Environmental Data Science (2024), 3: e9, 1–28
doi:10.1017/eds.2024.7

CAMBRIDGE UNIVERSITY PRESS

METHODS PAPER

Extreme heat wave sampling and prediction with analog Markov chain and comparisons with deep learning

George Miloshevich[1,2,3], Dario Lucente[1,4,5], Pascal Yiou[3] and Freddy Bouchet[1,6]

[1]ENSL, CNRS, Laboratoire de Physique, Lyon, France
[2]Centre for mathematical Plasma Astrophysics, Department of Mathematics, KU Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium
[3]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay & IPSL, CE Saclay Orme des Merisiers, Gif-sur-Yvette, France
[4]Department of Physics, University of Rome Sapienza, P.le Aldo Moro 2, 00185 Rome, Italy
[5]Institute for Complex Systems, CNR, P.le Aldo Moro 2, 00185 Rome, Italy
[6]Laboratoire de Météorologie Dynamique, UMR 8539 CNRS-ENS-X-Sorbonne Université, PSL & IPSL, Paris, France
Corresponding author: George Miloshevich; Email: George.miloshevich@kuleuven.be

Received: 07 July 2023; **Revised:** 20 January 2024; **Accepted:** 29 January 2024

Keywords: convolutional neural network; heat wave; Markov chain; prediction

Abstract

We present a data-driven emulator, a stochastic weather generator (SWG), suitable for estimating probabilities of prolonged heat waves in France and Scandinavia. This emulator is based on the method of analogs of circulation to which we add temperature and soil moisture as predictor fields. We train the emulator on an intermediate complexity climate model run and show that it is capable of predicting conditional probabilities (forecasting) of heat waves out of sample. Special attention is payed that this prediction is evaluated using a proper score appropriate for rare events. To accelerate the computation of analogs, dimensionality reduction techniques are applied and the performance is evaluated. The probabilistic prediction achieved with SWG is compared with the one achieved with a convolutional neural network (CNN). With the availability of hundreds of years of training data, CNNs perform better at the task of probabilistic prediction. In addition, we show that the SWG emulator trained on 80 years of data is capable of estimating extreme return times of order of thousands of years for heat waves longer than several days more precisely than the fit based on generalized extreme value distribution. Finally, the quality of its synthetic extreme teleconnection patterns obtained with SWG is studied. We showcase two examples of such synthetic teleconnection patterns for heat waves in France and Scandinavia that compare favorably to the very long climate model control run.

Impact Statement

Estimating conditional probabilities and rate of returns of extreme heat waves is important for climate change risk assessments. The most impactful heat waves are also long lasting. Numerical weather and climate models are expensive to run and may have biases, while observational datasets are too short to observe many of the potential extreme events and sample them accordingly. We present a relatively inexpensive yet powerful data-driven tool

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

for sampling and estimating probabilities of extreme prolonged heat waves and compare it with existing data-driven statistical approaches such as convolutional neural networks and extreme value theory estimations.

1. Introduction

It is expected that heat waves will be among the most impactful effects of climate change (Russo et al., 2015; Seneviratne et al., 2021) in the 21st century. While, generally, climate scientists have warned that global warming will increase the rate of heat waves, in some regions such as Europe, the trends have accelerated beyond what was expected (Christidis et al., 2020). Extreme heat waves, such as the Western European heat wave in 2003 (García-Herrera et al., 2010) and the Russian heat wave in 2010 (Barriopedro et al., 2011; Chen et al., 2022) had consequences on agriculture and posed health hazards. Such events have also affected other regions including Asia, for example, the Korean Peninsula (Choi et al., 2021) in the summers of 2013, 2016, and 2018 (Min et al., 2020). Thus, understanding the potential drivers, forecasting, and performing long-term projections of heat waves is necessary. These tasks, also relevant for climate change attribution studies (National Academies of Sciences Engineering and Medicine, 2016), are difficult since they require computation of small probabilities and thus massive ensemble simulations with expensive models or drawing conclusions from the relatively short observational record.

