: 2071–2100
- Spatial Resolution: Around 0.7°

As a result with higher greenhouse gas emission under SSP3-RCP7.0 scenario, and with a more distant future in the paper:

1. Future warming and circulation shifts might be more pronounced than our implementation.
2. Frequency changes in HB CTs may appear larger in the paper, simply because the climate change signal is stronger.

2.3.2 Data Preprocessing Steps

1. Reset spatial domain: only use data in Europe and the North Atlantic region (30°N–75°N latitude, 65°W–45°E longitude)
2. Spatial Regridding: All datasets were interpolated to a common 5° grid resolution, ensuring consistency and computational efficiency
3. Seasonal Centering: Applied to remove seasonal biases, helping the model better recognize atmospheric patterns independently of seasonal influences.

2.4 Models and Training

2.4.1 CNN Architecture

The classification model used in this study is based on a CNN, specifically designed to handle the image-like structure of atmospheric data (SLP and z500). The CNN architecture (Figure 2.2) comprises:

1. Input Layer: Two separate input channels corresponding to SLP and z500.
2. Convolutional Layers: Two convolutional layers applied to extract spatial patterns and features from input atmospheric fields. These layers use convolutional kernels to identify local and spatially coherent patterns.
3. Dropout Layer: Included for regularization to mitigate overfitting by randomly dropping out units during training, thus improving the model's generalization.
4. Fully Connected Layers: Two fully connected layers following the convolutional and dropout layers, used for integrating learned spatial features into final predictions of HB CTs.
5. Output Layer: Outputs softmax probabilities for each of the 29 HB CTs, which are then converted into classifications.

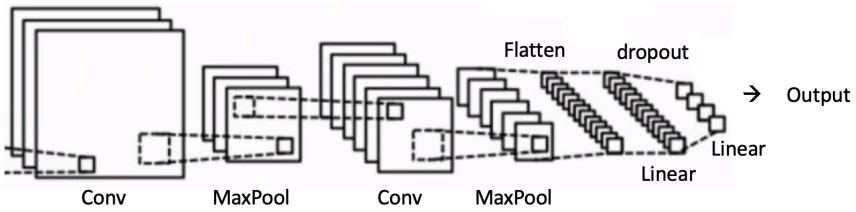


FIGURE 2.2: Schematic diagram of the CNN network

2.4.2 Training Protocol

2.4.2.1 Adam Optimizer

Adam (Adaptive Moment Estimation) (Mehta et al., 2019) is an optimization algorithm widely used for training neural networks. It combines the benefits of two other methods—AdaGrad (Ward et al., 2020), which adapts the learning rate for each parameter, and RMSProp (Zou et al., 2019), which considers the exponentially decaying average of squared gradients—to adjust the learning rates adaptively during training.

Adam optimizer efficiently handles noisy and sparse gradients, common in deep learning tasks, facilitating faster convergence and better handling of large parameter spaces typical of CNNs. This makes Adam particularly suitable for our CNN-based classification model, ensuring effective and efficient training.

2.4.2.2 Bayesian Optimization for Hyperparameter Tuning

To achieve an optimal configuration of hyperparameters for our CNN, we employed Bayesian optimization (Garnett, 2023), specifically using the Tree-structured Parzen Estimator (TPE) method (Ozaki et al., 2022) implemented in the Optuna library. Bayesian optimization is particularly suited for deep learning hyperparameter tuning due to its efficiency and ability to intelligently balance exploration and exploitation of the hyperparameter space, thus significantly reducing computational cost compared to grid or random search.

Specifically, the following hyperparameters were optimized:

1. Learning Rate: The learning rate controls the magnitude of updates to the network weights during training. We allowed this hyperparameter to vary logarithmically between 10^{-4} and 10^{-2}, enabling the optimizer to effectively explore both subtle and more substantial updates in network parameters.
2. Weight Decay: Serving as a regularization term to prevent overfitting, the weight decay was also sampled on a logarithmic scale ranging from 10^{-5} to 10^{-3}. This enabled fine control over model complexity and encouraged better generalization.
3. Dropout Rate: Dropout prevents overfitting by randomly dropping neurons during training, forcing the model to learn robust representations. We tuned dropout rate uniformly within a range from 0.2 to 0.6, allowing optimization of the model's regularization intensity.
4. Convolutional Layer Parameters:
    - Number of Output Channels in First Convolutional Layer (out_channels1): Chosen categorically from [4, 8, 16], optimizing the depth and feature extraction capability of initial convolutional operations.
    - Number of Output Channels in Second Convolutional Layer (out_channels2): Selected from [8, 16, 32], further refining the network's capability to learn hierarchical spatial patterns.
    - Kernel Size: Kernel size impacts the spatial context the model considers when learning features. We considered kernel sizes of [3, 5, 7], balancing between capturing detailed local patterns and broader spatial structures.
    - Fully Connected Layer Size: The number of neurons in the fully connected layer was chosen from [32, 64, 128], directly influencing the model's ability to integrate learned spatial features into higher-level abstractions suitable for classification.

The Bayesian optimization was performed over 100 trials, systematically exploring these hyperparameters. Each trial evaluated a distinct combination of parameters based on validation accuracy. The final selection of hyperparameters was determined by the combination that produced the highest validation accuracy, ensuring the CNN model's optimal performance on unseen data while minimizing the computational burden associated with hyperparameter tuning.

2.4.2.3 Nested Cross-Validation

Nested cross-validation (Zhong et al., 2023) involves two cross-validation loops:

1. Outer Loop: Evaluates the generalization performance of the model on independent test sets.
2. Inner Loop: Used for hyperparameter tuning and model selection within each training fold.

We use 5 folds of outer loop and 5 folds of inner loop.

2.4.2.4 Epochs and Early Stopping

Training was set to a maximum of 35 epochs, with early stopping criteria implemented, patience of 6 epochs without improvement, to prevent overfitting and reduce training time.

2.4.3 Predicting Technique

2.4.3.1 Ensemble Learning

Ensemble learning (Dietterich et al., 2002) combines predictions from multiple models to enhance predictive performance. In this case, multiple CNN models trained with different initial conditions. A deep ensemble of 30 independently trained CNN models, each initialized with different random weights, is utilized.

Predictions from individual CNN models are aggregated using a weighted averaging method, where each model's contribution is proportional to its validation performance. This enhances robustness, reduces prediction uncertainty, and provides stable and reliable classification results by capturing a wider range of atmospheric variability.

2.4.3.2 Transitional Smoothing

According to the definition, a HB CT must last at least three days (Hess and Brezowsky, 1952) to represent a stable atmospheric pattern. Without enforcing this constraint, the raw neural network predictions may contain short, noisy transitions that are meteorologically implausible. Thus, transition smoothing is necessary to:

1. Remove unrealistic, very short-lived classifications.
2. Produce physically consistent and stable time series of HB CTs.
3. Ensure that predictions align with domain-specific expectations.

According to the original paper:

1. Step 1: Identify all transitions where the predicted HB CT lasts fewer than three days.
2. Step 2: Check neighborhood consistency:
    - If the HB CT before and after the short transition is the same, assign this type to the transition days.

3. Step 3: If different types occur before and after:
    - Compare the predicted probabilities (confidence scores) for each neighboring type.
    - Assign the transition days to the neighbor with the higher predicted probability (i.e., stronger membership).

This method ensures that short transitions are corrected in a physically meaningful way, based on both label consistency and model confidence. Since the original code is not available, our transition_smoothing function follows the paper's logic but introduces minor adaptations for greater robustness and flexibility:

TABLE 2.2: Transitional smoothing implementation

Aspect	Paper Description	Our Code Implementation	Reason for Change
Minimum duration	3 days fixed	min_duration parameter (default=3)	Makes the method flexible for other settings.
Neighbor block length trust	Not specified clearly	min_neighbor_run_length parameter (default=2)	Ensures smoothing only when neighboring segments are stable enough.
Handling first and last segments	Implicitly ignored	Explicitly handled	Avoids out-of-bound errors at sequence edges.
Probability averaging	Single-point comparison	Average probability across transition segment	More robust by considering entire transition region, not a single timestep.
Fallback behavior when neighbors are not strong	Not detailed	Fallback to the only trusted neighbor if one side is strong enough	Avoids wrong smoothing when one neighbor is unreliable.

2.5 Results

2.5.1 Classification Performance

2.5.1.1 F1-score

The classification performance of our model was evaluated using a confusion matrix, F1-scores per class, macro F1-score (Opitz and Burst, 2019), and overall accuracy (Alberg et al., 2004), following the approach used in the reference study (Mittermeier et al., 2022). The overall accuracy achieved by our model on the cross-validation test set is 53.17%, and the macro F1-score is 55.14%. Among all HB CTs (Appendix Figure 2.9), the best performing classes in terms of F1-score are WZ (70.21), HM (63.7), and HB (62.78). These types exhibit distinct spatial patterns, aiding CNN identification. The lowest F1-scores are observed for classes like NA (33.61) and SEZ (44.37). These are likely due to class imbalance caused by rare occurrence. Generally, accuracy for each type is quite high, which may imply overfitting.

The confusion matrix (Appendix Figure 2.17) shows that most misclassifications occur among HB CTs with similar dynamical structures, consistent with expectations from atmospheric science. Diagonal dominance is visible but moderate, reflecting the challenging nature of this classification task.

Compared to the paper results, which reported an overall accuracy of 41.1% and a macro F1-score of 38.3% (Mittermeier et al., 2022), our model achieves slightly higher overall accuracy and macro F1-score.

TABLE 2.3: Metric comparison

Metric	Paper (%)	Our Model (%)
Macro F1-score	38.3	55.1
Overall accuracy	41.1	53.2

The boxplots of F1-score under each HB CT (Figure 2.6; Appendix Figure 2.13) confirm our conclusion:

1. Original study has considerably lower overall F1-scores (around 0.3–0.5), indicating challenges in reliable classification.
2. Our model demonstrates much higher stability with F1-scores generally between 0.80–0.95, suggesting improved reliability and robustness in identifying circulation patterns.

2.5.1.2 RMSE

To further evaluate the quality of the predictions, the Root-Mean-Square Error (RMSE) between the predicted HB CT composites and the true label composites was calculated for variable SLP.

Equation for the calculation of the RMSE with I being the predicted image (in our case: signature plot of the deep learning classifier) and K being the reference image (in our case: signature plot of the labels). M are number of rows and N the number of columns of the pictures to compare. The RMSE thus compares the pixel-wise values of two images. A value of zero indicates a perfect match (Mittermeier et al., 2022; Mueller et al., 2020).

$$\mathrm{RMSE} = \sqrt{\frac{1}{M \times N} \sum_{i=0,j=0}^{M-1,N-1} [I(i,j) - K(i,j)]^2}$$

Additionally, RMSEs for false positives and false negatives were computed separately.

The average RMSE for our implementation's predictions across all HB CTs is 0.88, while the false positives and false negatives yielded RMSEs of 1.08 and 1.23, respectively.

TABLE 2.4: RMSE comparison

RMSE Category	Paper	Implementation
Prediction RMSE	0.89	0.74
False positives RMSE	1.09	1.45
False negatives RMSE	1.28	0.69

Our implementation achieves a lower prediction RMSE (0.74) compared to the original paper (0.89) (Mittermeier et al., 2022). This indicates that, on average, the spatial patterns of the correctly predicted HB CTs are more similar to the true labels in our model than in the reference study.

However, the false positives RMSE is much higher (1.45) than in the paper (1.09) (Mittermeier et al., 2022). This suggests that when our implementation wrongly predicts a HB CT, the resulting spatial pattern is less similar to the true pattern than in the paper.

Our implementation shows a lower false negatives RMSE (0.69) compared to the paper (1.28) (Mittermeier et al., 2022). This means that when the model misses a true HB CT, the spatial signature remains relatively closer to the correct pattern than in the paper.

The overall pattern suggests that:

1. Our implementation is better at maintaining correct structures when making predictions or missing true types.
2. However, when making incorrect predictions (false positives), our implementation model tends to produce worse distortions than the model used in the paper.

2.5.2 Signature Plot Analysis

Figure 2.3 shows the signature plots for four HB CTs selected based on our RMSE results:

1. Two HB CTs with lowest RMSE: WZ and TRW (best performance)
2. Two HB CTs with highest RMSE: NA and HNFA (worst performance)

For each HB CT, four different composite plots are displayed:

1. Labels: Composite of true labels.
2. Predictions: Composite of model predictions.
3. False Positives: Days predicted as the HB CT but labeled differently.
4. False Negatives: Days labeled as the HB CT but predicted differently.

RMSE values are shown below each panel, comparing each composite to the true label composite.

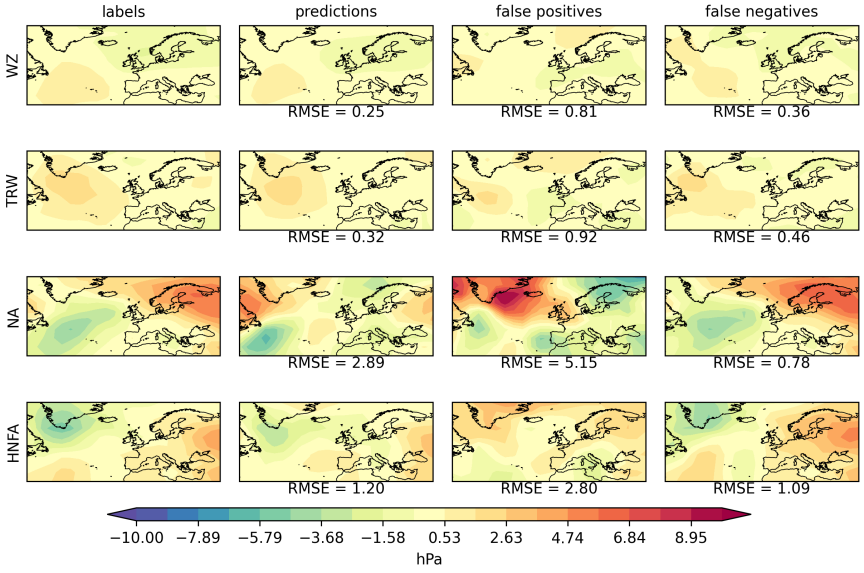


FIGURE 2.3: Signature plots of four selected circulation patterns at slp. Column 1: labels showing the indicated HB CT, Column 2: deep ensemble predictions showing the indicated HB CT, Column 3: signature pattern, when the deep ensemble predicts the indicated HB CT while labels state differently, Column 4: labels stating the indicated HB CT while deep ensemble predicts differently. The RMSE values are calculated by comparing the respective signature plot to the signature plot of the labels (column 1).

Based on our results, we can observe that:

1. WZ:
    - The predicted composite closely matches the label pattern (RMSE = 0.25).
    - False positives and false negatives show moderate RMSE (0.81 and 0.36), indicating stable identification of WZ.

2. TRW:
    - Predictions for TRW are also accurate (RMSE = 0.32).
    - False positives (0.92) and false negatives (0.46) suggest TRW is generally recognized reliably.
3. NA:
    - Predictions for NA have high RMSE (2.89), indicating difficulty in capturing the correct structure.
    - False positives (5.15) show very large errors, and false negatives (0.78) suggest missing NA patterns is common.
4. HNFA:
    - Predictions for HNFA also show high RMSE (1.20).
    - False positives (2.80) and false negatives (1.09) are large, implying HNFA is a difficult type for the model to classify.

Differences from the original paper (Appendix Figure 2.10) arise primarily from data variations, model differences, and random initialization.

Although the specific CTs selected differ, the general trend remains consistent with the findings of the paper:

1. Some HB CTs are easier to predict (low RMSE), others are harder (high RMSE).
2. False positives tend to have higher RMSE than correct predictions.
3. A larger discrepancy in the color distribution indicates a higher complexity of HB CT, and false negatives vary depending on HB CT complexity.

Thus, while specific HB CTs differ, our results fundamentally agree with the conclusions of the paper that certain patterns are systematically easier or harder to classify, and signature plots are effective to visualize these tendencies.

2.5.3 Frequency Distribution of HB CTs

Figure 2.4 shows the frequency distribution of the 29 HB CTs, expressed as the average number of days per year for the period 1950–1980. For each HB CT, the frequency derived from the true labels (blue bars) is compared to the frequency derived from the network predictions (yellow bars).

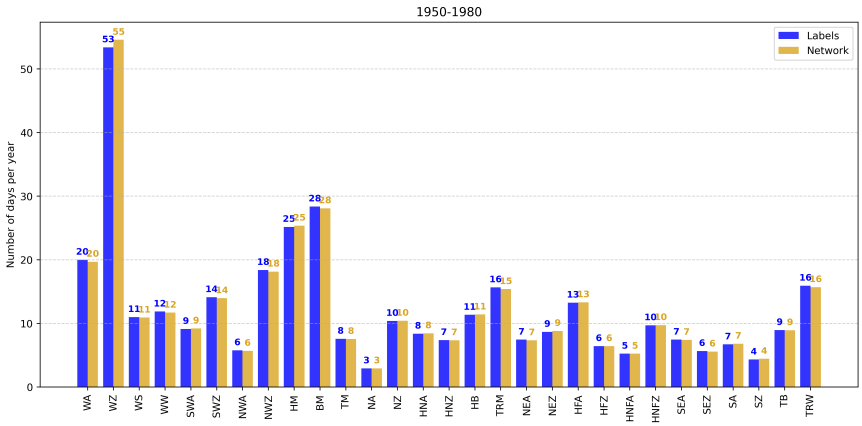


FIGURE 2.4: Frequency distribution of the 29 HB CTs in number of days per year for the training period 1950–1980 (blue) and for predictions of the CNN (yellow).

Overall, the CNN reliably reproduces observed HB CT frequencies:

1. Dominant HB CTs (e.g., WZ, HM) are captured accurately, reflecting robust model learning.
2. Rarer HB CTs (e.g., NA, SZ) exhibit minor deviations, suggesting challenges in accurately learning patterns due to fewer examples, which may indicate slight overfitting.

In comparison with the original study (Appendix Figure 2.11), despite differences in dataset periods (our study: 1950–1980; original: 1900–1980), frequency patterns remain consistent. Slight discrepancies in rarer HB CTs are likely due to dataset differences, emphasizing the importance of extended or more balanced training datasets for better generalization.

The agreement between the label frequencies and the network predictions, as well as the similarity to the paper's findings, indicates that our model effectively learns the relative occurrence rates of the different HB CTs. Overall, the frequency distribution results support the conclusions of the original study, confirming that the network can reliably reproduce the occurrence rates of various HB CTs.

2.5.4 Future Change

Figure 2.5 and Figure 2.6 show the relative changes in the frequency of occurrence of the 29 HB CTs between the future period (2031–2060) and the historical reference period (1971–2000), for the entire year, winter half-year, and summer half-year, based on 12 climate model realizations. Figure 2.5 is the conclusion directly from predicted results, while Figure 2.6 is obtained after transition smoothing. Positive values indicate an increase in occurrence; negative values indicate a decrease. The lower panels show the F1-scores of the classification model ensemble. Higher F1-scores indicate better classification stability and reliability. The grouping by wind direction allows visual comparison of performance across different HB CT categories.

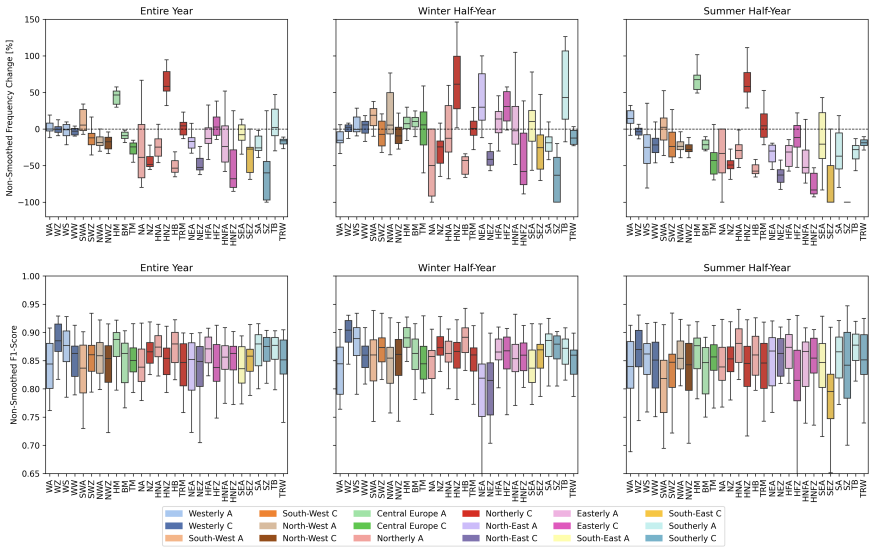


FIGURE 2.5: Boxplots of frequency change and f1 score without transitional smoothing. Upper plots show the change in the relative frequency of occurrence (%) of the HB CTs between the future 2031–2060 and the reference period 1971–2000 for the entire year, the winter half-year (ONDJFM) and the summer half-year (AMJJAS). Lower plots illustrate the spread of F1-scores.