The possibility of using atmospheric circulation analogs for predicting weather dates back to the works of Bjerknes (1921) and Lorenz (1969). The idea is quite intuitive: one expects weather patterns to repeat given similar initial conditions. However, estimating the amount of data (a *catalog* of analogs) actually necessary for short-range predictions revealed the superiority of physics-based numerical weather prediction (NWP) models (van den Dool, 2007; Balaji, 2021). On the other hand, for predicting certain types of large-scale patterns, analog forecasting may require less resources and can be competitive with or superior to NWP on time scales longer than Lyapunov time, such as subseasonal-to-seasonal scales (Van den Dool et al., 2003; Cohen et al., 2019) and especially long-range predictions. The examples include prediction of ENSO (Ding et al., 2019; Wang et al., 2020) and Madden–Julian oscillation (Krouma et al., 2023).

Analog forecasting involves constructing potential trajectories that the system could have explored from a given state. These methods, which assume a Markov property (Yiou, 2014), belong to the family of methods referred to as stochastic weather generators (SWGs). SWGs can be used as climate model emulators due to their ability to generate long sequences that can infer statistical properties of spatio-temporal dynamics and correlation structures of the system without running expensive general circulation models (GCMs) (Ailliot et al., 2015). Other applications include downscaling (Wilks, 1992; Rajagopalan and Lall, 1999) and data assimilation (Lguensat et al., 2017).

In the last decade, there has been a rapid advancement of other types of data-driven forecasting/emulators in earth system models based on deep learning (Reichstein et al., 2019). For example, in a seminal work (Ham et al., 2019), convolutional neural network (CNN) has achieved a positive skill for El Niño–Southern Oscillation (ENSO) prediction compared to NWP. These success stories are sometimes accompanied by similar skills displayed by the analog method (Ding et al., 2019; Wang et al., 2020). Recently, deep learning has been used to predict extreme heat waves (Chattopadhyay et al., 2020; Jacques-Dumas et al., 2022; Lopez-Gomez et al., 2022) targeting categorical scores related to hit rates[1]. In order to provide probabilistic prediction, the following studies were performed using random forests (van Straaten et al., 2022) and neural networks (Miloshevich et al., 2023a). Probabilistic forecasting in the context of uncertainty quantification plays an important role in the current development of machine learning-driven techniques applied to, for instance, post-processing of weather forecasts (Grönquist et al., 2021; Schulz and Lerch, 2022).

Geopotential height anomalies and soil moisture were chosen as the inputs in Miloshevich et al. (2023a) based on physical understanding of the precursors to heat waves. Persistent positive

[1] An example of categorical score is Matthews correlation coefficient (MCC), which was designed as a measure of categorical classification and is based on a combination of true/false positive/negatives appropriate for imbalanced datasets.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

geopotential height anomalies are known drivers, since they favor clear skies and produce subsidence. It has been argued that soil moisture has memory of previous land-atmospheric conditions (Koster and Suarez, 2001; Seneviratne et al., 2006). In fact, soil moisture was identified (Seneviratne et al., 2012) among the important drivers for European heat waves. In contrast, in Northern European regions, soil moisture may play a lesser role in preconditioning (Felsche et al., 2023). This association between soil moisture deficits and heat wave occurrence has also been indicated elsewhere (Shukla and Mintz, 1982; Rowntree and Bolton, 1983; D'Andrea et al., 2006; Fischer et al., 2007; Vautard et al., 2007; Lorenz et al., 2010; Seneviratne et al., 2010; Hirschi et al., 2011; Stéfanon et al., 2012; Miralles et al., 2014, 2019; Schubert et al., 2014; Zhou et al., 2019; Vargas Zeppetello and Battisti, 2020; Benson and Dirmeyer, 2021; Zeppetello et al., 2022), contributing to the understanding that in dry regimes heat waves could be amplified due to the impacts of evapotranspiration. In contrast, the SWG, designed for heat waves, did not take this crucial input and thus was not able to reproduce the corresponding feedback (Jézéquel et al., 2018). However, it is important to better understand the contributions of such drivers to improve projected climate change impacts on heat wave probabilities (Field et al., 2012; Berg et al., 2015; Horton et al., 2016). For instance, using circulation analogs in a preindustrial simulation, Horowitz et al. (2022) showed that circulation patterns explain around 80% of the temperature anomalies in the United States, while negative soil moisture anomalies added a significant positive contribution, especially mid-continent.