Without transition smoothing (Figure 2.5):

1. The boxplots show relatively smaller spreads for most HB CTs.
2. Most relative changes are within ±50%.
3. Fewer extreme outliers are observed, and the whiskers are relatively short.
4. The overall signal is dominated by noise, making it hard to detect consistent climate change trends.

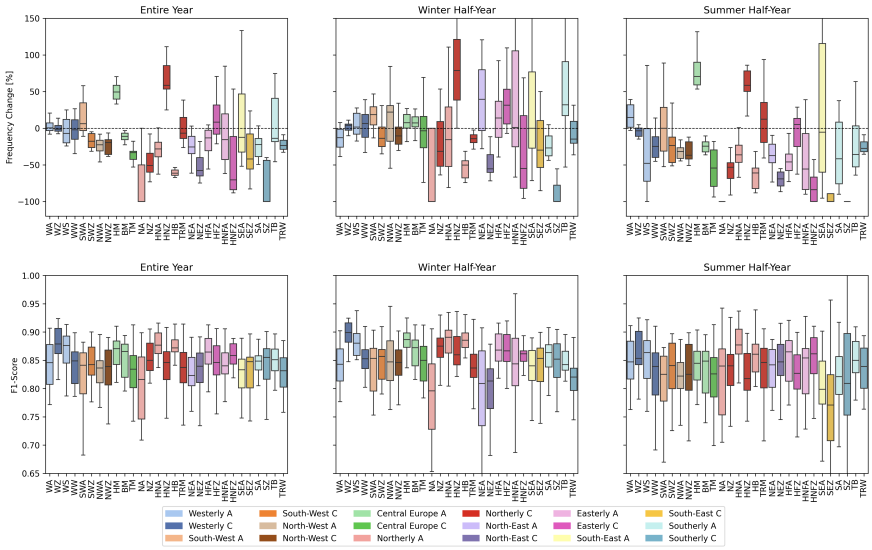


FIGURE 2.6: Boxplots of frequency change and f1 score after transitional smoothing. Upper plots show the change in the relative frequency of occurrence (%) of the HB CTs between the future 2031–2060 and the reference period 1971–2000 for the entire year, the winter half-year (ONDJFM) and the summer half-year (AMJJAS). Lower plots illustrate the spread of F1-scores.

With transition smoothing (Figure 2.6):

1. The boxplots show larger spreads, particularly for rare HB CTs.
2. Relative changes frequently extend beyond ±50%, and even approach ±100% in some cases.
3. More extreme outliers appear.
4. F1-score distributions are lower and broader, indicating less stable classification performance.

Overall, transition smoothing, while theoretically beneficial, introduces higher internal variability in our implementation, complicating clear signal identification. Variability differences from the original study stem from methodological factors (ensemble size, smoothing implementation) and underlying climate model characteristics.

Our findings reaffirm the value of a multi-ensemble approach and careful consideration of methodological choices when projecting future circulation trends, underscoring the need for robust ensemble analysis techniques to better discern climate-driven circulation changes from internal variability.

The original study (Appendix Figure 2.13) shows relatively controlled variability, mostly within ±50% (Mittermeier et al., 2022), and fewer extreme outliers, indicating relatively stable and moderate shifts in HB CTs. Our results (Figure 2.5) exhibit larger variability, particularly notable in specific HB CTs such as NA, SEA, HNFZ, and SZ, with frequency changes frequently exceeding ±50%, even reaching ±100%.

1. Entire Year:
   - Original study shows moderate shifts centered around zero, with clear signals for types like WA and SEA (Mittermeier et al., 2022).
   - Our implementation shows notably higher variability with clear signals (e.g., strong increase in HNZ, HM and significant decreases in HB and NEZ).
2. Winter Half-Year:
   - Original study depicts clear moderate increases for WW and HFA with relatively smaller variability (Mittermeier et al., 2022).

- Our data demonstrate significant changes with pronounced variability and stronger extremes, especially in HNZ and rare CTs like NA and SZ.

3. Summer Half-Year:
   - Original data indicate mild variability and clear, moderate increases and decreases in specific types (e.g., decreases in HNFZ, increases in WA) (Mittermeier et al., 2022).
   - Our results again illustrate increased variability with substantial increases (e.g., HNZ, HM) and decreases (e.g., HB, NEZ).

In both cases, the summer half-year has a greater influence on the entire year, despite the winter half-year showing higher variability. Rare CTs such as NA and SZ are observed with significant changes. It would be interesting to analyze how rare conditions evolve, however, due to the unbalanced data, the results are not very reliable.

Key CT Differences:

1. WA (Westerly Anticyclonic):
   - Our results show only slight increases, with less pronounced signals for WA, especially compared to significant changes in other types.
   - Original study consistently shows WA increasing clearly across all seasons (Mittermeier et al., 2022).
2. HNZ (Icelandic High Cyclonic):
   - In our results, HNZ consistently shows marked increases, particularly pronounced in the summer and winter half-years.
   - Original study presents less dramatic decreasing changes for HNZ, reflecting a critical methodological or scenario-driven difference (Mittermeier et al., 2022).
3. NA and SZ (Northerly Anticyclonic and Southerly Cyclonic):
   - Our results exhibit significantly larger variability and extreme reductions in NA, while SZ shows extreme negative changes particularly in summer.
   - Original study maintains relatively modest decreasing changes for these types (Mittermeier et al., 2022).

The differences in projected frequency changes between our results and the original study primarily stem from:

1. Time period difference (2031-2060 vs. 2071-2100).
2. Scenario differences: CanESM2 model (historical) in our study vs. original SMHI-LENS (SSP3-RCP7.0).
3. Model and methodological distinctions: Implementation differences (e.g., transitional smoothing, dataset variations, and resolution).
4. Ensemble size effects: Our smaller ensemble size (12 vs. original 50) could amplify internal variability signals.

These differences underline the sensitivity of CT analysis to methodological choices, highlighting the importance of transparent documentation of scenario assumptions and model details to ensure the accurate interpretation and comparability of future climate studies.

2.6 Conclusion and Outlook

In this study, we successfully implemented a CNN-based deep learning approach to classify large-scale CTs over Europe, building upon the established HB framework. Using high-resolution ERA5 data for model

training, our method achieved an overall accuracy of 53.17% and a macro F1-score of 55.14%, clearly outperforming traditional classification approaches.

The CNN model effectively learned dominant circulation patterns, yet faced challenges classifying rare or inherently complex CTs. These limitations indicate potential issues of class imbalance and inherent labeling uncertainties associated with manual classifications. One potential improvement to the statistical analysis would be reducing the number of subdivisions of CTs. Subdividing CT according to the wind direction of the jet stream, for example, could increase the robustness of the results. Additionally, the original study found that the most frequent misclassifications occurred between pairs of anticyclonic and cyclonic circulation types (Mittermeier et al., 2022), which could be addressed through refined classification methods. Furthermore, future research should be conducted to quantify human-level errors in labelling to better understand the impacts of manual classification biases.

Besides the primary training session and results shown in this report, we conducted another model training experiment using a narrower range of hyperparameters. The results, which are illustrated in the presentation, demonstrated noticeably different classification performance compared to the setup we introduced here. In particular, I suspect that the size of the fully connected linear layers is the mean reason for this difference. Since in earlier experiments with narrower range of hyperparameters, we set the fully connected layer size as fixed 50 dimensions, however, in the latest optimization round, where more extensive hyperparameter tuning were applied, the model consistently favored a dimension size of 128, which leads to high accuracy for each HB CTs. Therefore, the model shows good performance on the current dataset, it is important to note that this does not necessarily guarantee the absence of overfitting.

Due to limited computational resources, we were unable to perform a more thorough investigation to validate these observations or further optimize the model architecture. Future work should focus on systematically assessing the impact of fully connected layer sizes and further regularization strategies to ensure model robustness and generalization.

The analysis of frequency changes in both our implementation and original paper reveals that for most HB CTs, the changes fall within a range of ±5 days (relatively ±50%) (Appendix Figure 2.12; Figure 2.6; Appendix Figure 2.14; Appendix Figure 2.13). In the original study, the most significant changes in frequency were observed in the WA circulation type (Appendix Figure 2.13) (Mittermeier et al., 2022), while in the own-built model, the HM and HNZ circulation types experienced the largest shifts in frequency. Notably, our results differ from previous studies, primarily due to distinct methodological choices such as dataset selection, scenario assumptions, and smoothing techniques. These variations underline the sensitivity of HB CTs analysis and stress the importance of transparent methodological reporting.

When focusing only on the frequency of HB CTs, without considering their duration, it is important to highlight that the duration of a HB CT significantly influences the amplitude of weather extremes. This underscores the need to integrate both frequency and duration for a more comprehensive understanding of weather patterns. Duration of a HB CT impacts the amplitude of a weather extreme.

To further evaluate the uncertainties in frequency changes, it would be beneficial to combine multi-model and single-model ensembles under different forcing scenarios. This approach could help assess the uncertainties arising from various climate models and the assumptions in different forcing scenarios.

For future studies, we recommend quantifying human-level errors in labeling to understand better the impacts of manual classification biases. Additionally, models that explicitly capture temporal continuity, such as Deep Hidden Markov Models (Yu et al., 2015) or temporal-aware CNN architectures (e.g., ConvLSTM (Moishin et al., 2021)), could enhance prediction accuracy by directly modeling the persistence characteristic inherent in circulation patterns (Mittermeier et al., 2022).

2.7 Appendix

2.7.1 Statement

The implementation was carried out using Python version 3.9.18.

2.7.2 Figures

Synoptic patterns of the 29 HB CTs using ERA20C (Figure 2.7) (Mittermeier et al., 2022)

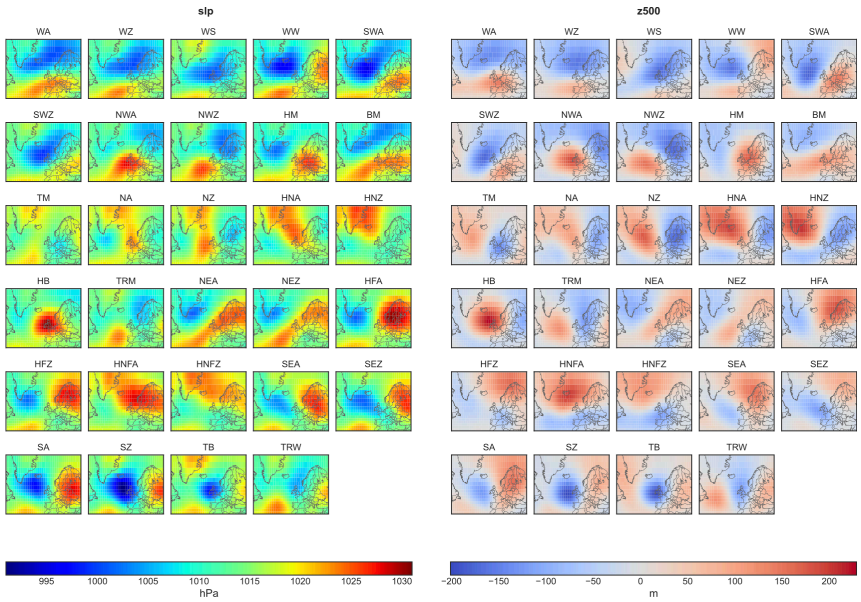


FIGURE 2.7: Synoptic patterns of the 29 HB CTs created using ERA-20C data averaged over the period 1900 -1980, showing SLP on the left and z500 on the right side.

HB CTs (Figure 2.8)

F1 score comparison table (Figure 2.9)

Signature plots of paper - SLP (Figure 2.10) (Mittermeier et al., 2022)

Frequency distribution of paper (Figure 2.11) (Mittermeier et al., 2022)

Boxplot of absolute future change and F1 score of our implementation (Figure 2.12)

Boxplot of future change and F1 score of paper (Figure 2.13) (Mittermeier et al., 2022)

Boxplot of absolute future change and F1 score of paper (Figure 2.14) (Mittermeier et al., 2022)

RMSE table of our implementation - SLP (Figure 2.15)

RMSE table of paper - SLP (Figure 2.16)

Confusion matrix of our implementation (Figure 2.17)

Confusion matrix of paper (Figure 2.18) (Mittermeier et al., 2022)

Full signature plot of our implementation - SLP (Figure 2.19)

Full signature plot of paper - SLP (Figure 2.20) (Mittermeier et al., 2022)

Acronym	Original name (German)	Translated name (English)
WA	Westlage, antizyklonal	Anticyclonic Westerly
WZ	Westlage, zyklonal	Cyclonic Westerly
WS	Südliche Westlage	South-Shifted Westerly
WW	Winkelförmige Westlage	Maritime Westerly (Block E. Europe)
SWA	Südwestlage, antizyklonal	Anticyclonic North-Westerly
SWZ	Südwestlage, zyklonal	Cyclonic South-Westerly
NWA	Nordwestlage, antizyklonal	Anticyclonic North-Westerly
NWZ	Nordwestlage, zyklonal	Cyclonic North-Westerly
HM	Hoch Mitteleuropa	High over Central Europe
BM	Hochdruckbrücke (Rücken) Mitteleuropa	Zonal Ridge across Central Europe
TM	Tief Mitteleuropa	Low (Cut-Off) over Central Europe
NA	Nordlage, antizyklonal	Anticyclonic Northerly
NZ	Nordlage, zyklonal	Cyclonic Northerly
HNA	Hoch Nordmeer-Island, antizyklonal	Icelandic High, Ridge C. Europe
HNZ	Hoch Nordmeer-Island, zyklonal	Icelandic High, Trough C. Europe
HB	Hoch Britische Inseln	High over the British Isles
TRM	Trog Mitteleuropa	Trough over Central Europe
NEA	Nordostlage, antizyklonal	Anticyclonic North-Easterly
NEZ	Nordostlage, zyklonal	Cyclonic North-Easterly
HFA	Hoch Fennoskandien, antizyklonal	Scandinavian High, Ridge C. Europe
HFZ	Hoch Fennoskandien, zyklonal	Scandinavian High, Trough C. Europe
HNFA	Hoch Nordmeer-Fennoskandien, antizykl.	High Scandinavia-Iceland, Ridge C. Europe
HNFZ	Hoch Nordmeer-Fennoskandien, zyklonal	High Scandinavia-Iceland, Trough C. Europe
SEA	Südostlage, antizyklonal	Anticyclonic South-Easterly
SEZ	Südostlage, zyklonal	Cyclonic Southerly
SA	Südlage, antizyklonal	Anticylonic Southerly
SZ	Südlage, zyklonal	Cyclonic Southerly
TB	Tief Britische Inseln	Low over the British Isles
TRW	Trog Westeuropa	Trough over Western Europe

FIGURE 2.8: HB CTs

	WA	WZ	WS	WW	SWA	SWZ	NWA	NWZ	HM	BM	TM	NA	NZ	HNA	HNZ
Implementation	57.19	70.21	59.03	55.88	51.34	57.18	50.17	53.26	63.70	55.54	52.25	33.61	58.27	60.89	51.54
Paper	44.60	47.08	45.39	37.70	35.36	30.86	38.88	37.07	51.24	47.29	37.23	24.85	44.32	45.57	27.11

	HB	TRM	NEA	NEZ	HFA	HFZ	HNFA	HNFZ	SEA	SEZ	SA	SZ	TB	TRW	Macro
Implementation	62.78	52.30	52.02	51.16	60.57	51.99	55.17	57.32	47.69	44.37	60.27	57.72	59.81	55.87	58.09
Paper	50.99	27.86	41.44	33.12	45.32	24.81	33.35	34.02	38.09	37.93	39.84	38.19	42.11	29.34	38.30

FIGURE 2.9: Macro F1

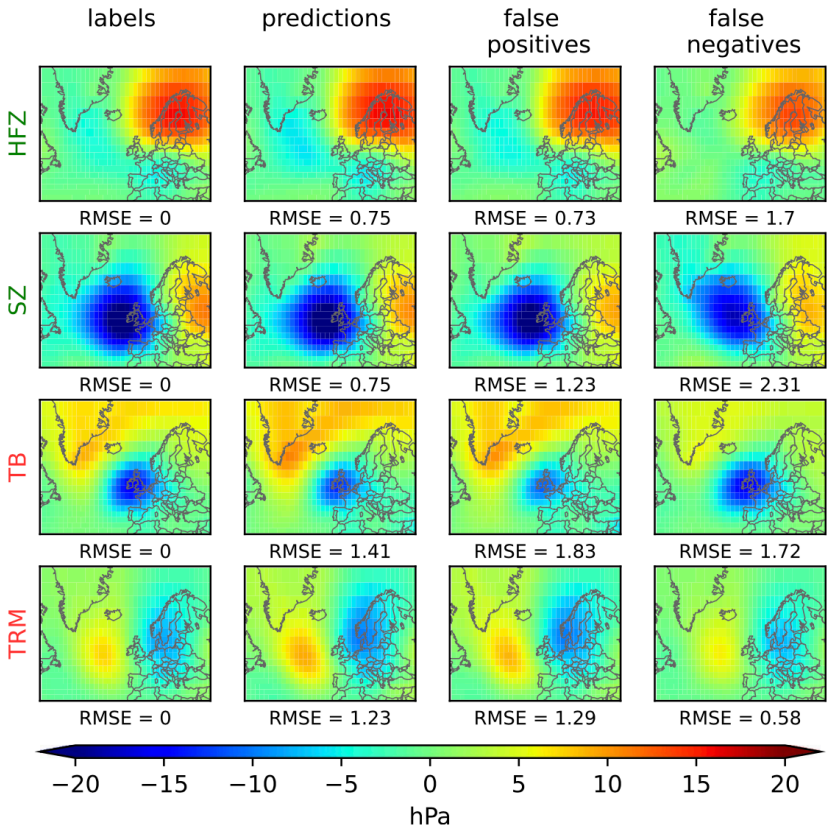


FIGURE 2.10: Signature plots of four selected circulation patterns at slp. Column 1: labels showing the indicated HB CT, Column 2: deep ensemble predictions showing the indicated HB CT, Column 3: signature pattern, when the deep ensemble predicts the indicated HB CT while labels state differently, Column 4: labels stating the indicated HB CT while deep ensemble predicts differently. The RMSE values are calculated by comparing the respective signature plot to the signature plot of the labels (column 1). The four HB CTs are chosen as positive (green) and negative (red) examples.

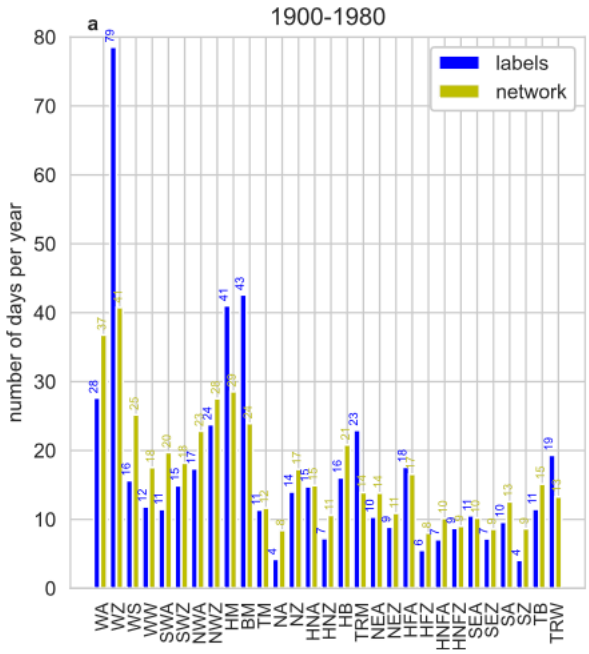


FIGURE 2.11: Frequency distribution of the 29 HB CTs in number of days per year for the training period 1900–1980 (blue) and for predictions of the CNN (yellow).

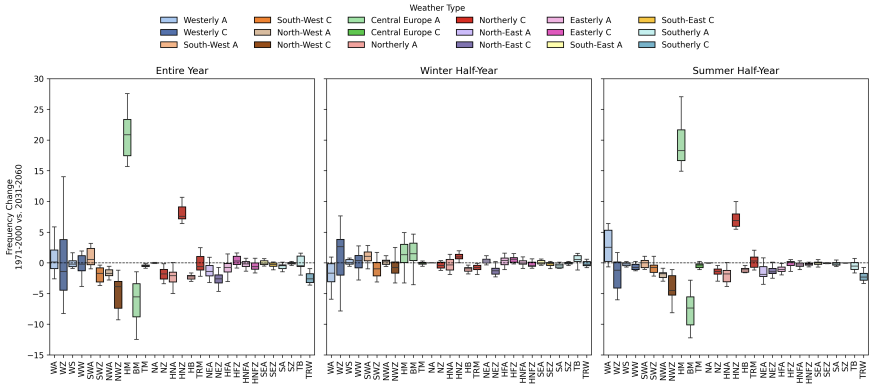


FIGURE 2.12: Boxplots of frequency change and f1 score. Upper plots show the change in the absolute frequency of occurrence (days) of the HB CTs between the future 2031–2060 and the reference period 1971–2000 for the entire year, the winter half-year (ONDJFM) and the summer half-year (AMJJAS). Lower plots illustrate the spread of F1-scores.