Recently, the analog method was successfully adapted for predicting chaotic transitions in a low-dimensional system (Lucente et al., 2022a). In fact, the analog method is generally expected to perform better when the number of relevant degrees of freedom is not too high. This is natural given that it belongs to the family of k-nearest neighbors algorithms, which suffer from the curse of dimensionality (Beyer et al., 1999). Consequently, for our problem of heat wave prediction, we will consider linear and nonlinear dimensionality reduction techniques and evaluate the performance of SWG in real versus latent space. This approach is partially motivated by the emergence of generative modeling for climate and weather applications, for example, studies combining deep learning architectures with extreme value theory (EVT) for generating extremes (Bhatia et al., 2021; Boulaguiem et al., 2022) and realistic climate situations (Besombes et al., 2021).

Beyond finite-horizon probabilistic prediction, data-driven methods can be used to create emulators capable of extracting risks for "black-swan" events, events that are so far removed from the climatology that they have not been observed yet but their impact may be devastating. Such risk assessments are often carried out using EVT based on reanalysis, with long runs of high-fidelity models (methods such as ensemble boosting (Gessner et al., 2021) or rare event algorithms (Ragone et al., 2018; Ragone and Bouchet, 2021)). More recently, other statistical approaches such as Markov state models (Finkel et al., 2023) have been used to estimate the rates of extreme sudden stratospheric warming events based on ensemble hindcasts of European Center of Medium Range Weather Forecasting.

The first goal of this study is to compare the performance of two data-driven probabilistic forecasting approaches for extreme heat waves in two European regions: France and Scandinavia. We aim to explore the use of SWG, which relies on the method of analog Markov chain, optimize it for our task, and compare or combine it with more modern deep learning approaches such as CNN. We will investigate ways to accelerate the computation of analogs using dimensionality reduction techniques, such as training a variational autoencoder (VAE) to project the state of the system to a small-dimensional latent space. Our second goal will consist of generating synthetic time series with the help of the SWG trained on a short model run and comparing statistics of under-resolved extreme events (in this short run) to a long control run. We will work with PlaSim data, which is an intermediate complexity GCM. The choice of the right metrics is important since many of the current data-driven forecasts are trained based on mean square error, which is not suited for the evaluation of extremes. However, extremes actually represent the most immediate societal risks (Watson, 2022) and are a subject of this manuscript. The choice of GCM is motivated by the ability of PlaSim to generate inexpensive long runs and the recent study of Miloshevich et al. (2023b), who showed that the composite statistics of the large-scale 500 hPa geopotential height

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

fields, conditioned on heat waves, revealed similar patterns for PlaSim as for higher fidelity models such as CESM and the ERA5 reanalysis.

The paper is organized as follows. Section 2.1 describes the details of our dataset generated by PlaSim simulation. Section 2.2 defines the notation and heat wave events. Section 2.3 delineates the goal of the probabilistic inference, the scoring function that determines the goodness of the predictions, the training and validation protocol. Section 2.4 reviews the CNN introduced in Miloshevich et al. (2023a). Section 2.5 introduces the analog Markov chain, that is, SWG, and the corresponding steps involving coarse-graining, definition of metric, how to make probabilistic prediction with SWG, and how to construct return time plots. Section 3 spells out the results consisting of Section 3.1 that covers probabilistic forecasting and Section 3.2 discusses extending return time plots and teleconnections. Finally, Section 4 concludes the findings of this study.

2. Methods and data

2.1. PlaSim model data

The data have been generated from 80 batches[2] that are independent 100-year-long stationary simulations of PlaSim (Fraedrich et al., 2005, 1998). PlaSim is an intermediate complexity GCM, suitable for methodological developments, such as the one performed here. It can be run to generate long simulations at a lower computational cost than the new generation of CMIP6 ensemble.