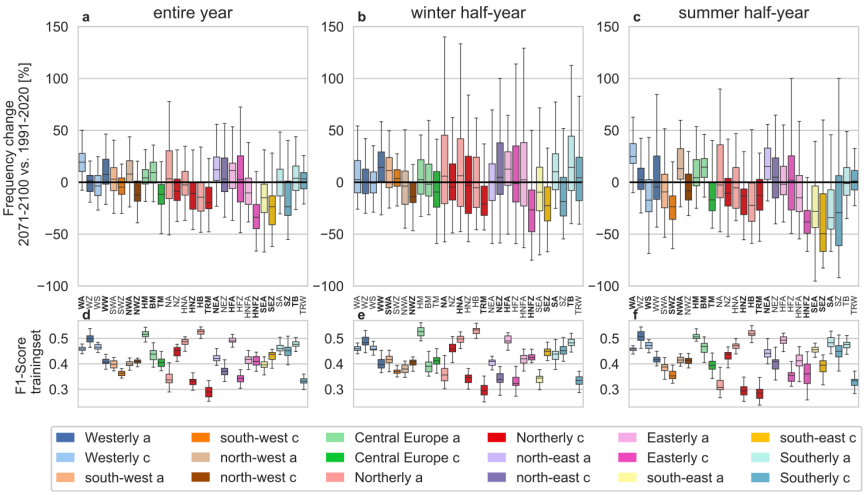


FIGURE 2.13: Boxplots of frequency change and f1 score. Upper plots show the change in the relative frequency of occurrence (%) of the HB CTs between the future 2031–2060 and the reference period 1971–2000 for the entire year, the winter half-year (ONDJFM) and the summer half-year (AMJJAS). Lower plots illustrate the spread of F1-scores.

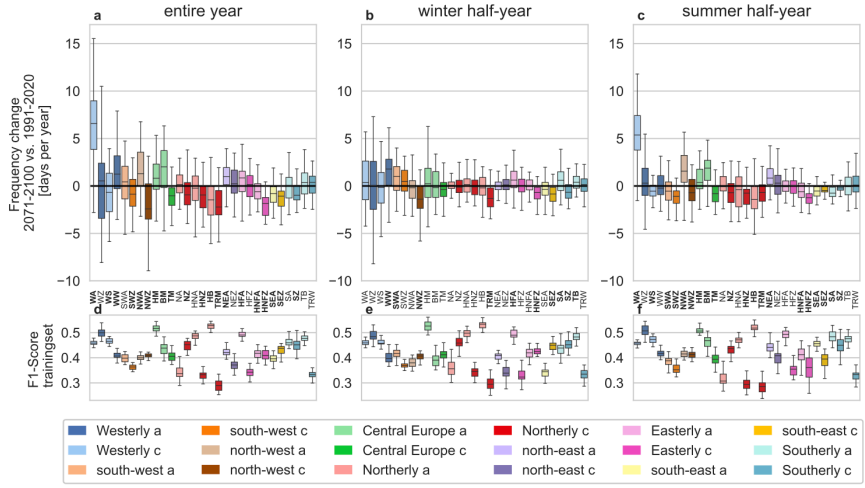


FIGURE 2.14: Boxplots of frequency change and f1 score. Upper plots show the change in the absolute relative frequency of occurrence (days) of the HB CTs between the future 2031–2060 and the reference period 1971–2000 for the entire year, the winter half-year (ONDJFM) and the summer half-year (AMJJAS). Lower plots illustrate the spread of F1-scores.

RMSE	WA	WZ	WS	WW	SWA	SWZ	NWA	NWZ	HM	BM	TM	NA	NZ	HNA	HNZ
Pred	0.39	0.25	0.47	0.54	0.48	0.51	1.17	0.51	0.40	0.52	0.90	2.89	0.54	0.56	0.70
FP	0.88	0.81	1.49	0.91	0.83	1.28	1.34	1.00	0.92	0.93	1.54	5.15	0.96	1.04	1.22
FN	0.50	0.36	0.73	0.55	0.39	0.83	1.01	0.49	0.36	0.34	0.83	0.78	0.66	0.92	0.71

RMSE	HB	TRM	NEA	NEZ	HFA	HFZ	HNFA	HNFZ	SEA	SEZ	SA	SZ	TB	TRW	Ø
Pred	0.50	0.57	0.72	1.07	0.69	0.86	1.20	0.68	0.70	1.10	0.71	1.01	0.64	0.32	0.74
FP	1.30	1.04	1.33	1.13	1.55	1.88	2.80	1.36	1.25	1.85	1.64	2.42	1.20	0.92	1.45
FN	0.62	0.33	0.54	1.16	0.50	0.71	1.09	0.88	0.76	0.56	1.04	1.10	0.83	0.46	0.69

FIGURE 2.15: RMSE - slp

RMSE	WA	WZ	WS	WW	SWA	SWZ	NWA	NWZ	HM	BM	TM	NA	NZ	HNA	HNZ
Pred	0.93	0.82	1.00	0.88	0.77	0.82	0.73	1.06	0.78	1.04	0.91	0.87	0.39	0.85	1.29
FP	0.78	1.11	1.21	1.08	1.23	1.00	0.83	1.06	0.80	0.84	0.92	0.92	0.89	1.41	1.41
FN	1.25	0.84	2.10	1.42	1.62	1.24	0.83	1.36	0.87	1.12	1.03	1.50	0.80	1.04	1.30

RMSE	HB	TRM	NEA	NEZ	HFA	HFZ	HNFA	HNFZ	SEA	SEZ	SA	SZ	TB	TRW	Ø
Pred	0.66	1.23	0.58	0.89	0.89	0.75	0.74	0.66	0.99	0.91	1.11	0.75	1.41	1.14	0.89
FP	1.11	1.29	0.76	0.96	1.41	0.73	0.84	0.94	0.98	1.22	1.69	1.23	1.83	1.27	1.09
FN	1.57	0.58	1.00	0.99	1.25	1.70	1.99	1.00	1.31	1.58	1.32	2.31	1.72	0.61	1.28

FIGURE 2.16: RMSE - slp

	WA	WZ	WS	WW	SWA	SWZ	NWA	NWZ	HM	BM	TM	NA	NZ	HNA	HNZ	HB	TRM	NEA	NEZ	HFA	HFZ	HNFA	HNFZ	SEA	SEZ	SA	SZ	TB	TRW	∑	Precision
WA	350.0	101.0	0.0	3.0	11.0	3.0	10.0	19.0	29.0	71.0	0.0	1.0	0.0	3.0	0.0	1.0	8.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0	620.0	0.58
WZ	80.0	1260.0	36.0	28.0	15.0	33.0	2.0	59.0	19.0	35.0	3.0	0.0	8.0	1.0	3.0	0.0	36.0	0.0	2.0	1.0	1.0	0.0	2.0	0.0	0.0	0.0	1.0	7.0	23.0	1655.0	0.65
WS	0.0	83.0	183.0	2.0	1.0	16.0	0.0	6.0	2.0	1.0	9.0	0.0	1.0	3.0	3.0	0.0	2.0	0.0	1.0	0.0	2.0	0.0	8.0	1.0	1.0	0.0	2.0	7.0	7.0	341.0	0.66
WW	2.0	54.0	4.0	209.0	5.0	9.0	0.0	2.0	5.0	16.0	1.0	0.0	1.0	2.0	0.0	0.0	5.0	0.0	4.0	9.0	6.0	1.0	3.0	2.0	5.0	1.0	4.0	8.0	10.0	368.0	0.55
SWA	15.0	37.0	1.0	5.0	134.0	22.0	0.0	0.0	35.0	13.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	4.0	1.0	0.0	0.0	0.0	0.0	5.0	0.0	5.0	5.0	283.0	0.56
SWZ	6.0	74.0	6.0	14.0	28.0	239.0	0.0	5.0	10.0	8.0	0.0	0.0	1.0	0.0	2.0	0.0	2.0	0.0	0.0	0.0	3.0	0.0	2.0	0.0	1.0	3.0	7.0	7.0	20.0	438.0	0.6
NWA	20.0	5.0	0.0	0.0	0.0	0.0	74.0	19.0	14.0	23.0	0.0	1.0	3.0	2.0	0.0	15.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	179.0	0.64
NWZ	20.0	103.0	2.0	2.0	0.0	2.0	9.0	302.0	8.0	24.0	0.0	0.0	28.0	0.0	1.0	15.0	38.0	3.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	8.0	570.0	0.54
HM	22.0	25.0	0.0	3.0	14.0	8.0	3.0	4.0	522.0	90.0	1.0	1.0	0.0	13.0	2.0	17.0	3.0	2.0	5.0	25.0	0.0	1.0	1.0	6.0	0.0	4.0	2.0	2.0	4.0	780.0	0.61
BM	66.0	25.0	2.0	9.0	11.0	9.0	3.0	28.0	66.0	549.0	5.0	0.0	5.0	6.0	1.0	23.0	14.0	10.0	13.0	12.0	2.0	1.0	2.0	2.0	0.0	0.0	0.0	2.0	14.0	880.0	0.5
TM	0.0	4.0	8.0	3.0	0.0	0.0	0.0	3.0	0.0	4.0	116.0	0.0	4.0	1.0	7.0	0.0	20.0	2.0	14.0	2.0	8.0	1.0	14.0	2.0	5.0	0.0	0.0	6.0	11.0	235.0	0.56
NA	1.0	4.0	3.0	0.0	0.0	0.0	2.0	4.0	14.0	6.0	0.0	20.0	3.0	11.0	5.0	8.0	2.0	1.0	0.0	1.0	1.0	2.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	91.0	0.71
NZ	1.0	13.0	0.0	0.0	0.0	0.0	2.0	42.0	1.0	9.0	7.0	0.0	178.0	7.0	7.0	12.0	23.0	2.0	5.0	0.0	0.0	0.0	5.0	0.0	1.0	0.0	0.0	0.0	7.0	322.0	0.62
HNA	0.0	3.0	1.0	0.0	0.0	0.0	2.0	1.0	11.0	7.0	2.0	1.0	5.0	165.0	12.0	21.0	0.0	4.0	1.0	2.0	0.0	1.0	16.0	4.0	0.0	0.0	0.0	0.0	1.0	260.0	0.59
HNZ	0.0	8.0	6.0	0.0	0.0	5.0	0.0	3.0	3.0	8.0	7.0	0.0	20.0	12.0	109.0	3.0	6.0	1.0	3.0	2.0	0.0	3.0	23.0	1.0	0.0	0.0	0.0	1.0	5.0	229.0	0.56
HB	3.0	0.0	0.0	0.0	0.0	0.0	7.0	10.0	14.0	32.0	1.0	2.0	5.0	19.0	1.0	237.0	5.0	7.0	5.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	352.0	0.59
TRM	6.0	61.0	5.0	6.0	0.0	5.0	0.0	39.0	6.0	25.0	17.0	0.0	12.0	0.0	9.0	9.0	250.0	1.0	12.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	18.0	486.0	0.53
NEA	2.0	2.0	0.0	3.0	0.0	0.0	0.0	2.0	9.0	33.0	0.0	1.0	2.0	1.0	1.0	13.0	1.0	103.0	14.0	25.0	3.0	4.0	5.0	1.0	3.0	0.0	0.0	0.0	3.0	231.0	0.62
NEZ	0.0	2.0	0.0	5.0	0.0	0.0	1.0	4.0	1.0	37.0	4.0	0.0	3.0	0.0	5.0	18.0	8.0	10.0	132.0	9.0	9.0	1.0	13.0	1.0	3.0	1.0	0.0	0.0	2.0	269.0	0.53
HFA	1.0	2.0	0.0	12.0	3.0	2.0	0.0	1.0	25.0	25.0	0.0	0.0	0.0	3.0	0.0	1.0	2.0	8.0	5.0	275.0	8.0	6.0	8.0	17.0	4.0	3.0	0.0	0.0	0.0	411.0	0.55
HFZ	0.0	1.0	1.0	7.0	0.0	1.0	0.0	0.0	2.0	7.0	4.0	0.0	0.0	0.0	0.0	0.0	2.0	3.0	6.0	31.0	98.0	4.0	9.0	7.0	7.0	1.0	2.0	0.0	6.0	199.0	0.55
HNFA	0.0	1.0	0.0	0.0	0.0	2.0	1.0	0.0	4.0	7.0	0.0	0.0	0.0	8.0	3.0	4.0	0.0	4.0	2.0	22.0	1.0	80.0	14.0	4.0	0.0	3.0	0.0	2.0	1.0	163.0	0.63
HNFZ	0.0	1.0	7.0	3.0	0.0	1.0	0.0	1.0	5.0	3.0	13.0	1.0	3.0	13.0	10.0	6.0	1.0	2.0	7.0	11.0	7.0	11.0	184.0	2.0	3.0	1.0	0.0	2.0	3.0	301.0	0.54
SEA	0.0	1.0	4.0	6.0	1.0	1.0	0.0	0.0	17.0	15.0	1.0	0.0	0.0	5.0	1.0	0.0	0.0	0.0	0.0	33.0	4.0	8.0	8.0	98.0	5.0	17.0	5.0	0.0	1.0	231.0	0.54
SEZ	0.0	3.0	1.0	17.0	0.0	3.0	0.0	1.0	0.0	7.0	2.0	0.0	0.0	1.0	1.0	0.0	5.0	0.0	5.0	13.0	12.0	1.0	12.0	13.0	65.0	2.0	0.0	5.0	6.0	175.0	0.55
SA	0.0	1.0	0.0	9.0	10.0	1.0	0.0	0.0	26.0	7.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	14.0	3.0	0.0	1.0	8.0	4.0	113.0	4.0	2.0	4.0	208.0	0.68
SZ	0.0	2.0	2.0	11.0	0.0	10.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	8.0	5.0	7.0	71.0	6.0	3.0	134.0	0.63
TB	0.0	14.0	4.0	11.0	2.0	15.0	0.0	1.0	3.0	4.0	5.0	0.0	4.0	3.0	3.0	0.0	2.0	0.0	3.0	1.0	3.0	0.0	1.0	3.0	1.0	2.0	11.0	154.0	28.0	278.0	0.65
TRW	9.0	44.0	3.0	12.0	4.0	11.0	0.0	8.0	6.0	30.0	11.0	0.0	3.0	1.0	7.0	0.0	32.0	0.0	3.0	2.0	4.0	0.0	6.0	0.0	4.0	4.0	3.0	18.0	269.0	494.0	0.57
∑	604.0	1934.0	279.0	380.0	239.0	398.0	116.0	564.0	859.0	1097.0	209.0	28.0	289.0	282.0	194.0	403.0	470.0	165.0	247.0	497.0	178.0	127.0	341.0	180.0	118.0	167.0	112.0	237.0	469.0	11183.0	0.59
Recall	0.56	0.76	0.54	0.57	0.47	0.55	0.41	0.53	0.67	0.62	0.49	0.22	0.55	0.63	0.48	0.67	0.51	0.45	0.49	0.67	0.49	0.49	0.61	0.42	0.37	0.54	0.53	0.55	0.54	0.53	

FIGURE 2.17: Confusion matrix

Table A1. Confusion matrix of our proposed smoothed approach, averaged over the test sets of nested cross-validation. Correctly classified classes are highlighted in bold.

		Labels																													
		WA	WZ	WS	WW	SWA	SWZ	NWA	NWZ	HM	BM	TM	NA	NZ	HNA	HNZ	HB	TRM	NEA	NEZ	HFA	HFZ	HNFA	HNFZ	SEA	SEZ	SA	SZ	TB	TRW	$\sum$	Precision
Outputs	WA	**102**	62	1	3	4	2	10	8	22	26	0	1	1	0	0	1	3	1	1	0	0	0	0	0	0	0	1	3	253	0,40
	WZ	13	**195**	12	5	2	6	1	11	4	2	0	1	2	1	0	0	6	1	0	0	0	0	0	0	0	0	0	2	5	272	0,72
	WS	1	51	**67**	3	0	8	0	4	1	1	4	0	0	1	3	0	6	0	0	0	0	0	1	0	3	0	1	6	5	170	0,40
	WW	4	24	3	**42**	2	4	1	3	6	6	1	0	0	0	0	0	3	2	3	2	1	0	0	2	2	3	2	2	6	125	0,33
	SWA	14	18	1	4	**34**	11	0	0	25	7	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	4	0	1	2	125	0,27
	SWZ	4	36	9	5	8	**30**	0	1	4	1	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	2	1	4	5	115	0,26
	NWA	14	8	0	1	0	0	**58**	18	12	20	0	1	4	2	1	12	3	2	2	0	0	0	0	0	0	0	0	0	1	159	0,37
	NWZ	10	56	2	2	0	0	15	**67**	2	4	2	1	7	1	0	1	21	1	1	0	1	0	0	0	1	0	0	1	3	198	0,34
	HM	8	4	1	1	4	1	5	1	**142**	23	0	1	0	2	0	4	0	2	1	7	1	1	0	2	0	3	0	0	1	216	0,66
	BM	15	5	0	3	1	0	7	2	25	**104**	0	0	1	1	0	2	3	3	3	3	0	1	0	1	0	2	0	0	2	187	0,56
	TM	0	8	4	1	0	0	0	3	0	1	**34**	1	6	0	4	0	10	1	3	1	1	1	5	0	1	0	1	4	6	95	0,36
	NA	2	4	0	0	0	0	3	1	4	1	0	**12**	7	12	3	3	1	2	2	1	0	1	1	0	0	0	0	0	0	59	0,20
	NZ	2	7	0	0	0	0	6	12	1	2	6	4	**52**	4	5	3	15	1	2	0	0	0	1	0	0	0	0	0	1	125	0,42
	HNA	1	4	1	0	0	0	2	0	11	3	1	6	3	**54**	5	10	1	1	1	2	0	3	2	2	0	0	0	1	0	116	0,47
	HNZ	0	5	3	0	0	1	1	0	1	1	7	2	7	10	**17**	1	3	0	1	0	0	1	7	1	0	0	0	1	2	72	0,23
	HB	1	1	0	0	0	0	19	4	13	11	0	2	2	8	0	**65**	2	5	2	2	0	1	0	0	0	0	0	0	0	141	0,46
	TRM	3	14	2	1	0	1	1	17	1	4	6	0	9	1	0	0	**35**	1	2	0	0	0	0	0	1	0	0	1	8	109	0,32
	NEA	1	2	0	2	0	0	4	1	10	10	1	1	1	3	0	5	1	**43**	13	10	2	2	1	1	0	1	0	0	1	114	0,38
	NEZ	1	2	1	1	0	0	5	1	1	5	4	0	2	1	1	2	5	10	**26**	2	2	1	2	0	1	0	0	0	1	79	0,33
	HFA	1	1	0	1	1	0	0	0	21	5	0	0	0	3	0	1	1	9	2	**63**	5	4	1	9	2	4	0	0	1	135	0,47
	HFZ	0	1	1	2	0	0	0	1	1	1	3	0	0	0	0	0	1	3	8	12	**12**	1	4	2	5	1	0	1	1	61	0,20
	HNFA	0	1	0	0	0	0	0	0	2	3	1	1	2	11	1	2	1	3	3	12	1	**21**	8	3	1	0	0	1	1	79	0,27
	HNFZ	0	3	1	0	0	0	0	0	1	1	7	0	2	4	7	0	1	0	2	4	2	6	**24**	3	2	0	0	1	2	74	0,32
	SEA	0	1	1	1	1	0	0	0	7	2	0	0	0	3	0	0	0	0	1	12	2	2	3	**32**	9	8	1	1	1	87	0,36
	SEZ	0	2	3	4	0	0	0	1	0	1	2	0	0	0	0	0	2	1	2	1	4	0	3	6	**24**	2	3	2	3	68	0,35
	SA	1	2	0	3	4	3	0	0	16	5	0	0	0	1	0	0	0	0	0	6	0	0	0	7	1	**34**	4	1	3	94	0,36
	SZ	0	1	5	4	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	4	8	**19**	7	5	63	0,30
	TB	2	20	7	4	2	5	0	3	3	1	3	0	0	0	1	0	3	0	0	1	0	0	1	1	2	2	2	**44**	12	120	0,37
	TRW	4	18	3	3	1	3	1	4	2	4	2	0	1	0	0	0	12	0	1	0	1	0	0	1	1	1	1	7	**32**	104	0,31
	$\sum$	205	557	126	97	68	81	140	164	340	255	88	35	111	123	51	115	140	94	79	143	36	49	67	78	59	76	36	89	116	3616	—
	Recall	0,50	0,35	0,53	0,43	0,50	0,37	0,42	0,41	0,42	0,41	0,39	0,34	0,47	0,44	0,33	0,57	0,25	0,46	0,33	0,44	0,33	0,44	0,36	0,40	0,41	0,45	0,53	0,49	0,28	0,41	—