PlaSim consists of an atmospheric component coupled to ice, ocean, and land modules. The atmospheric component solves the primitive equations of fluid dynamics adapted to the geometry of the Earth in the spectral space. Unresolved processes such as boundary heat fluxes, convection, clouds, and so forth are parameterized. The interactions between the soil and the atmosphere are crucial for this study. They are governed by the bucket model of Manabe (1969), which controls the soil water content by replenishing it from precipitation and snowmelt and depleting by the surface evaporation. The soil moisture capacity is prescribed based on geographical distribution.

The model was run with conditions corresponding to the Earth climate in 1990s with fixed greenhouse gas concentrations and boundary conditions, which include incoming solar radiation, sea surface temperatures, and sea ice cover that are cyclically repeated each year. The parameters are chosen to reproduce the climate of the 1990s. We also include the daily cycle in this work (like in Miloshevich et al. (2023a)). The model is run at T42 resolution, which corresponds to cells that are 2.8 by 2.8 degrees and results in 64 by 128 resolution over the whole globe. The vertical resolution is 10 layers. The fields are sampled every 3 hours and daily averages are taken.

In this paper, we work with different subsets of the full 8000-year-long dataset defined in Table 1.

Table 1. *Break-up of the input data into training and validation for different subsets of the full D8000 dataset. In D100 and D500, we follow fivefold cross-validation, while in D8000, exceptionally 10-fold cross-validation was employed. The latter implies that we slide the test window 10 times so that it explores the full dataset, while the training set is always the complement*

Abbreviation	Training years	L
*D*100	80	20
*D*500	400	100
*D*8000	7200	800

[2] In this context, "batch" refers to independent runs, that is, 80 independent initial conditions. In contrast, it does not refer to the procedure of splitting a dataset into portions that is performed during the gradient descent when training neural network.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

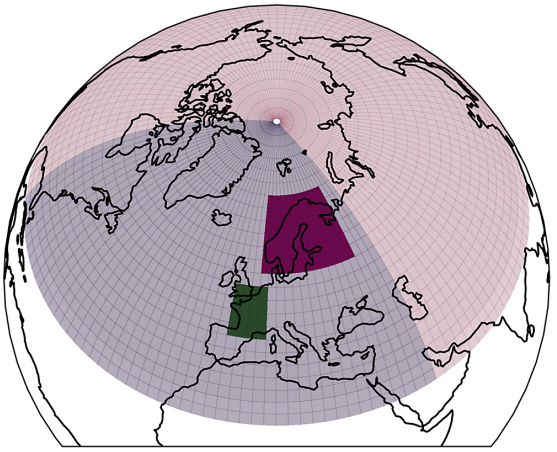


Figure 1. *The map shows relevant areas. Blue: North Atlantic–Europe (NAE). The dimensions of this region are 22 by 48. Red+Blue: North Hemisphere (above 30 N). The dimensions of this region are 24 by 128. Green: France; Purple: Scandinavia.*

2.2. Physical fields and geographical domains

The inputs to our various data pipelines will consist of fields $\mathcal{F}(\mathbf{r},t)$ such as

- $\mathcal{T}$: 2-meter temperature
- $\mathcal{Z}$: 500 hPa geopotential height
- $\mathcal{S}$: soil moisture

at time t, and discrete points $\mathbf{r} \in \mathbb{D}$, where $\mathbb{D}$ represents the domain of interest. We will consider three domains:

- $\mathbb{D}_{NAE}$: North Atlantic, Europe (24 by 48 cells);
- $\mathbb{D}_{France}$: the area of France masked over the land;
- $\mathbb{D}_{Scandinavia}$: the area of Scandinavia masked over the land.

The areas of France and Scandinavia that will correspond to the geographical areas of heat waves are defined in Figure 1. Data-driven methods that will be introduced below will accept typically some stacked version of the fields that will be represented by the letter $\mathcal{X}$

$$\mathcal{X} = (\mathcal{Z}(\mathbf{r},t), \mathcal{T}(\mathbf{r},t), \mathcal{S}(\mathbf{r},t)), \tag{1}$$

where each component corresponds to a specific field.