FIGURE 2.18: Confusion matrix

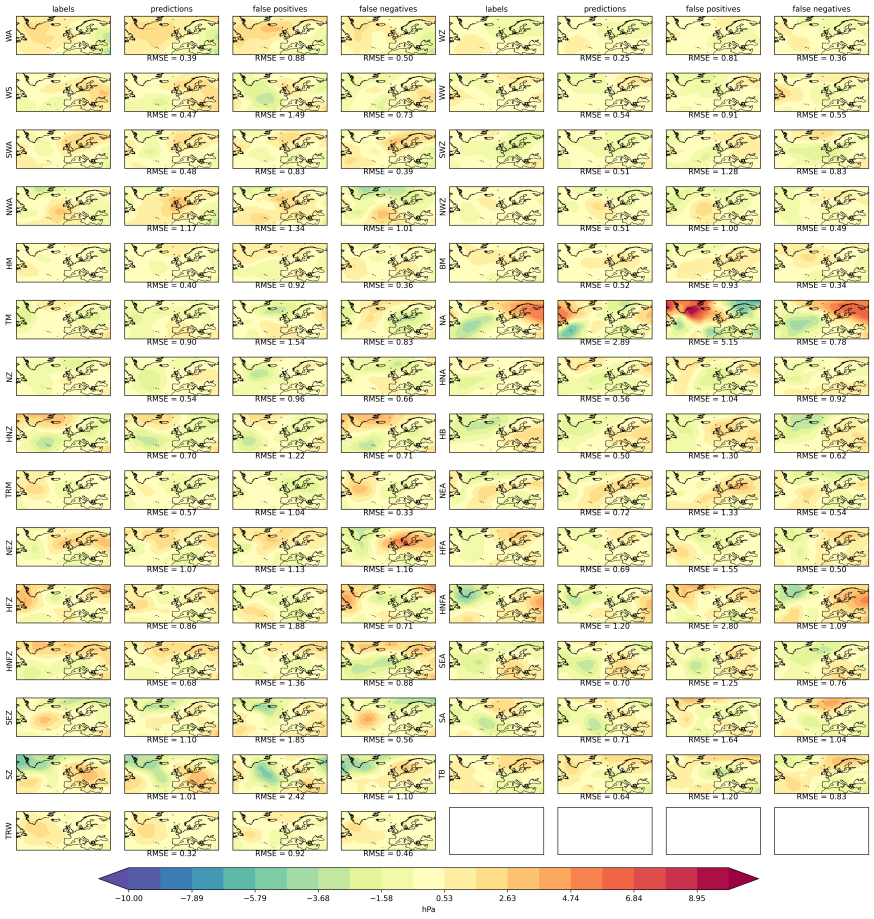


FIGURE 2.19: Signature plot

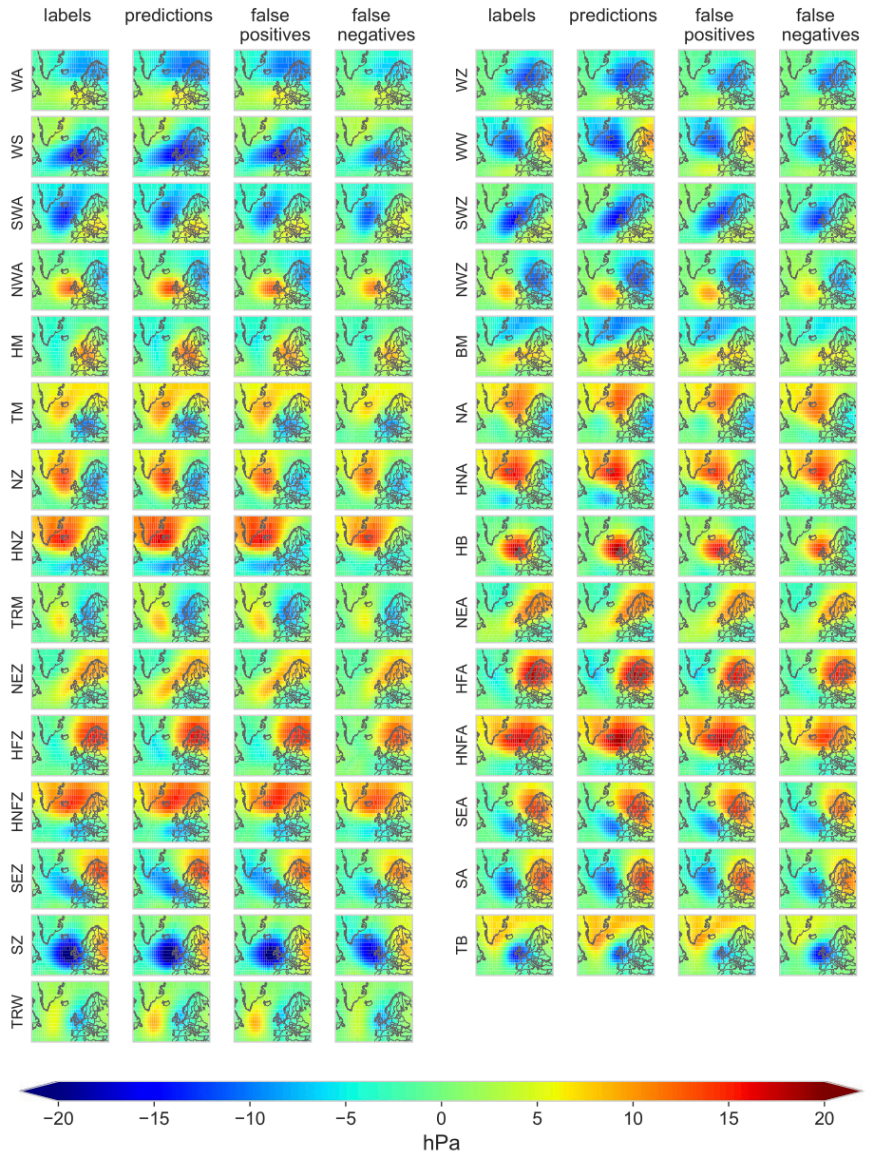


FIGURE 2.20: Signature plot

3

Causal Discovery - Constrain uncertainties in climate projections

Author:Marta Caserio & Jonas Ameluxen

Supervisor: Henri Funk

3.1 Abstract

Understanding and evaluating the performance of climate models is essential for improving predictions of future climate variability. Traditional evaluation techniques often fall short in identifying deep-seated structural biases in models. This study introduces a novel process-oriented evaluation approach using causal discovery methods, specifically the PCMCI algorithm, to assess global drought teleconnections. By applying the PCMCI algorithm to SPI-12 precipitation indices from both reanalysis (ERA5) and climate model (CSIRO ACCESS ESM 1.5) datasets, we extract causal networks that reveal underlying climate modes and their interactions. A Varimax-rotated principal component analysis (PCA) was used to reduce data dimensionality, and selected components were analyzed to evaluate the consistency between observed and simulated teleconnections. Results highlight significant differences in causal link structures between the datasets, particularly in ocean-dominated climate modes, suggesting that while PCMCI has limitations in physical interpretation, it holds strong potential for comparative climate model diagnostics. Our findings underscore the importance of integrating causal inference tools into the climate model evaluation toolbox to better constrain model uncertainty and improve future projections.

3.2 Introduction to climate models

Understanding our climate system and how it responds to certain external or internal inputs has always been a key part of scientific research. However, due to the high complexity and nonlinearity of a system as large as the earth that operates on timescales from seconds to decades using experimental methods to understand the earth system is not feasible (Edwards (2011);Runge et al. (2019)). For this reason, models have been used to represent, abstract and simplify the most important drivers of the earth system. The earliest beginnings of conceptual climate models can be traced back to ancient scholars like Ptolemy who distinguished different climatic zones based on the maximal daylength and latitude (Edwards (2011); Sanderson (1999)). Complex mathematical climate models started to emerge in the 19th and 20th centuries through scientists like Milutin Milanković who managed to explain a large part of natural climate variability through periodic cycles of earth's eccentricity, axial tilt and precession (Edwards (2011)). With the rapid technological and scientific breakthroughs since the 20th century the complexity and accuracy of climate models increased tremendously on one hand through better understanding of the underlying principles but also through collection of decades of observational data from satellites and field measurements (Runge et al., 2019).

Modern global climate models can be divided into two subgroups. The first one are general circulation models (GCMs) which simulate the dynamics of atmosphere (AGCM) and oceans (OGCM) following the laws of fluid

motion, thermodynamics and momentum conservation (eva (2013); Nowack et al. (2020)). In these models atmosphere and oceans are divided into grid cells for which the dynamical equations describing the evolution of variables like temperature or vapor pressure are solved with numerical methods (cli (2008)). An extension to these models are Earth System Models (ESMs) which expand the GCMs by including biogeochemical cycles such as the carbon or nitrogen cycle or atmospheric chemistry (cli (2008)).

Anthropogenic climate change has changed Earth's climate at unprecedented rates. To better understand this change and model future climate pathways the Coupled Model Intercomparison Project (CMIP) was organized by the Working Group on Coupled Modelling (WGCM) in an effort the compare state of the art GCMs and ECMs and tackle important questions regarding climate change ((Eyring et al., 2016a)). Over the last 30 years CMIP went through 6 different Phases, each including more models and addressing a wider range of research questions. With its standardized framework CMIP allows detailed multi-model evaluation which, over the years, revealed model specific systemic differences between individual model groups and observations ((Eyring et al., 2019)). For this reason, rigid climate model evaluation is crucial and has been a rapidly advancing field over the past decades. While more and more routine evaluation metrics and tools like the Earth System Model Evaluation Tool ((Eyring et al., 2016b)) using metrics such as means, variances and trend analysis have been developed, these methods often fail to identify underlying model biases ((Nowack et al., 2020)).

A novel approach to constrain uncertainties in climate models is a process-oriented causal model evaluation (CME) approach introduced by (Nowack et al., 2020). This method utilizes causal discovery methods developed by (Runge et al., 2019) (Detecting causal associations in large nonlinear time series datasets, Sci. Adv., 5, eaau4996 2019) to systematically exclude common driver effects and indirect links ((Nowack et al., 2020)), resulting in a network of causal global connections. (Nowack et al., 2020) applied CME to show that inter-model comparison and comparison to observational data of the resulting causal networks can identify biases in climate models and thus help reducing uncertainties for climate predictions.

In this work we introduce the method proposed by (Nowack et al., 2020) and show one potential use by applying it to global drought datasets based on reanalysis and global climate model precipitation datasets.

3.3 Process

The use of Causal Networks and especially the PCMCI algorithm, as introduced in the previous section, to help evaluate and better understand large scale climate data timeseries and model outputs has gained some popularity over the last years. One topic where PCMCI algorithms have been applied multiple times are global weather teleconnections. It has been shown multiple times that weather patterns and weather extremes like precipitation and temperature can have significant influence on weather in regions thousands of kilometers aways. One example of such a long distance weather teleconnection is the impact of the El Niño Southern Oscillation (ENSO) on North American precipitation and weather patterns(Ropelewski and Halpert, 1986). An example of a teleconnection between two hydrometeorological extremes is the 2010 floods in Pakistan that were shown to be connected to a heatwave in western Russia (Lau and Kim, 2012). In both cases the driver behind these teleconnections were atmospheric wave trains (Rossby Waves) that lead to a hydrometeorological connection between the distant regions (Lau and Kim (2012); Ropelewski and Halpert (1986)). Such teleconnections can work in both directions but can also be one-directional as in the ENSO-North-American case (Detecting causal associations in large nonlinear time series datasets, Sci. Adv., 5, eaau4996 2019). The PCMCI algorithm can be used to help discover or confirm suspected teleconnections, however caution when interpreting such results is necessary. One study used PCMCI to show the significant role global teleconnections play in the synchronization of extreme rainfall events (Boers et al., 2019).

In this work we will apply the PCMCI algorithm to global standard precipitation index (SPI) datasets to assess drought teleconnections. Introduced in 1993, the SPI is a commonly used measure to define droughts by fitting long-term baseline precipitation values to a probability distribution (usually gamma distribution)

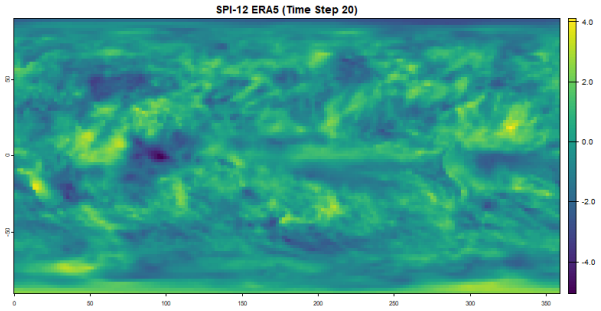


FIGURE 3.1: SPI-12 patterns on ERA5 reanalysis data showing observed drought patterns.

and then transforming the probability values into a standard normal variable with = 0 and =1 (The relationship of drought frequency and duration to time scales 1993). SPI values were calculated based on 12 month cumulative precipitation values (SPI-12) and hence represent long lasting droughts or wet-periods (Chauhan et al., 2024). In a previous study it was shown that especially oceans play a significant role on modulating global droughts (Chauhan et al., 2024). Similar to (Nowack et al., 2020) we applied the PCMCI algorithm to a reanalysis dataset and one climate model dataset. The underlying assumption being that the causal network from the reanalysis dataset represents real world teleconnections. Comparing this to the causal network of the climate model can uncover where the model fails or succeeds at reproducing those teleconnections.

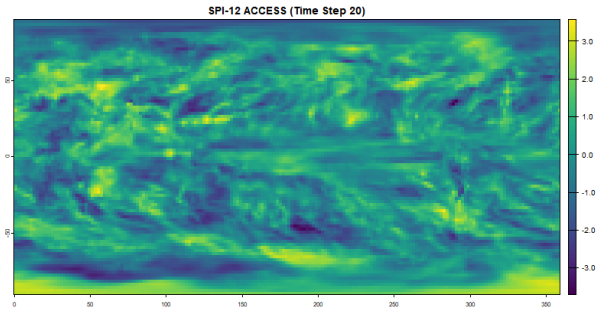


FIGURE 3.2: SPI-12 patterns on ACCESS ESM 1.5 climate model simulation.

3.4 Statistical background

3.4.1 Principal Component Analysis and Varimax Rotation

Our PCA implementation transforms the high-dimensional SPI-12 data into a set of linearly uncorrelated variables that maximize explained variance. This approach effectively identifies coherent drought patterns across geographical regions while substantially reducing computational complexity for subsequent causal analysis. The Kaiser criterion was employed specifically because it provides an objective threshold for component retention by selecting only those components with eigenvalues exceeding unity, which by definition contribute more information than a single original variable.

The Varimax rotation technique redistributes the component loadings to achieve "simple structure," where each grid cell preferentially loads strongly onto a single component. Mathematically, Varimax maximizes the sum of the variances of squared loadings within each component:

$$V = \sum_k (\sum_i (l_{ik}^2 - \bar{l}_k^2)^2)$$

Where l_{ik} represents the loading of variable i on component k, and $\bar{l}_k^2$ is the mean of the squared loadings for component k. This optimization produces more spatially distinct patterns compared to unrotated PCA, making it particularly valuable for identifying climatologically meaningful teleconnection patterns. The $|0.4|$ threshold for significant loadings was selected based on conventional practice in climate research, representing a balance between noise reduction and retention of meaningful spatial signals.

3.4.2 PCMCI Algorithm Technical Implementation

The PCMCI algorithm addresses fundamental limitations of traditional correlation analyses by systematically controlling for autocorrelation, common drivers, and indirect causal effects. The PC step implements a condition-selection algorithm where for each time series Y, potential causal parents X are identified through iterative conditional independence testing. The algorithm begins with a full set of potential parents (all time series at all considered lags) and progressively removes links that fail conditional independence tests with increasing conditioning set sizes.

The MCI test then evaluates the conditional independence between each potential cause-effect pair $(X_{t-\tau}, Y_t)$ while controlling for both the past of Y and all other potential common causes, using the formula:

$$\text{MCI} : X_{t-\tau} \perp\!\!\!\perp Y_t | Z_t^Y, Z_{t-\tau}^X$$

Where Z_t^Y represents the parents of Y excluding $X_{t-\tau}$, and $Z_{t-\tau}^X$ represents the parents of $X_{t-\tau}$. This formulation allows PCMCI to distinguish between direct and indirect causal relationships, reducing spurious connections often found in traditional correlation analyses.

In our implementation, we used partial correlation as the conditional independence test with standardized time series data. The time lag parameters were specifically configured with `tau_min = 1` and `tau_max = 5`, allowing us to capture causal relationships occurring between 1 and 5 months.This range is sufficient to detect both relatively rapid atmospheric teleconnections and slower oceanic teleconnection patterns. We deliberately employed a stringent significance threshold with `pc_alpha = 0.0001` to ensure high confidence in the detected causal links, effectively minimizing false positives while accepting a potentially higher rate of false negatives. This conservative approach prioritizes the reliability of identified teleconnections over their quantity, particularly important when comparing model outputs to observational data. The resulting causal networks represent directional relationships between drought patterns, providing insights into the causal mechanisms driving global drought teleconnections and enabling rigorous evaluation of climate model performance in reproducing these teleconnection structures.

3.5 Results

The reanalysis dataset we used is the Copernicus ERA5 post-processed daily statistics on single levels from 1940 to present. CSIRO ACCESS ESM 1.5 data was used as the climate model dataset. For both datasets the years 1950 until (including) 1990 were used. The ERA5 dataset contains daily precipitation values at 0.25° x 0.25° resolution. ACCESS ESM 1.5 data contains daily precipitation rate values at 1.875° x 1.25° resolution. Both datasets were aggregated to total monthly precipitation in mm and regridded to a 2° x 2° resolution using bilinear interpolation.

To reduce the high dimensionality of the datasets we used a Varimax rotated principal component analysis (PCA) to identify large-scale patterns of SPI and thus drought variability. The PCA was applied to monthly SPI-12 values across all grid cells from 1950 to 1990. Thirty principal components (PCs) were retained based on the Kaiser criterion (eigenvalues >1). Varimax rotation was then applied to enhance interpretability by maximizing the variance of squared loadings within each component. The loadings threshold was set at |0.4| to determine significant contributions from individual grid cells. From the rotated components, we selected those with variance contributions exceeding 4% for detailed causal analysis—five for ERA5 and six for ACCESS.

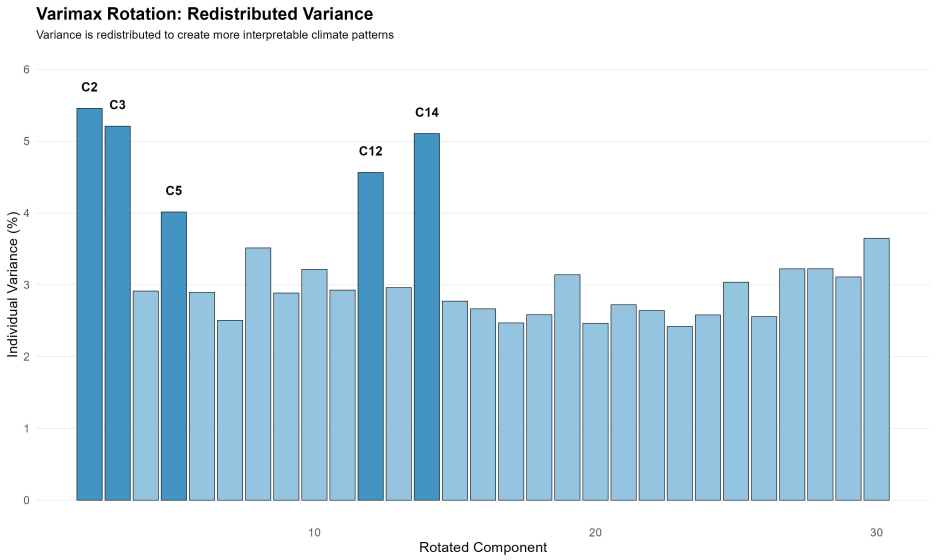


FIGURE 3.3: Varimax rotated principal components for ERA5 dataset.

In the ERA5 reanalysis dataset (Figure 3.3), the variance distribution is characterized by five prominent components (C2, C3, C5, C12, and C14) that individually explain more than 4% of the total variance. Components C2 and C3 show particularly high explanatory power, accounting for approximately 5.4% and 5.2% of the variance respectively. These high-variance components likely represent dominant global drought teleconnection patterns with substantial spatiotemporal coherence. The remaining 25 components each explain approximately 2.5-3.5% of the variance, resulting in a more uniform distribution of explanatory power across the rotated component space. This pattern suggests that after accounting for the major teleconnection modes, the remaining drought variability is distributed across numerous localized or regional patterns of similar importance.