2.3. Definition of heat wave events

As in Miloshevich et al. (2023a), we are concerned with the prediction of $T = 15$ day-long heat waves. Thus, our events of interest are related to the time and space integrated anomaly of 2-meter temperature ($\mathcal{T}$), where we use symbol $\mathbb{E}$ to imply average over the years conditioned to the calendar day:

$$A(t) = \frac{1}{T}\int_t^{t+T} \frac{1}{|\mathbb{D}|}\int_{\mathbb{D}} (\mathcal{T} - \mathbb{E}(\mathcal{T}))(\boldsymbol{r},t')\,\mathrm{d}\boldsymbol{r}\,\mathrm{d}t', \tag{2}$$

which depending on the threshold $A(t) \geq \alpha$ defines a class of heat waves. Symbol $\mathbb{D}$ corresponds to the domain of interest described in Section 2.2. $|\mathbb{D}|$ is the surface area of $\mathbb{D}$ and equation (2) contains an area

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

weighted average. T is the chosen duration (in days) of the heat waves. For probabilistic prediction, we set $T = 15$ days, while for return times we relax this requirement to study a family of return time plots.

The paper will involve two slightly different definitions of heat waves based on equation (2) depending on the task that will be performed. In the next Section 2.4.1, we will introduce probabilistic prediction, for which heat waves will be defined as exceeding a certain fixed threshold a set fixed for all heat waves, which allows many heat waves per summer. Concurrently, in Section 2.6.6, we will work with a block-maximum definition, which permits only single heat wave per summer, so each year gets unique value a. Moreover, a range of T values will be considered.

2.4. Probabilistic prediction and validation

2.4.1. Committor function

Our goal is a probabilistic forecast, a *committor function* (Lucente et al., 2022a) in the nomenclature of rare event simulations. Based on equation (10), we define the label $y = y(t)$ for each state as a function of the corresponding time t as $y(t) \in \{0,1\}$, so that $y = 1$ iff $A > a$, where a is 95th percentile of A. Our objective is to find the probability p of $y = 1$ at time t given the state $\mathcal{X}$ which we observe at an earlier time $t - \tau$ (see Section 2.6.3):

$$p = \Pr(y(t) = 1 | \mathcal{X}(t-\tau), \tau). \tag{3}$$

Parameter τ is referred to as lead time or sometimes lag time. We stress that for nonzero τ it is the initial observed state $\mathcal{X}$ that is shifted in time, rather than the labels y, thus the events of interest (both $y = 1$ or $y = 0$) are kept the same, allowing controlled comparisons among different values of τ. Moreover, the season of interest for which we intend to compute the committor functions and compare them to the ground truth should always involve the same days: from June 1 to August 16, because the last potential heat wave may last $T = 15$ days and, this way, end in August 30 (the last day of summer in PlaSim).

2.4.2. Normalized logarithmic skill score

The quality of the prediction will be measured with a *normalized logarithmic score* (NLS) (Miloshevich et al., 2023a) due its attractive properties for rare events (Benedetti, 2010) and the appropriateness for probabilistic predictions (Wilks, 2019). In the following, we will briefly introduce the NLS. Since the dataset is composed of N states $\{\mathcal{X}_n\}_{1 \geq n \geq N}$, each of them labeled by $y_n \in \{0,1\}$, it can be equivalently represented by N independent pairs $(\mathcal{X}_n, y_n)$. The aim of the prediction task is to provide an estimate $\widehat{p}_y$ of the unknown probability $p_{\overline{y}}(X) = \mathbb{P}(y = \overline{y} \,|\, \mathcal{X} = X|)$. It has been argued (Benedetti, 2010; Miloshevich et al., 2023a) that among the possible scores to assess the quality of a probabilistic prediction, the most suitable is the logarithmic score

$$S_N(\widehat{p}_y) = -\frac{1}{N}\sum_{n=1}^{N} \log\left(\widehat{p}_{y_n}(\mathcal{X}_n)\right). \tag{4}$$