The ACCESS ESM 1.5 climate model (Figure 3.4) exhibits a somewhat different variance structure, with six components (C1, C2, C3, C7, C13, and C21) exceeding the 4% variance threshold. Component C13 displays the highest explanatory power at approximately 5.2%, followed closely by C2 at 5.0%. A notable difference from the ERA5 results is the appearance of C21 as a significant component, suggesting that the climate model

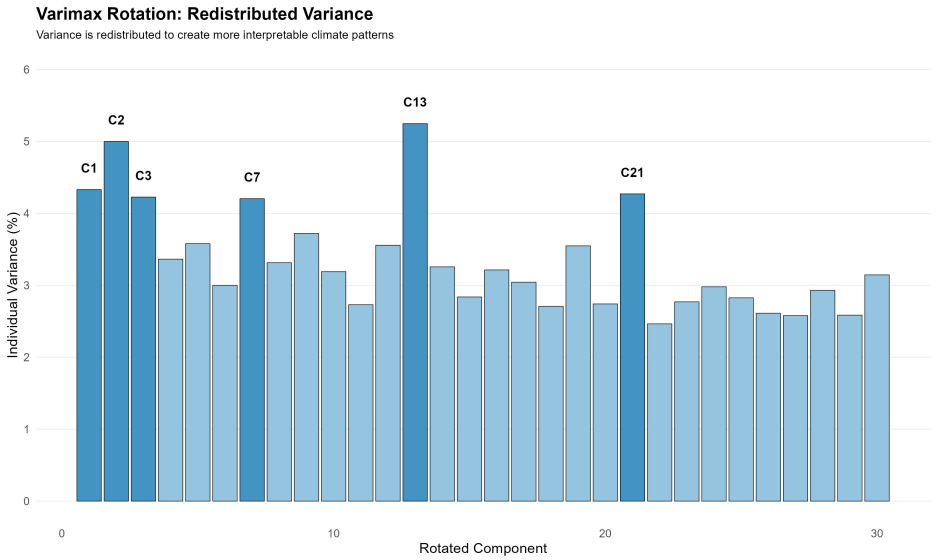


FIGURE 3.4: Varimax rotated principal components for ACCESS dataset.

simulates an important teleconnection pattern that manifests in a higher-order component. This structural difference in the variance distribution between observed and modeled data provides an initial indication that ACCESS ESM 1.5 may represent certain climate processes differently than observed in reanalysis data.

The variance threshold of 4% was selected as an objective criterion to isolate the most influential drought teleconnection patterns while maintaining computational tractability for the subsequent causal analysis. The rotated components exceeding this threshold were selected for detailed causal network analysis using the PCMCI algorithm.

It is important to note, as highlighted by Hannachi et al. (2007), that while Varimax rotation enhances pattern interpretability, the resulting components represent statistical constructs that may not perfectly align with physically coherent climate phenomena. Nevertheless, these rotated components provide valuable insights into the spatial organization of drought variability and offer a robust foundation for subsequent causal analysis of teleconnection structures. The differences in component structure between ERA5 and ACCESS ESM 1.5 datasets suggest potential model biases in representing the spatial organization and relative importance of global drought patterns, a finding that will be further explored through causal network comparison.

3.5.1 Spatial Patterns of Rotated Components

The spatial distributions of the five high-variance ERA5 components (Figure 3.5) reveal distinct geographical footprints that correspond to known ocean-atmosphere coupled systems. Notably, all five components are predominantly oceanic in nature, underscoring the critical role of sea surface temperature patterns in modulating global drought teleconnections.

Component 2 exhibits a clear tropical Pacific signature, with significant loadings concentrated in the central-eastern equatorial Pacific. This pattern strongly resembles the canonical Eastern Pacific El Niño (EP-ENSO) pattern, characterized by maximum sea surface temperature anomalies in the eastern tropical Pacific (Di Lorenzo et al., 2013). Component 5 displays a broader Pacific footprint extending from the eastern to the central Pacific with notable loadings along the equatorial and subtropical Pacific, suggesting influences from both ENSO and the Pacific Decadal Oscillation (PDO). Component 14 shows a distinctive pattern with significant loadings in both the western-central tropical Pacific and the southeastern Pacific, along with notable signals in the Mediterranean and southeastern Asia, potentially representing a combination of Central Pacific El Niño (CP-ENSO) and Indo-Pacific teleconnection patterns.

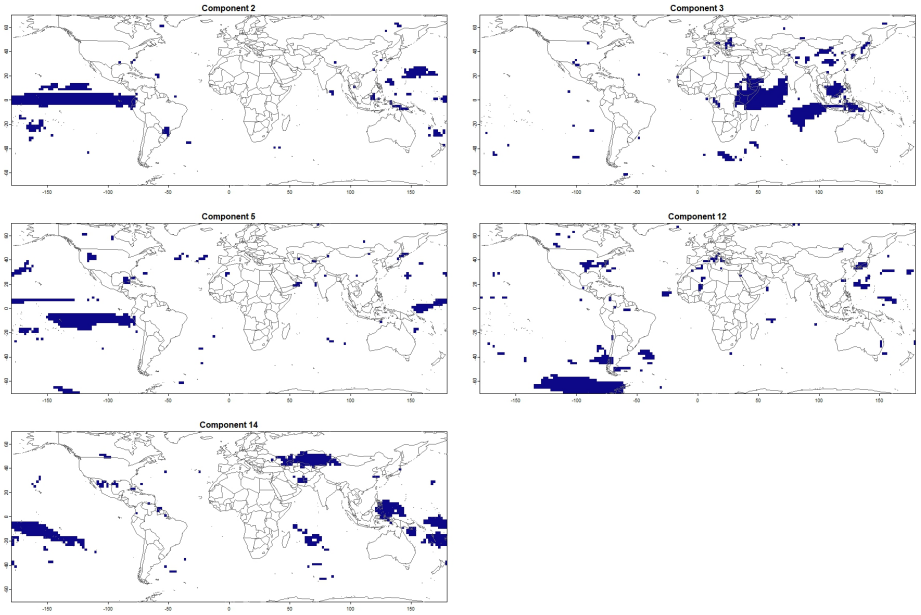


FIGURE 3.5: World map showing the location of the different rotated components from ERA5 dataset

Component 3's spatial distribution extends from the Western Indian Ocean (WIO) across the maritime continent to the Indo-Pacific Warm Pool (IPWP). This pattern likely represents the Indian Ocean Dipole (IOD) in conjunction with IPWP variability, systems known to significantly influence precipitation patterns across the Indo-Pacific region through modulation of the Walker circulation (Zhang and Han, 2020; Newton et al., 2006). Component 12 shows predominant loadings in the Arctic Ocean region, which may partially reflect coordinate projection effects that can amplify variance near the poles in global gridded datasets, as higher latitudes are represented by more grid cells per unit area due to meridional convergence.

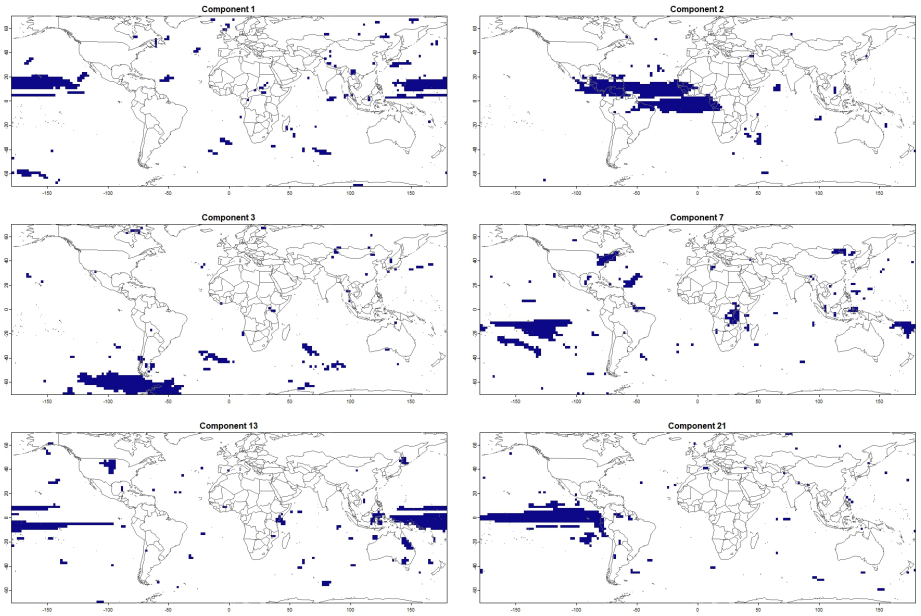


FIGURE 3.6: World map showing the location of the different rotated components from ACCESS dataset

The ACCESS ESM 1.5 dataset (Figure 3.6) reproduces several Pacific Ocean teleconnection patterns similar

to those identified in ERA5, though with notable structural differences. Components 1, 7, and 13 capture various aspects of Pacific variability, with Component 1 showing a strong equatorial Pacific signal comparable to EP-ENSO, Component 7 exhibiting a broader eastern Pacific pattern with extensions into the North Atlantic and Indian Ocean, and Component 13 displaying a pan-Pacific pattern reminiscent of the PDO. Component 21 reveals an intriguing pattern connecting the tropical eastern Pacific with the subtropical North Atlantic, potentially representing an inter-basin teleconnection mechanism.

A critical difference between the datasets emerges in the representation of Indian and Atlantic Ocean teleconnection patterns. While ERA5 identifies a strong WIO-IPWP component (Component 3), ACCESS ESM 1.5 shows no comparable high-variance component in this region. Instead, ACCESS ESM 1.5 exhibits a prominent Atlantic Ocean component (Component 2) that is not evident among the high-variance ERA5 components. This structural difference suggests a potential model bias in representing the relative importance of Indian Ocean versus Atlantic Ocean teleconnection systems, which could significantly impact simulated drought patterns across adjacent continental regions. Component 3 in ACCESS is predominantly concentrated in the Arctic region, similar to Component 12 in ERA5, though with greater extension into northern Eurasia.

Both datasets highlight the dominance of oceanic climate modes in explaining global drought variability, aligning with findings from Chauhan et al. (2024) that oceanic influences play a crucial role in synchronized drought events across multiple regions. The numerous smaller, scattered regions of significant loadings visible in both datasets likely represent mathematical artifacts of the Varimax rotation process rather than physically coherent teleconnections. These secondary signals should be interpreted cautiously as they may not represent robust geophysical connections to the primary climate mode represented by each component.

The differences in spatial patterns between ERA5 and ACCESS ESM 1.5 components provide valuable insights into potential model biases in simulating the spatial structure and relative importance of major climate teleconnection systems.

3.5.2 Causal Network

The PCMCI algorithm was applied to all 30 components from both ERA5 and ACCESS ESM 1.5 datasets, yielding comprehensive causal networks shown in Figures 3.7 and 3.8. Both networks exhibit a dominant pattern of strong autocorrelation, indicated by the self-loops (circular arrows) at each node, consistent with findings from Nowack et al. (2020) in their analysis of sea level pressure datasets. This autocorrelation reflects the inherent memory in climate systems, where drought patterns typically persist over multiple months due to soil moisture feedback mechanisms and the relatively slow evolution of ocean temperature anomalies that drive teleconnection patterns.

Beyond autocorrelation, the full networks reveal complex interconnections between different components, with varying correlation strengths and directionality. The color scale indicates correlation strength and sign, with red representing positive correlations (where an increase in one component leads to a subsequent increase in the connected component) and blue representing negative correlations (where an increase leads to a subsequent decrease). The full networks exhibit a mix of both positive and negative causal links, reflecting the complex feedback mechanisms in global climate teleconnections, where both reinforcing and dampening interactions can occur.

While a comprehensive analysis of all causal links in these full networks is beyond the scope of this study, visual inspection reveals some structural differences between ERA5 and ACCESS ESM 1.5 networks. The ERA5 network appears to have a slightly denser structure of inter-component links compared to ACCESS, particularly in the lower-variance components. This difference suggests that the climate model may not fully capture the complexity of interactions between secondary drought teleconnection patterns, potentially simplifying some of the more subtle causal relationships present in observational data.

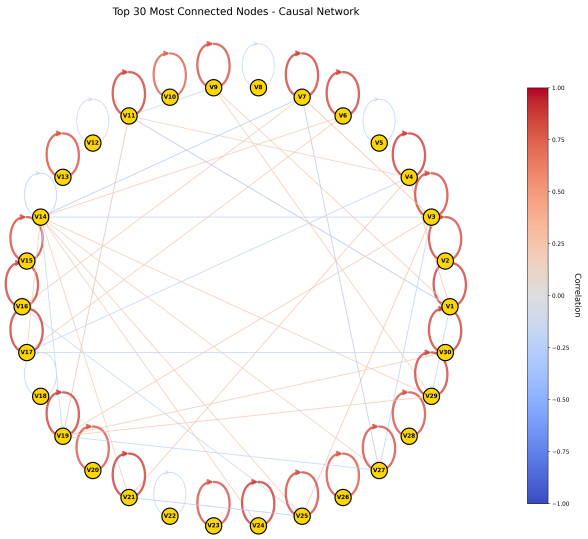


FIGURE 3.7: Causal network of all 30 rotated components from ERA5 dataset showing significant causal links detected by the PCMCI algorithm. Red links indicate positive correlations while blue links represent negative correlations. The prominence of self-loops (autocorrelation) is evident across most nodes.

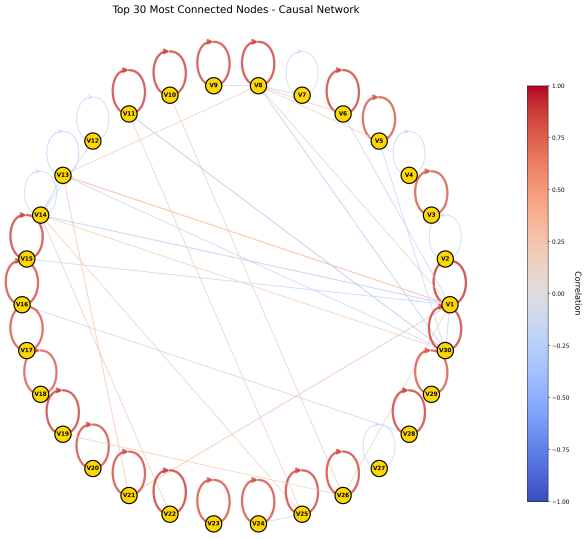


FIGURE 3.8: Causal network of all 30 rotated components from ACCESS ESM 1.5 dataset. Compared to ERA5, the ACCESS model shows a slightly different pattern of inter-component connectivity, though with similarly dominant autocorrelation signals.

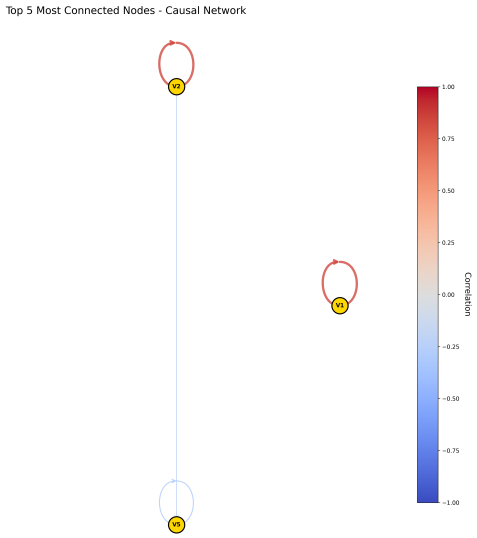


FIGURE 3.9: Causal network of the five selected high-variance components from ERA5 dataset. Node labels correspond to components (V1=C2, V2=C3, V3=C5, V4=C12, V5=C14). Note the single weak negative causal link between the Western Indian Ocean/Indo-Pacific Warm Pool component (V2) and the Pacific component (V5).

3.5.3 High-Variance Component Networks

Focusing on the high-variance components (Figures 3.9 and 3.10), we observe notable differences in causal structure between ERA5 and ACCESS ESM 1.5. In the ERA5 network (Figure 3.9), the five selected components (V1=C2, V2=C3, V3=C5, V4=C12, V5=C14) exhibit minimal inter-component causal connectivity, with only one significant causal link detected: a negative correlation from component 3 (V2, representing the Western Indian Ocean/Indo-Pacific Warm Pool) to component 14 (V5, representing Pacific variability patterns). This negative relationship aligns with known climate dynamics, where warming in the Western Indian Ocean/Indo-Pacific Warm Pool can induce atmospheric wave patterns that influence Pacific circulation, typically manifesting as a dampening effect consistent with the negative correlation observed. This finding corroborates Chauhan et al. (2024), who identified similar teleconnections between these ocean basins in their global drought analysis.

In contrast, the ACCESS ESM 1.5 network (Figure 3.10) exhibits a more interconnected structure among its six high-variance components (V1=C1, V2=C2, V3=C3, V4=C7, V5=C13, V6=C21). The model simulates a triangular causal structure between three Pacific-dominated components (V1, V5, and V6, corresponding to components 1, 13, and 21), with positive correlations of moderate strength. This interconnected Pacific structure suggests that the climate model simulates stronger intra-basin coupling within the Pacific than is evident in the observational data. Notably, component V1 (C1) exerts causal influence on both V5 (C13) and V6 (C21), while V5 and V6 also share a direct causal link, forming a closed causal triangle that implies potential feedback mechanisms within Pacific climate variability.

The striking difference between these causal structures—ERA5 showing primarily inter-basin teleconnection (Indian-Pacific) versus ACCESS showing stronger intra-basin connections (within Pacific)—reveals a key bias in the climate model's representation of drought teleconnection mechanisms. The model appears to overemphasize Pacific internal dynamics while underrepresenting the crucial teleconnections between the Indian and Pacific Oceans that are evident in observational data. This finding has important implications for the model's ability to simulate drought propagation patterns, particularly for regions influenced by Indian Ocean climate variability.

It is worth noting that while both datasets exhibit strong autocorrelation in their components (self-loops with

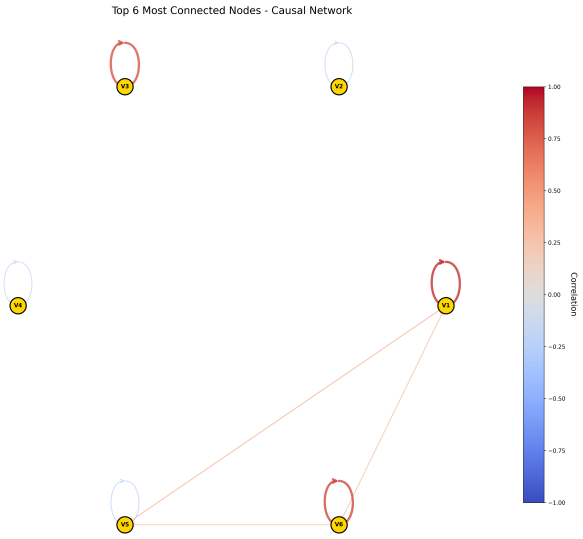


FIGURE 3.10: Causal network of the six selected high-variance components from ACCESS ESM 1.5 dataset. Node labels correspond to components (V1=C1, V2=C2, V3=C3, V4=C7, V5=C13, V6=C21). The model simulates a triangular causal structure between Pacific Ocean components (V1, V5, V6), which differs substantially from the observed ERA5 pattern.

positive correlations), the relative strength of these autocorrelations differs between ERA5 and ACCESS. The model generally simulates stronger autocorrelation in its Pacific components, which may indicate that the model's representation of Pacific climate modes exhibits greater persistence than observed in nature. This bias in temporal dynamics could affect the model's ability to accurately simulate the timing and duration of drought events associated with Pacific variability.

These differences in causal network structure between ERA5 and ACCESS ESM 1.5 highlight the utility of causal discovery methods for climate model evaluation, revealing specific biases in teleconnection representation that might not be apparent from traditional evaluation metrics focused on means, variances, or spatial patterns alone. The detection of these structural differences in causal mechanisms provides valuable insights for targeted model improvement efforts, particularly in the representation of ocean-atmosphere coupling processes that drive global drought teleconnections.

3.6 Limitations

This study has explored the application of the PCMCI algorithm as a causal discovery method for understanding teleconnection structures in global drought patterns. While our results demonstrate the potential of this approach, they also highlight several important methodological challenges that must be addressed for effective implementation in climate science applications.

The application of causal discovery methods to high-dimensional climate datasets necessitates dimensionality reduction techniques such as the Varimax-rotated PCA employed in this study. However, these statistical approaches do not inherently guarantee physically meaningful representations of the climate system. The rotated components, while statistically optimized for variance explanation and interpretability, may combine or separate physical processes in ways that do not align with actual climate dynamics. This creates a fundamental dependency on expert knowledge to either validate the physical relevance of statistically derived

components or to pre-select regions of interest based on established climate science understanding (Nowack et al., 2020).

Similarly, the causal links identified through PCMCI analysis require careful interpretation within the context of known atmospheric and oceanic processes. A statistically significant causal relationship between two components does not automatically translate to a well-understood physical mechanism. The blue link between Western Indian Ocean/Indo-Pacific Warm Pool and Pacific components in our ERA5 analysis, while statistically robust, requires corroboration from atmospheric dynamics theory to establish its physical validity. This interpretive requirement limits the explanatory power of causal networks as standalone discovery tools, particularly when applied to complex, multi-scale phenomena like global drought teleconnections.

The stringent statistical threshold (`pc_alpha = 0.0001`) applied in our implementation further constrains the detection of weaker but potentially important causal connections, as evidenced by the sparse causal network in our ERA5 analysis. While this conservative approach minimizes false positives, it may suppress the identification of emerging or secondary teleconnection pathways but nevertheless contribute significantly to drought propagation dynamics.

3.7 Conclusions

Despite these limitations as an explanatory method, our findings strongly support the value of causal discovery approaches for comparative model evaluation purposes. The notable differences between ERA5 and ACCESS ESM 1.5 causal structures—particularly the contrasting inter-basin versus intra-basin teleconnection patterns—reveal specific model biases that might remain undetected through traditional evaluation metrics focused on means, variances, or spatial patterns.

As demonstrated by Nowack et al. (2020) and corroborated by our results, causal networks provide a process-oriented diagnostic framework that can identify where models succeed or fail in reproducing the underlying causal mechanisms that drive climate variability. This capability becomes especially valuable when evaluating climate projections across multiple models, where observational validation is not possible. In such contexts, a model's ability to reproduce known causal structures in historical simulations may serve as an important indicator of its reliability in projecting future climate states.

Furthermore, the structural differences detected in our analysis suggest specific areas for targeted model improvement, particularly in the representation of Indian Ocean-Pacific Ocean teleconnections that appear underemphasized in the ACCESS ESM 1.5 model. Such diagnostic insights can guide focused development efforts to enhance the physical representation of key teleconnection mechanisms in climate models.

4

Cold Extremes

Author: Zhuoyang Li and Katrin Strößner

Supervisor: Henri Funk

4.1 Abstract

Extreme cold events have long been associated with severe societal impacts on energy systems, infrastructure, and public health. Therefore, it remains important to explore the potential for such events to occur in the future and develop appropriate measures in advance. In the context of global warming, even cold winters in Central Europe have been affected by rising temperatures. In this study, we investigated whether extremely cold winters—such as the coldest winter in Germany in 1963—could still occur under a warming climate.

We first applied a dynamical adjustment approach combined with elastic net regression to confirm that atmospheric circulation was the main driver of temperature anomalies. This method also captured a decreasing tendency in the frequency and magnitude of cold extremes under global warming conditions. Furthermore, by examining the most extreme cold storylines from the supplementary boosted data provided by (Sippel et al. (2024)), we found that extremely cold winters—such as the event observed in 1963—remain physically plausible despite a warming climate.

4.2 Introduction and Geographic Background

Extreme cold events, or cold waves, are periods of unusually low temperatures that can severely impact society, leading to increased mortality, energy crises, and disruptions to infrastructure, transportation, and agriculture (Pinto et al. (2024)). Europe has experienced several significant cold waves in recent history, including February 2012 in Eastern and Central Europe (Planchon et al. (2015)), January 2017 in Southeastern Europe (Anagnostopoulou et al. (2017)), March 2018 across Northern and Western Europe (Karpechko et al. (2018)), and the winter of 2023, which was exacerbated by energy shortages linked to the Russian-Ukrainian war (Quesada et al. (2023)). Despite long-term warming trends, cold waves remain a major concern due to their unpredictable nature and severe socio-economic consequences.

Cold waves typically arise from persistent atmospheric circulation patterns that direct the cold Arctic or Eurasian air into Europe (Quesada et al. (2023)). Key mechanisms include Scandinavian blocking, sudden stratospheric warming events, North Atlantic sea surface temperature anomalies, and snow-albedo feedbacks, which can amplify and prolong cold conditions. One of the most extreme examples is the winter of 1962/1963, the coldest on record in many Central European countries ((Eichler, 1970; Sippel et al., 2024)). This winter was characterized by prolonged high-pressure blocking over Northwestern Europe, which diverted the usual westerly flow and allowed persistent easterly winds to bring frigid air into the continent (Loikith and Neelin (2019)). Extensive snow cover reinforced the cold through high albedo effects, leading to the freezing of major

European rivers and lakes, including the Rhine, Rhône, IJsselmeer, and large parts of the Baltic Sea (Groisman et al. (1994)). The resulting extreme conditions had severe impacts on human health, infrastructure, and energy systems, highlighting the risks posed by such events even in a warming climate (Eichler (1970)).

This study builds on the work of (Sippel et al. (2024)), who investigated the question: "Could an extremely cold central European winter such as 1963 happen again despite climate change?" Their research addressed two key questions: (1) If a winter atmospheric circulation similar to 1963 were to re-occur in present-day climate, what would be the intensity in terms of cold temperatures? and (2) Is a winter as cold as 1963 or colder still possible in Central Europe today? The present study focuses in greater detail on the second question, assessing the potential for such an extremely cold winter in the current climate and its implications for future extreme events.

4.3 Data processing

In order to investigate potential worst-case cold winter conditions in Germany with a particular focus on Bavaria, we utilized the ERA5 reanalysis dataset (Hersbach et al. (2020)) to analyze temperature anomalies during the winter months, specifically from December to February (DJF). To detect anomalies, we applied a 90-day moving average to remove the seasonal cycle, using 1981–2010 as the reference period. Seasonal temperature anomalies were calculated from daily anomalies.

4.4 Methods

4.4.1 Dynamical Adjustment using Elastic Net Regression

Dynamical adjustment is a technique in climate science, that aims to estimate the influence of atmospheric circulation on a target surface climate variable, such as surface air temperature ((Wallace et al., 1995; Smoliak et al., 2015; Deser et al., 2016)). Here, we first apply dynamical adjustment to explore the influence of circulation patterns on temperature anomalies and to better understand the results obtained from other methods. Formally, the temperature anomaly at time t, denoted $T(t)$, is expressed as:

$$T(t) = T_{\text{circ}}(t) + T_{\text{resid}}(t),$$

where:

1. **Circulation-induced component** ($T_{\text{circ}}(t)$) : Represents the part of the temperature anomaly that is driven by large-scale atmospheric circulation patterns.
2. **Residual component** ($T_{\text{res}}(t)$) : Captures thermodynamical effects, including externally forced warming and other unknown influences not explained by circulation patterns.

Atmospheric circulation is typically difficult to measure directly, as it is not a single, easily defined quantity. Instead, it influences observable variables such as sea level pressure (SLP) and geopotential height, which are commonly used as proxies for large-scale circulation ((Smoliak et al., 2015; Sippel et al., 2019)). These variables capture essential aspects of circulation patterns, including the strength and position of high- and low-pressure systems, the configuration of jet streams, and the occurrence of blocking events. To extract the circulation-induced component, we use sea level pressure (SLP) patterns as a proxy for atmospheric circulation. By applying statistical regression techniques, we estimate the part of temperature variability

that these circulation patterns can explain. However, this method assumes a linear separation between circulation and thermodynamical effects. In reality, climate processes can be more complex, making this a limitation of the approach.

In our study, we pursue two different approaches to dynamical adjustment. Both methods aim to estimate the circulation-induced component of daily mean winter temperature over our study region Bavaria,using a regularized linear regression technique, called "elastic net regression" (Zou and Hastie (2005)). The first approach is based on the ERA5 reanalysis dataset, where an elastic net regression model is trained using sea level pressure (SLP) grid cells as predictors.

We also use a second dynamical adjustment approach, in which the regression model is trained on the CESM2-LE, using the same predictors as in the ERA5-based model. We subtract the domain-average mean trend of geopotential height patterns to account for the long-term column expansion due to warming (Sippel et al. (2024)), which allows for a greater focus on the interannual variability of atmospheric circulation and its impact on temperature. The resulting regression model, trained entirely on CESM2-LE, is independent of the observational data and is subsequently applied to the ERA5 dataset for comparison.

4.4.1.1 Elastic Net Regression

The model estimates the coefficient vector β by minimizing the following penalized least squares objective function:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \left[\alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \right] \right\}$$

where:

- $\mathbf{y}$ denotes the vector of surface temperature anomalies,
- $\mathbf{X}$ represents the matrix of circulation-related predictors (e.g., gridded SLP values),
- $\lambda \geq 0$ is a tuning parameter controlling the overall penalty strength,
- $\alpha \in [0, 1]$ determines the balance between the L1 and L2 penalties.

This formulation combines two regularization methods: Lasso (L1) and Ridge (L2) regression. The L1 penalty results in sparsity by shrinking specific coefficients to zero exactly, which can be interpreted as making variable selection by including only the most important predictors in the model. The L2 penalty shrinks the coefficients more evenly and stabilizes the model if predictors are highly correlated. By adjusting the mixing parameter α, Elastic net blends the advantages of both methods, enabling variable selection while maintaining model stability and predictive accuracy in multicollinearity. These properties make elastic nets particularly suitable for modeling temperature responses to spatially structured and interdependent circulation fields.

4.4.2 Ensemble Boosting

Cold extremes pose significant challenges in climate science due to their substantial socio-economic impacts. Traditional climate models struggle to capture the rare and intense nature of these events, necessitating advanced methodologies such as ensemble boosting. Hence, (Sippel et al. (2024)) explored the principles of ensemble boosting and its application to evaluate whether a worst-case cold winter such as 1963 is still possible. They focused on a 30-member CESM2 initial condition large ensemble (CESM2-ETH) from 2005 to 2035 to generate physically plausible worst-case scenarios of extremely cold winters.

Ensemble boosting is a technique designed to enhance the representation of extreme weather and climate events in model simulations. The core concept involves perturbing an initial state within a climate model,

allowing different yet physically consistent realizations of an extreme event. By systematically re-initializing the model with minuscule perturbations, it becomes possible to explore the tail behavior of the event distribution. In the context of climate modeling, boosting follows a two-step approach:

1. **First-order boosting** – Re-initialization occurs approximately 5-20 days before an identified extreme event using a round-off perturbation. This yields multiple ensemble members, each evolving uniquely but within the constraints of atmospheric dynamics.
2. **Second-order boosting** – After identifying the coldest simulations from the first-order boosted ensemble, additional perturbations are applied to these extreme cases, further refining the representation of worst-case scenarios.

This approach enables a more comprehensive understanding of potential extreme cold events by expanding the dataset of plausible realizations beyond those found in standard climate model ensembles, such as single model initial-condition large ensemble (SMILEs).

4.4.2.1 Data and Methodology of Ensemble Boosting

In the study of (Sippel et al. (2024)), the CESM2-ETH large ensemble, spanning 900 winter seasons (December-January-February, DJF) from 2005 to 2035, serves as the foundational dataset. This dataset follows the CMIP6 historical forcing (2005-2014) and the SSP3-7.0 scenario (2015-2035). Each of the 30 ensemble members originates from a transient historical simulation with a round-off perturbation in atmospheric initial conditions. To analyze extreme cold events, a boosting methodology was applied:

- **First-order boosting**: The coldest December during the 2020s in the CESM2-ETH ensemble was identified. This simulation was then perturbed and re-initialized for each day from December 1-15, generating 50 ensemble members per day. This resulted in a total of 750 simulations, capturing a well-constrained representation of early winter cold conditions.
- **Second-order boosting**: To further explore extreme cold persistence into January, the two coldest simulations from the first-order boosted set were selected. These were subsequently re-initialized daily from January 1-15, with 50 ensemble members per day, leading to 1500 additional simulations.

The perturbation methodology maintained physical consistency by applying small modifications to the specific humidity field (q) at each grid point, with a magnitude of 10^{-13}. These perturbations ensured mass, energy, and momentum conservation up to the precision of a round-off error. The coupled model was then run for 60 days, with ensemble spread remaining small for the first 4-5 days before diverging significantly.

Instead of replicating the exact methods used by (Sippel et al. (2024)), this study focuses on leveraging the provided supplementary boosted data by (Sippel et al. (2024)) to examine the most extreme cold storylines. Specifically, the study analyzed: (i) The three coldest storylines (minimum temperature and average temperature) from the BSSP370cmip6.0480013.zip dataset. (ii) The three coldest storylines (minimum temperature and average temperature) from the BSSP370cmip6.0230013.zip dataset.

These datasets consist of (i) first-order boosting simulations originating from ensemble member 13 of CESM2-ETH, initialized on December 6, 2022, and December 15, 2022, as well as second-order boosting simulations branching off from specific first-order boosted members ensemble member 23 on December 6, 2022, and ensemble member 48 on December 15, 2022). By analyzing these datasets, this study aims to answer the research question whether winters such as 1963 are still possible in today's climate.

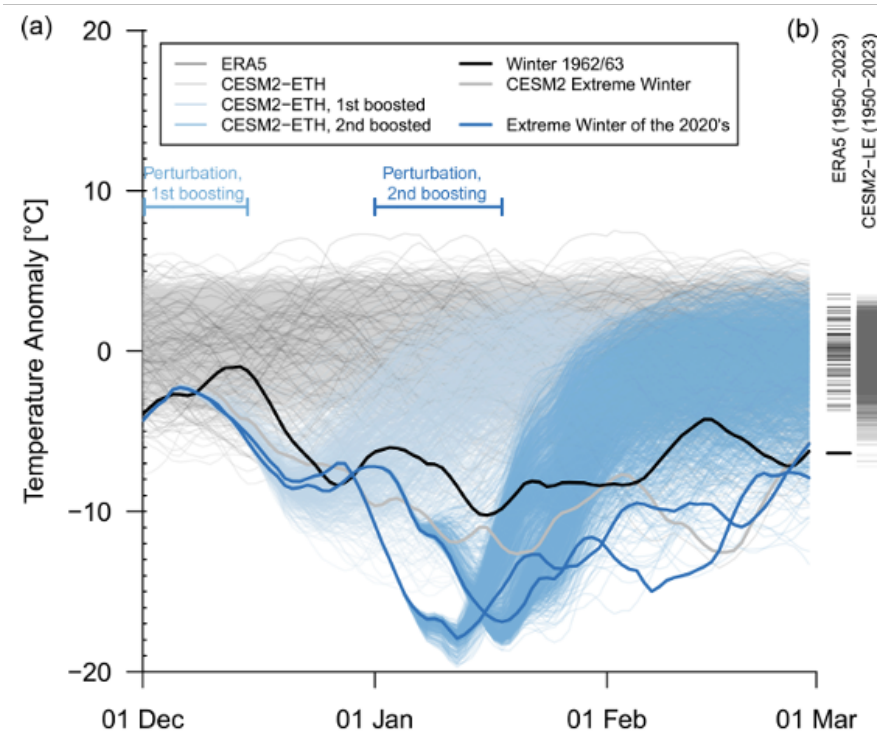


FIGURE 4.1: This figure (a) provides an illustrative example of model boosting, adapted from Sippel et al. (2024)

4.5 Results and Discussion

Here, the key findings derived from the applied methods are presented, followed by a reflection on their scientific implications and a discussion of the methodological limitations and uncertainties involved in the analysis.

4.5.1 Dynamical Adjustment

The circulation-induced component of temperature variability is clearly separated from the residual component, which is not explained by circulation and likely reflects thermodynamical effects (Fig.4.2). The residual time series shows a consistent upward trend (Fig.4.2 bottom), indicating that thermodynamical warming plays a significant role in addition to circulation changes. Besides, the circulation-induced variability shows a strong Pearson correlation of R = 0.8 with the observed, detrended DJF temperature anomalies over Bavaria (Fig.4.2 top), thus supporting the conclusion that circulation is the main driver of inter-annual winter temperature variability as suggested by the reference study. (Sippel et al. (2024))

- **Top:** 1951–2024 winter (DJF) temperature anomalies and the contribution of atmospheric circulation (blue line)
- **Bottom:** residual temperature anomaly time series when atmospheric circulation contributions are removed and the trend of this "circulation conditional" residual.

The dark blue line (adjusted using ERA5) shows a clear upward trend, suggesting a decrease in the frequency of cold spells (Fig.4.2 top). This indicates that, in addition to thermodynamical effects, changes in

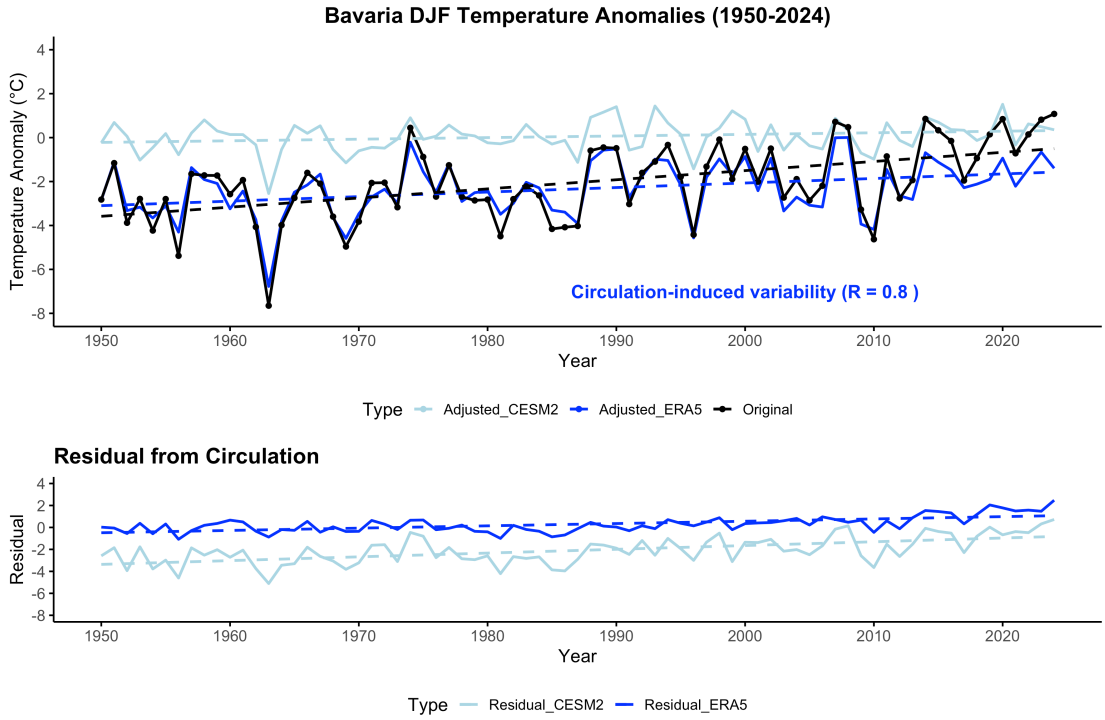


FIGURE 4.2: Winter temperature anomaly time series over Bavaria and long-term trends, dashed lines show linear trends in the original time series (black) and the circulation-induced and residual component (blue).

atmospheric circulation have also contributed to winter warming over Bavaria. However, the CESM2-based light blue line remains relatively flat, showing little to no evidence of strong externally forced changes and suggesting that the future of regional circulation changes under external forcing remains highly uncertain.

This raises a critical uncertainty, as discussed in (Sippel et al. (2024)): whether the circulation trend observed in recent decades represents an externally forced signal or just natural variability. If the trend is indeed forced but not captured by the models, extreme cold winters like 1963 may become less likely in the future. Conversely, if the observed trend is primarily due to natural variability, it could reverse, and similar cold extremes may occur again. This uncertainty remains a key challenge in climate modeling, and understanding it better is crucial for improving future predictions.

4.5.2 Ensemble Boosting

The three coldest storylines, in terms of both minimum temperature and average temperature, were analyzed for ensemble members 23 and 48 from the first-order boosted simulations.

For ensemble member 23, the lowest recorded temperatures in the second-order boosted simulations reached a minimum temperature of -26.3°C, while the average temperature over the winter period from January to March was -12.1°C. Figure 4.3 illustrates the three coldest minimum temperature storylines in the left panel and the three coldest average temperature storylines in the right panel. The ensemble members associated with these extreme conditions-ens010, ens040, and ens042 for minimum temperature, and ens037, ens039, and ens047 for average temperature—exhibit pronounced cold events, highlighting the capacity of the boosting technique to explore the statistical tail of extreme winter conditions.

For ensemble member 48, the coldest recorded temperatures in the second-order boosted simulations included a minimum temperature of -25.9°C, which consistently occurred at the start of the initialization, and an

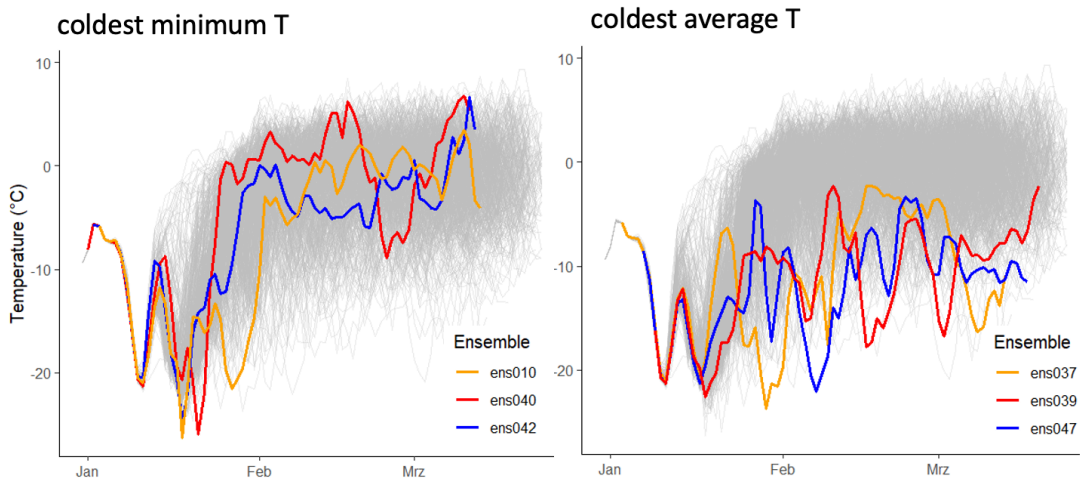


FIGURE 4.3: The three coldest storylines (based on minimum and average temperature) are derived from ensemble member 23. The data are extracted from the file BSSP370cmip6.0230013.zip, provided by Sippel et al. (2024)

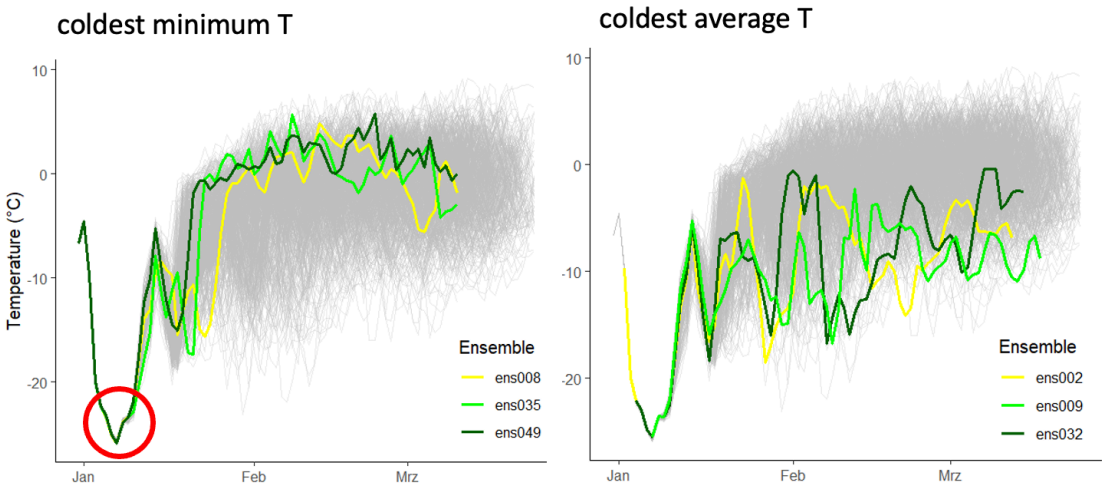


FIGURE 4.4: The three coldest storylines (based on minimum and average temperature) are derived from ensemble member 48. The data are extracted from the file BSSP370cmip6.0480013.zip, provided by Sippel et al. (2024)

average temperature of -10.2°C over the winter period from January to March (Fig.4.4). The fact that the lowest minimum temperatures always appeared at the beginning of the initialization phase suggests potential influences of edge conditions, temporal dependencies, or initialization bias. To address these factors, early January was excluded from the analysis. When focusing on mid-to-late winter, the lowest minimum temperature was observed in simulation ens037, reaching -23.1°C.

The ensemble boosting results demonstrate that extremely cold temperatures, such as in the winter of 1962/1963 or colder, can still be reached today under an SSP3-7.0 scenario. This contrasts with the findings by (Quesada et al. (2023)), who observed a general decline in the frequency and severity of cold events across Europe. However, the presented results align with other studies suggesting that under certain conditions—such as a weakening Atlantic Meridional Overturning Circulation (AMOC)—cold extreme intensities may increase in the future (Meccia et al. (2023)). This apparent contradiction illustrates the complex and multifaceted nature of cold event dynamics. It also reflects the low seasonal predictability identified in recent research, which attributes this uncertainty to chaotic atmospheric forcings, such as variability in westerly winds (Kautz et al. (2021)). The finding that extreme cold events remain possible is further supported by (Brunner et al. (2018)), who showed that approximately 70% of central European cold extremes coincide with atmospheric blocking between 60°W and 30°E—highlighting the continued relevance of large-scale weather patterns as a dominant driver.

Last, there are important limitations to consider. The distribution and frequency of “boosted” cold events are sensitive to the specific events selected for enhancement, which may influence the representativeness of the results. Additionally, re-initializing the model with different atmospheric conditions could potentially generate even colder outcomes, indicating that the full range of extreme winter possibilities might not be fully captured.

5

Introduction

Author:

Supervisor:

5.1 Intro About the Seminar Topic

5.2 Outline of the Booklet

6

Introduction

Author:

Supervisor:

6.1 Intro About the Seminar Topic

6.2 Outline of the Booklet

7

Co-occurence of extremes

Author: Chongjun

Supervisor: Henri Funk

7.1 Abstract

This report investigates the spatial dependence of extreme co-occurrence globally by employing hypothesis testing. The analysis aims to assess whether extreme events, such as temperature and precipitation extremes, exhibit spatial dependencies across the Earth's land surface. Using observational data from the Climatic Research Unit (CRU), the Berkeley Earth (BE) dataset, and the Global Precipitation Climatology Centre (GPCC), we explore both positive and negative spatial dependencies between extreme temperature and precipitation events. The study quantifies the intensity of spatial dependence by introducing the Probability Multiplication Factor (PMF), which provides a numerical representation of the strength of the spatial correlation between extreme events. Furthermore, we examine how spatial dependence varies with distance and latitude, offering insights into the geographic and climatological factors that influence the co-occurrence of extreme events. The findings demonstrate significant patterns of both positive and negative spatial dependencies, highlighting the complex relationships between extreme weather events and their spatial distribution across different regions.

7.2 Introduction

Climate extremes, such as heatwaves, heavy precipitation events, droughts, and tropical cyclones, pose severe threats to both natural and human systems. Their intensification in recent decades has led to escalating economic losses, ecosystem disruptions, and public health crises. These extremes are not random or isolated occurrences but exhibit distinct statistical dependencies. While extreme events can cluster over time (temporal dependence), they are also spatially dependent, meaning that extreme conditions in one region are often correlated with those in neighboring or even distant areas due to large-scale climate dynamics. Understanding this spatial dependence is crucial for improving risk assessments, predicting multi-region impacts, and formulating climate adaptation strategies.

Although temporal dependence plays a role in the persistence of climate extremes, this report focuses exclusively on their spatial dependence—the extent to which extreme climate conditions co-occur across different locations. Spatial dependence arises due to shared meteorological drivers, such as atmospheric circulation patterns, teleconnections, and geographic influences. For instance, heatwaves are often linked across multiple regions by high-pressure blocking systems, while extreme precipitation events can span large areas due to stalled weather fronts. Recognizing and quantifying these spatial relationships is critical for accurately modeling extreme event risks and preparing for simultaneous disasters that affect multiple regions.

One striking example of spatial dependence in climate extremes is the 2010 Eastern Russian heatwave,

which was not confined to Eastern Russia alone but also influenced temperatures across Eastern Europe and Central Asia. The event was driven by a persistent atmospheric blocking pattern, which sustained extreme temperatures over an extensive spatial scale.

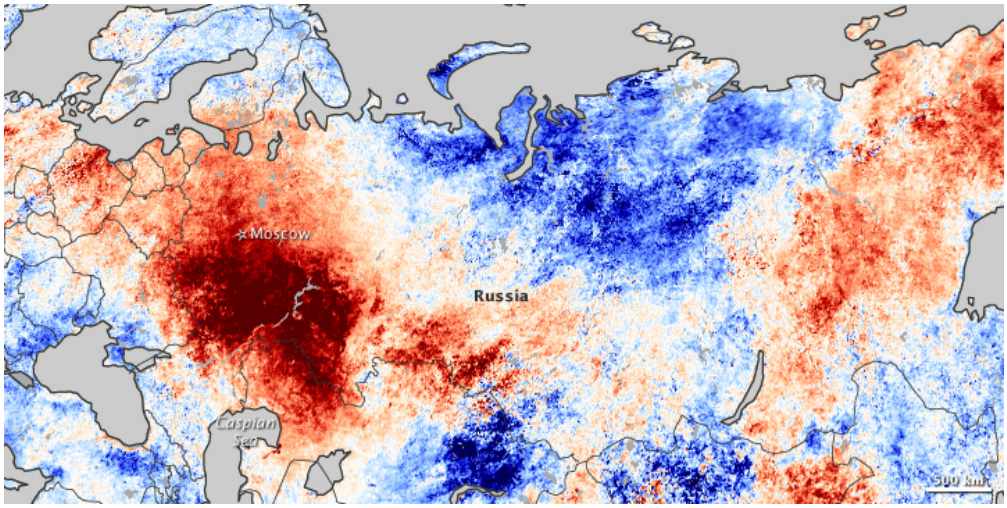


FIGURE 7.1: Source: [https://earthobservatory.nasa.gov/images/45069/heatwave-in-russia]. Heat waves in Eastern Russia, Eastern Europe and Central Asia in 2010.

Another example is tropical cyclones in the Atlantic, where clusters of hurricanes can develop due to favorable oceanic and atmospheric conditions, causing sequential or simultaneous destruction across multiple coastal regions. Similarly, compound flooding events, where storm surges coincide with extreme rainfall across interconnected watersheds, demonstrate the importance of spatially correlated extremes in shaping disaster impacts.

The study of spatial dependence in climate extremes has gained increasing attention as traditional models often assume independent occurrences, leading to the underestimation of multi-region risks. Modern statistical approaches, such as extreme value theory (EVT), copula models, and geostatistical methods, provide powerful tools for quantifying spatial dependence and improving predictions of concurrent extreme events. Additionally, the role of climate teleconnections, such as the El Niño-Southern Oscillation (ENSO) and the North Atlantic Oscillation (NAO), further highlights how distant regions can experience correlated extremes.

Despite advancements in understanding spatial dependencies, significant challenges remain. Non-stationarity in climate extremes—driven by global warming, land-use changes, and shifting atmospheric patterns—complicates the identification of stable spatial dependence structures. Furthermore, as climate extremes become more frequent and intense, their spatial footprints may expand, necessitating continuous refinement of statistical models. Addressing these challenges is crucial for enhancing resilience against climate risks and improving predictive capabilities in disaster-prone regions.

This report will discuss methodologies for quantifying the spatial dependence of climate extremes and examine key characteristics of this dependence on a global scale. By deepening our understanding of how extreme climate events co-occur across space, we can better identify large-scale patterns, improve predictive models, and refine risk assessments to support climate adaptation strategies.

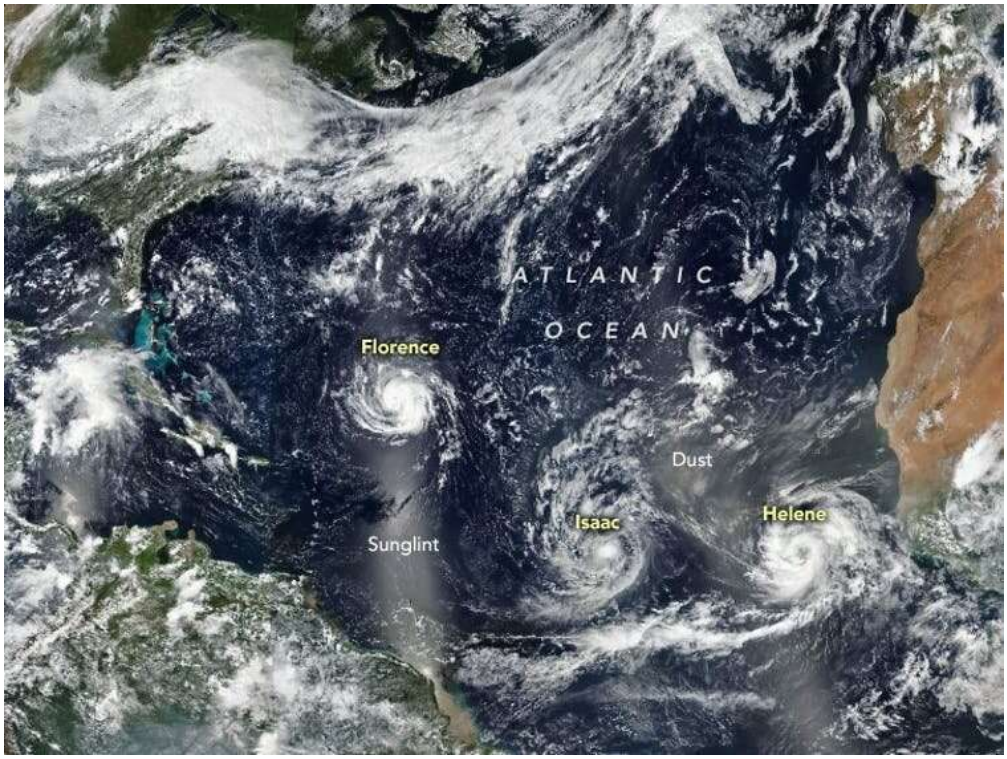


FIGURE 7.2: Source: [https://phys.org/news/2019-03-big-storm-clusters-hurricane-hotspots.html]. Hurricanes occur simultaneous in the Atlantic in 2019.

7.3 Tail Dependence

Tail dependence refers to the relationship between the extreme values of two or more variables, specifically whether extreme values of one variable tend to coincide with extreme values of another variable. Unlike traditional correlation $r(G_i, G_j)$, which measures the linear relationship between variables, tail dependence focuses on the co-occurrence of extreme events, such as the occurrence of severe climate conditions in different regions simultaneously. The objective of this research is not to analyze correlation between regions but to explore how extreme values in one region might depend on extreme values in another region, even if there is no linear correlation.

Statistically, tail dependence is a measure used to quantify the likelihood of simultaneous extreme values in the tails of distributions. There are various methods to measure tail dependence, with the most commonly used being the upper and lower tail dependence coefficients. These coefficients describe the probability of one variable exceeding a threshold given that the other variable has already exceeded a similar threshold, capturing the essence of tail events and their joint occurrences. Tail dependence is crucial in risk management, particularly in fields such as finance and climate science, where extreme events (e.g., market crashes or heatwaves) can have substantial consequences.

It is important to note that while correlation can provide insights into the likelihood of tail dependence, a high correlation between two grid cells (G_i and G_j) suggests a higher probability of their extreme values occurring together. However, a high correlation does not necessarily imply that extreme values will co-occur, nor does the absence of correlation rule out the possibility of tail dependence. In other words, tail dependence can manifest even when two variables are not correlated, and vice versa.

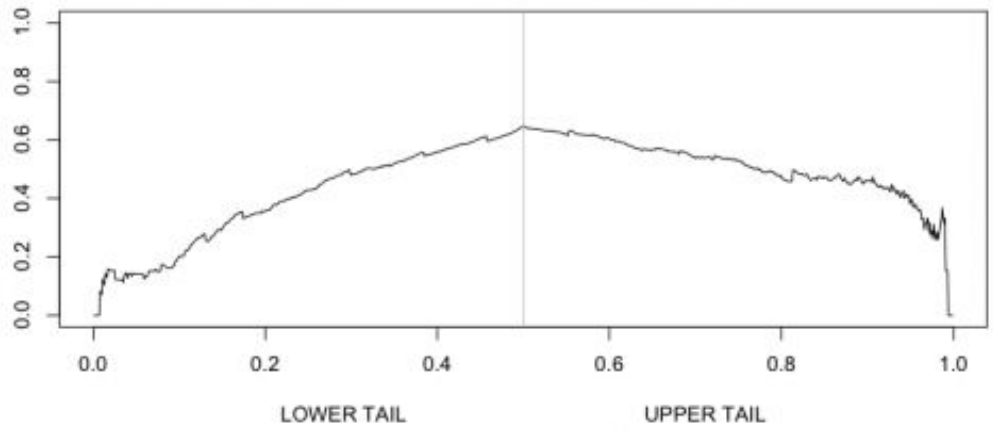


FIGURE 7.3: Source: [https://freakonometrics.hypotheses.org/2435]. Tail dependence of 2 variables.

7.4 Hypothesis Test

We conducted a hypothesis test to assess the spatial dependence of climate extremes. The null hypothesis, H_0, is that climate extremes occur independently at two grid cells, G_i and G_j, where $i \neq j$, at a significance level of $\alpha = 0.05$. We aim to identify spatial dependence by rejecting H_0. The identification of extremes in a grid cell is defined by whether the value exceeds the 95th percentile or falls below the 5th percentile of the distribution. Therefore, the probability that a single grid cell, G_i, is extreme is $P(G_i \text{ is extreme}) = 0.1$ (10%).

Next, we consider the joint probability of two grid cells, G_i and G_j, being extreme. Specifically, the probability that both G_i and G_j are extreme is calculated based on two cases: both cells exceeding the 95th percentile, or both cells falling below the 5th percentile. This defines positive concurrence of extremes. For negative concurrence, we consider the case where G_i is greater than the 95th percentile and G_j is less than the 5th percentile, or vice versa. In mathematical terms:

- Positive concurrence: $P(G_i > 0.95 \text{ and } G_j > 0.95)$ or $P(G_i < 0.05 \text{ and } G_j < 0.05)$
- Negative concurrence: $P(G_i > 0.95 \text{ and } G_j < 0.05)$ or $P(G_i < 0.05 \text{ and } G_j > 0.95)$

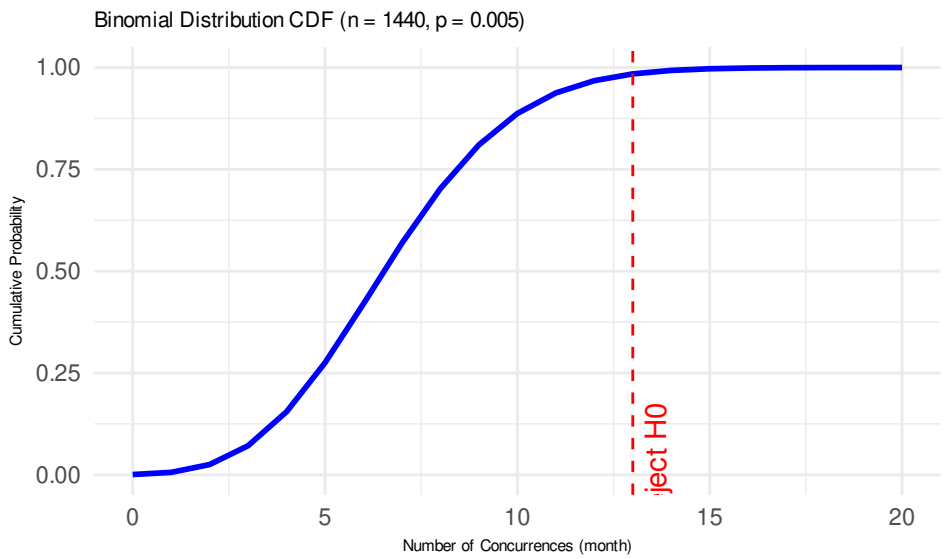
Under the assumption of no spatial dependence, the probability of either positive or negative concurrence, p_0, is computed as:

$$p_0 = 2 \times 0.0025 = 0.005$$

This is the probability of concurrent extremes occurring at both grid cells, regardless of whether the concurrence is positive or negative, under the null hypothesis of no spatial dependence.

Next, we consider a time frame of 120 years, or 1440 months. The expected number of months during which both grid cells experience concurrent extremes is $E(X) = n \times p_0 = 1440 \times 0.005 = 7.2$ months. To test for spatial dependence, we model the number of concurrent extreme months using a binomial distribution:

$$X \sim \text{Binomial}(n = 1440, p = 0.005)$$



Using the binomial distribution's cumulative probability, we reject H_0 if the number of months with concurrent extremes, x, is greater than or equal to 13. That is, we reject H_0 at the 0.05 significance level when $x \geq 13$. This corresponds to the condition where:

$$P(X \geq 13) \leq 0.05$$

Thus, if the number of months with concurrent extremes exceeds 13 in the 120-year period, we conclude that there is spatial dependence between the two grid cells. The corresponding probability threshold for rejecting H_0 is when $p_0 \geq 0.009$.

7.5 Probability Multiple Factor

In order to assess the strength of spatial dependence, the study introduces the Probability Multiplication Factor (PMF), which quantifies the degree of extreme concurrence between two grid cells. The PMF is defined as the ratio of the actual probability of extreme concurrence to the baseline probability of extreme concurrence under the null hypothesis, denoted as p_0. Mathematically, this can be expressed as:

$$\text{PMF} = \frac{P(\text{extreme concurrence})}{p_0}$$

The threshold for the PMF is calculated by dividing $p_0 = 0.005$ by the baseline probability of extreme concurrence $p_0 = 0.005$, yielding a threshold value of:

$$\frac{0.009}{0.005} = 1.8$$

When the PMF exceeds or equals this threshold, i.e., when:

$$\text{PMF} \geq 1.8$$

it indicates the presence of spatially dependent extremes. Furthermore, the larger the PMF, the stronger the spatial dependence between the extreme events in the two grid cells.

7.6 Results

The study utilizes temperature and precipitation products from the Climatic Research Unit (CRU), the temperature product from Berkeley Earth (BE), and the precipitation product from the Global Precipitation Climatology Centre (GPCC). These datasets provide time series data spanning a 120-year period (1901–2020), enabling a comprehensive analysis of extreme event co-occurrence and spatial dependence.

The study divides the globe into $2° \times 2°$ grid cells, selecting only those with a land area greater than 30% for analysis. Consequently, the research primarily focuses on spatial dependence over land, excluding oceanic regions.

Analysis of temperature and precipitation extremes reveals distinct patterns of spatial dependence. Temperature extremes predominantly exhibit positive spatial co-occurrence, with a mean Probability Multiplication Factor (PMF) of 3.25 across 58% of grid cell pairs. In contrast, negative dependence is detected in only 15% of grid cell pairs. Precipitation extremes display both positive and negative spatial dependencies, observed in 27% and 25% of grid cell pairs, respectively, with mean PMF values of 2.64 and 2.40.

The strength of positive concurrence is particularly pronounced in geographically proximate regions and diminishes with increasing distance. Specifically, for both temperature (Figure 4. A) and precipitation (Figure 4. B) extremes, PMF remains consistently higher when the distance between grid cells is less than 5000 km. Beyond this threshold, PMF exhibits only a marginal decline as distance increases, suggesting a limited attenuation of spatial dependence at larger scales. However, in cases of extreme distances, a PMF exceeding 1.8 alone is not sufficient to confirm spatial dependence, as atmospheric circulation patterns, including heat advection and precipitation transport, can facilitate the transfer of extreme conditions between distant grid cells within a monthly timeframe. Nevertheless, the fact that PMF does not significantly decrease beyond 5000 km indicates that distance is not the primary determinant of spatial dependence. This implies the presence of additional underlying factors contributing to the strong spatial correlation observed between geographically distant regions.

The negative concurrence strength of temperature extremes (Figure 4. C) is also pronounced in neighboring regions. In contrast, no distinct spatial pattern is observed for the negative concurrence of precipitation extremes (Figure 4. D).

The study further identified that extreme events tend to co-occur more strongly in the tropics and high-latitude regions of both hemispheres. In the analysis, solid lines represent the probability mass function (PMF) of regions with positive concurrence, while dashed lines indicate regions with negative concurrence. For both temperature and precipitation extremes, the mean PMF for negative concurrence remains consistently lower than that for positive concurrence, suggesting that positively dependent extremes are more prominent.

The impact of long-term warming trends can be evaluated by comparing the red and blue lines in the figures. In the tropics, the strong positive dependence of temperature extremes (Figure 5.E) is largely due to warming trends, as indicated by PMF values reaching up to 6 when the trend is included. However, after removing the trend, the PMF drops to 4.5, emphasizing the substantial influence of warming on the observed concurrence of temperature extremes. In contrast, negative dependence in temperature extremes is more commonly found in the northern high-latitude regions. The historical simulations from CMIP6 models show a closer match with observed data when trends are included, whereas the historical natural (hist nat) simulations better align with observations when trends are removed.

For precipitation extremes (Figure 5.F), the spatial pattern of dependence is similar to that of temperature extremes. However, in this case, long-term warming trends have only a minor effect on PMF, suggesting that other factors play a more significant role in shaping the spatial dependence of precipitation extremes. Another key difference from temperature extremes is that the results from historical simulations and historical natural

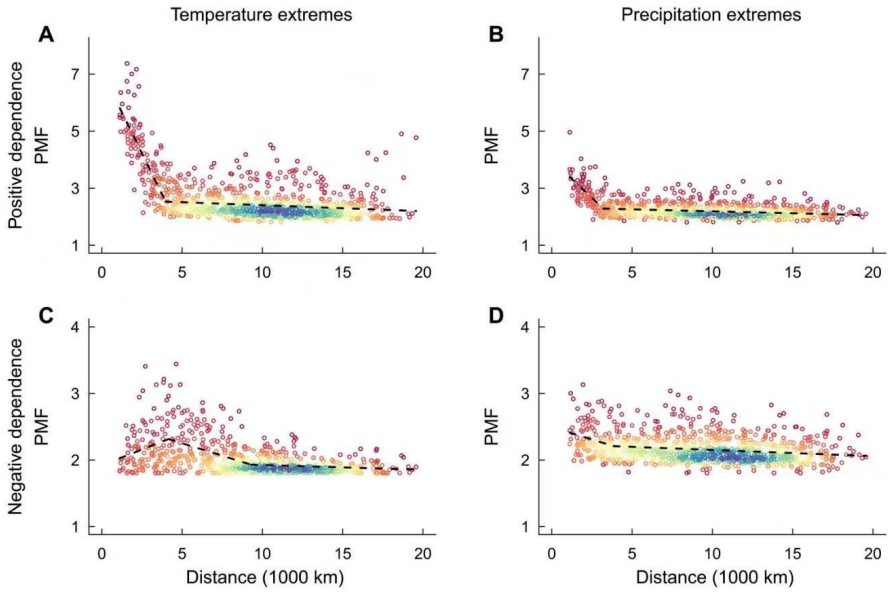


FIGURE 7.4: Source: [https://www.science.org/doi/10.1126/sciadv.abo1638]. The distribution of PMF values across different grid cell pairs at varying distances. Each scatter point on the plot represents the PMF of a grid cell pair, while the color indicates the density of point distribution. Blue signifies the highest density, indicating more concentrated pairs, whereas red represents the lowest density, corresponding to more dispersed pairs. The data used is observed data.

(hist nat) simulations are nearly identical. This suggests that the warming trend does not substantially alter the spatial dependence of precipitation extremes, unlike its effect on temperature extremes.

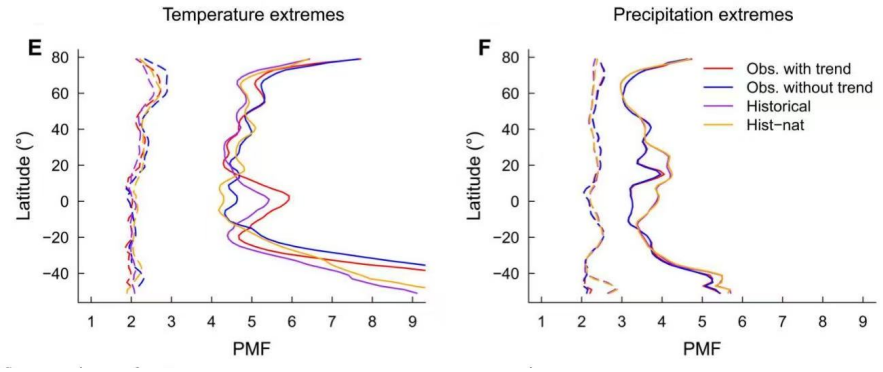


FIGURE 7.5: Source: [https://www.science.org/doi/10.1126/sciadv.abo1638]. The distribution of PMF values across different grid cell pairs at varying latitudes. Solid lines represent the PMF of positively concurrent regions, dashed lines indicate negatively concurrent regions. Four scenarios: observation with/without trend, historical simulated data and hist-nat simulated data (natural forcing only).

7.7 Sensitivity Test

This study also investigated how spatial dependence responds to different thresholds used to define extreme events. The hypothesis test was conducted with a threshold of 0.05/0.95, while the sensitivity analysis explored a range from 0.01/0.99 to 0.1/0.9. The 0.01/0.99 threshold represents the strictest criterion, meaning only the most extreme events are considered, while the 0.1/0.9 threshold is the least strict, allowing more events to be included. The findings suggest that as the threshold becomes stricter, fewer grid cell pairs exhibit extreme concurrence, leading to a smaller number of identified regions. At the same time, the mean probability mass function (PMF) increases, indicating that spatial dependence is stronger under stricter threshold conditions.

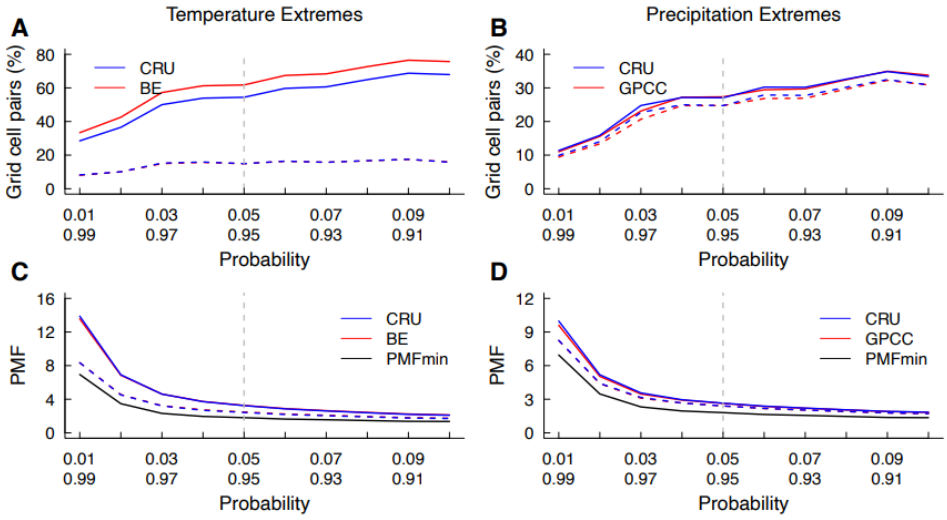


FIGURE 7.6: Source: [https://www.science.org/doi/10.1126/sciadv.abo1638]. The proportion of land grid cell pairs where climate extremes exhibit significant positive (solid lines) or negative (dashed lines) dependence, assessed using probability thresholds ranging from 0.9/0.1 to 0.99/0.01, based on observational data from 1901 to 2020. (C,D) Corresponding results for (A,B), but depicting the mean probability multiplication factor (PMF) for positively (solid lines) and negatively (dashed lines) dependent climate extremes.

8

Introduction

Author:

Supervisor:

8.1 Intro About the Seminar Topic

8.2 Outline of the Booklet

9

Acknowledgements

The most important contributions are from the students themselves. The success of such projects highly depends on the students. And this book is a success, so thanks a lot to all the authors! The other important role is the supervisor. Thanks to all the supervisors who participated! Special thanks to Helmut Küchenhoff[1] who enabled us to conduct the seminar in such an experimental way, supported us and gave valuable feedback for the seminar structure. Thanks a lot as well to the entire Department of Statistics[2] and the LMU Munich[3] for the infrastructure.

The authors of this work take full responsibilities for its content.

[1] https://www.stablab.stat.uni-muenchen.de/personen/leitung/kuechenhoff1/index.html
[2] https://www.statistik.uni-muenchen.de/
[3] http://www.en.uni-muenchen.de/index.html

Bibliography

(2008). Climate models: an assessment of strengths and limitations. Available online: https://calhoun.nps.edu/bitstream/handle/10945/40146/climate%20models_%20an%20assessment%20of%20strengths%20and%20limitations.pdf;sequence=4.

(2013). Evaluation of climate models. Available online: https://pure.mpg.de/pubman/faces/viewitemoverviewpage.jsp?itemid=item_1977534.

Alberg, A. J., Park, J. W., Hager, B. W., Brock, M. V., and Diener-West, M. (2004). The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *Journal of general internal medicine*, 19(5p1):460–465.

Anagnostopoulou, C., Tolika, K., Lazoglou, G., and Maheras, P. (2017). The exceptionally cold january of 2017 over the balkan peninsula: A climatological and synoptic analysis. *Atmosphere*, 8(12):252.

Baur, F., Hess, P., and Nagel, H. (1944). *Kalender der Großwetterlagen Europas 1881–1939*. Forschungsinstitut für langfristige Witterungsvorhersage, Bad Homburg.

Bissolli, P. (2001). Wetterlagen und großwetterlagen im 20. jahrhundert. *DWD Klimastatusbericht*.

Bissolli, P. and Dittmann, E. (2001). The objective weather type classification of the german weather service and its possibilities of application to environmental and meteorological investigations. *Meteorologische Zeitschrift*, 10(4):253–260.

Boers, N., Goswami, B., Rheinwalt, A., Bookhagen, B., Hoskins, B., and Kurths, J. (2019). Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566(7744):373–377.

Brunner, L., Schaller, N., Anstey, J., Sillmann, J., and Steiner, A. K. (2018). Dependence of present and future european temperature extremes on the location of atmospheric blocking. *Geophysical research letters*, 45(12):6311–6320.

Chauhan, T., Chandel, V., and Ghosh, S. (2024). Global land drought hubs confounded by teleconnection hotspots in equatorial oceans. *npj Clim Atmos Sci*, 7(1):1–11.

Deser, C., Terray, L., and Phillips, A. S. (2016). Forced and internal components of winter air temperature trends over north america during the past 50 years: Mechanisms and implications. *Journal of Climate*, 29(6):2237–2258.

Di Lorenzo, E., Combes, V., Keister, J., Strub, P. T., Thomas, A., Franks, P., et al. (2013). Synthesis of pacific ocean climate and ecosystem dynamics. *Oceanography*, 26(4):68–81.

Dietterich, T. G. et al. (2002). Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125.

Edwards, P. N. (2011). History of climate modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 2(1):128–139.

Eichler, W. (1970). Der strenge winter 1962/1963 und seine vielschichtigen biologischen auswirkungen in mitteleuropa. Zool.-Bot. Ges. Österreich. Last accessed: 15 July 2024.

European Centre for Medium-Range Weather Forecasts (2017). ERA5: Data Documentation.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016a). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., et al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, 9(2):102–110.

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., et al. (2016b). Esmvaltool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of earth system models in cmip. *Geoscientific Model Development*, 9(5):1747–1802.

Friedlingstein, O. B., Webb, M., Gregory, J., et al. (2008). A summary of the cmip5 experiment design.

Garnett, R. (2023). *Bayesian optimization*. Cambridge University Press.

Government of Canada (2025). Canadian climate data and scenarios - canesm2 predictions.

Groisman, P. Y., Karl, T. R., and Knight, R. W. (1994). Observed impact of snow cover on the heat balance and the rise of continental spring temperatures. *Science*, 263(5144):198–200.

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, 27(9):1119–1152.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049.

Hess, P. (2005). Katalog der grosswetterlagen europas nach paul hess und helmuth brezowsky:(1881-2004).

Hess, P. and Brezowsky, H. (1952). *Katalog der Großwetterlagen Europas*. Number 33 in Berichte des Deutschen Wetterdienstes in der US-Zone. Deutscher Wetterdienst.

Hua, W., Chen, H., Sun, S., and Zhou, L. (2015). Assessing climatic impacts of future land use and land cover change projected with the canesm2 model. *International Journal of Climatology*, 35(12).

Häckel, H. (2021). *Meteorologie*. Verlag Eugen Ulmer, Stuttgart, 9., vollständig überarbeitete und erweiterte auflage edition.

Karpechko, A. Y., Charlton-Perez, A., Balmaseda, M., Tyrrell, N., and Vitart, F. (2018). Predicting sudden stratospheric warming 2018 and its climate impacts with a multimodel ensemble. *Geophysical Research Letters*, 45(24):13–538.

Kautz, L.-A., Martius, O., Pfahl, S., Pinto, J. G., Ramos, A. M., Sousa, P. M., and Woollings, T. (2021). Atmospheric blocking and weather extremes over the euro-atlantic sector–a review. *Weather and Climate Dynamics Discussions*, 2021:1–43.

Lau, W. K. M. and Kim, K.-M. (2012). The 2010 pakistan flood and russian heat wave: Teleconnection of hydrometeorological extremes. *Journal of Hydrometeorology*, 13(1):392–403.

Loikith, P. C. and Neelin, J. D. (2019). Non-gaussian cold-side temperature distribution tails and associated synoptic meteorology. *Journal of Climate*, 32(23):8399–8414.

Meccia, V. L., Fuentes-Franco, R., Davini, P., Bellomo, K., Fabiano, F., Yang, S., and von Hardenberg, J. (2023). Internal multi-centennial variability of the atlantic meridional overturning circulation simulated by ec-earth3. *Climate Dynamics*, 60(11):3695–3712.

Mehta, S., Paunwala, C., and Vaidya, B. (2019). Cnn based traffic sign classification using adam optimizer. In *2019 international conference on intelligent computing and control systems (ICCS)*, pages 1293–1298. IEEE.

Mittermeier, M., Weigert, M., Rügamer, D., Küchenhoff, H., and Ludwig, R. (2022). A deep learning based classification of atmospheric circulation types over europe: projection of future changes in a cmip6 large ensemble. *Environmental Research Letters*, 17(8):084021.

Moishin, M., Deo, R. C., Prasad, R., Raj, N., and Abdulla, S. (2021). Designing deep-based learning flood forecast model with convlstm hybrid algorithm. *Ieee Access*, 9:50982–50993.

Mueller, M. U., Ekhtiari, N., Almeida, R. M., and Rieke, C. (2020). Super-resolution of multispectral satellite images using convolutional neural networks. *arXiv preprint arXiv:2002.00580*.

Newton, A., Thunell, R., and Stott, L. (2006). Climate and hydrographic variability in the indo-pacific warm pool during the last millennium. *Geophysical Research Letters*, 33(19).

Nowack, P., Runge, J., Eyring, V., and Haigh, J. D. (2020). Causal networks for climate model evaluation and constrained projections. *Nature Communications*, 11(1):1415.

Opitz, J. and Burst, S. (2019). Macro f1 and macro f1. *arXiv preprint arXiv:1911.03347*.

Ozaki, Y., Tanigaki, Y., Watanabe, S., Nomura, M., and Onishi, M. (2022). Multiobjective tree-structured parzen estimator. *Journal of Artificial Intelligence Research*, 73:1209–1250.

Pinto, I., Rantanen, M., Ødemark, K., Tradowsky, J., Kjellström, E., Barnes, C., Otto, F., Heinrich, D., Pereira Marghidan, C., Vahlberg, M., et al. (2024). Extreme cold will still occur in northern europe, although less often–risking decreasing preparedness and higher vulnerability. *Cambridge University Press*, 10:108899.

Planchon, O., Quénol, H., Irimia, L., and Patriche, C. (2015). European cold wave during february 2012 and impacts in wine growing regions of moldavia (romania). *Theoretical and Applied Climatology*, 120:469–478.

Poli, P., Hersbach, H., Dee, D., Siminski, J., Laloyaux, P., Tan, D., Peubey, C., Thépaut, J.-N., Trémolet, Y., Hólm, E. V., Bonavita, M., Isaksen, L., and Fisher, M. (2016). Era-20c: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, 29(11):4083–4097.

Quesada, B., Vautard, R., and Yiou, P. (2023). Cold waves still matter: characteristics and associated climatic signals in europe. *Climatic Change*, 176(6):70.

Ropelewski, C. F. and Halpert, M. S. (1986). North american precipitation and temperature patterns associated with the el niño/southern oscillation (enso). *Monthly Weather Review*, 114(12):2352–2362.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in earth system sciences. *Nature Communications*, 10:2553.

Sanderson, M. (1999). The classification of climates from pythagoras to koeppen. *Bulletin of the American Meteorological Society*, 80(4):669–673.

Sippel, S., Barnes, C., Cadiou, C., Fischer, E., Kew, S., Kretschmer, M., Philip, S., Shepherd, T. G., Singh, J., Vautard, R., et al. (2024). Could an extremely cold central european winter such as 1963 happen again despite climate change? *Weather and Climate Dynamics*, 5(3):943–957.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A. G., Fischer, E., and Knutti, R. (2019). Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, 32(17):5677–5699.

Smoliak, B. V., Wallace, J. M., Lin, P., and Fu, Q. (2015). Dynamical adjustment of the northern hemisphere surface air temperature field: Methodology and application to observations. *Journal of Climate*, 28(4):1613–1629.

Wallace, J. M., Zhang, Y., and Renwick, J. A. (1995). Dynamic contribution to hemispheric mean temperature trends. *Science*, 270(5237):780–783.

Ward, R., Wu, X., and Bottou, L. (2020). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30.

Werner, P. C. and Gerstengarbe, F. W. (2010). *Katalog der Großwetterlagen Europas (1881–2009) nach Paul Hess und Helmut Brezowsky*. Number 119 in PIK Report. Potsdam-Institut für Klimafolgenforschung.

Wyser, K., Koenigk, T., Fladrich, U., Fuentes-Franco, R., Karami, M. P., and Kruschke, T. (2021). The smhi large ensemble (smhi-lens) with ec-earth3.3.1. *Geoscientific Model Development*, 14(9):4781–4796.

Yu, D., Deng, L., Yu, D., and Deng, L. (2015). Deep neural network-hidden markov model hybrid systems. *Automatic speech recognition: A deep learning approach*, pages 99–116.

Zhang, X. and Han, W. (2020). Effects of climate modes on interannual variability of upwelling in the tropical indian ocean. *Journal of Climate*, 33(4):1547–1573.

Zhong, Y., Chalise, P., and He, J. (2023). Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data. *Communications in statistics-simulation and computation*, 52(1):110–125.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. (2019). A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11127–11135.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.