It should be noted that the logarithmic score coincides with the empirical cross entropy widely used in machine learning. The score S_N is positive and negatively oriented (the smaller the better). The value of S_N is not indicative in itself, but is useful for comparing the quality of two different predictions. Thus, the NLS is defined as an affine transformation of the logarithmic score, which allows us to compare the data-driven forecast with the climatological one. To be more precise, let $\overline{p}$ be the climatological frequency of the extremes ($\overline{p} = \mathbb{P}(y = 1)$) and $S_N^{(c)}$ the logarithmic score for this prediction, that is

$$S_N^{(c)}(\overline{p}) = -\frac{1}{N}\sum_{n=1}^{N}\left[\delta_{y_n,1}\log(\overline{p}) + \delta_{y_n,0}\log(1-\overline{p})\right] = -\overline{p}\log(\overline{p}) - (1-\overline{p})\log(1-\overline{p}). \tag{5}$$

Then, the NLS score is defined as

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

$$\mathrm{NLS}\left(\widehat{p}_y\right)=\frac{S_N^{(c)}-S_N\left(\widehat{p}_y\right)}{S_N^{(c)}(\overline{p})}. \tag{6}$$

Thus, NLS is positively oriented (the larger the better) and bounded above by 1. Since we identify $y=1$ with 95th percentile of $A(t)$ this implies that $\overline{p}=0.05$. In other words, we always compare the results to the baseline[3], which would result in NLS equal to zero. Any predictions with smaller NLS are completely useless from the probabilistic perspective. Finally, we stress that other scores devised for rare events, such as MCC (Matthews, 1975), most notably used in the study (Jacques-Dumas et al., 2022) and others, are useful for categorical prediction but do not necessarily translate into high NLS scores when training neural networks, especially if early stopping[4] is used, since the optimal epoch may vary depending on the chosen score.

2.4.3. Training/validation protocol for committor function

We split the dataset into training (TS) and validation (VS). The ML algorithm of choice is trained on TS and various hyperparameters are optimized on VS depending on the target metric (see Section 2.4.2). Different methods can be compared on VS, which will be performed in this paper. The splits will be based on Table 1 and correspond to the selection of the first number of years, for example, D100 implies choosing the first 100 years of the simulation.

To have a better idea about the spread of the skill (equation (6)), we will rely on fivefold cross-validation. This means that the TS/VS split is performed five times, while sliding the VS window consequently through the full data interval, chosen for the particular study (D100, D500, etc.). Each TS/VS split we refer to as "fold," that is, there are five folds (numbered 0–4). For instance, for D100, fold number 3, VS starts in year 60 and ends in year 79, while TS is as usual the complement of that set. Notice that years are also numbered from 0. In order to ensure that each VS has the same number of extreme events, we perform custom stratification (Miloshevich et al., 2023a). This means that we shuffle the years of the simulation between different folds so that each interval receives the same (or almost the same) number of extreme events. The procedure does not modify the mean benchmarks but tends to decrease the error bars, which strongly depend on the number of extreme events within the VS.

2.5. Convolutional neural network

The CNN we take for this study has been proposed by Miloshevich et al. (2023a). Since the precise inputs are slightly different in this paper (as well as a whole new region of Scandinavia that is considered in this paper for the first time), the training had to be repeated, but the hyperparameters were left unchanged.

In this manuscript, the CNN accepts an input that is 24 by 48 grid-points over three fields (see Eq. (1)). We perform masking (Miloshevich et al., 2023a) of the temperature and soil moisture, that is, values outside of the respective heat wave area (France or Scandinavia depending on which heat waves we study, see Figure 1) are set to zero to avoid some spurious correlations. In this paper, we also study the heat waves occurring in Scandinavia, and for consistency perform masking over the corresponding area in that case. The inputs are then normalized so that each grid-cell of each field has mean 0 and standard deviation 1.

The input is fed through three convolutional and two max-pooling layers; the intermediate output is flattened and passed through a dense layer, which is connected to the final output. The final output consists of two neurons and a soft-max activation, which enforces the outputs within $(0,1)$ interval consistent with the domain of probabilities. The architecture is trained with Adam optimizer to minimize cross-entropy as

[3] In this case, baseline would be always guessing the heat wave will occur with 5% probability independent of X.

[4] Early stopping refers to the callback that stops adapting the weights of the neural network once an objective criterion has been reached, which is often measured by monitoring the loss function.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

is appropriate for classification task and early stopping is applied, which stops the training when the average foldwise NLS reaches the maximum.

2.6. Analog Markov chain

2.6.1. SWG algorithm

The principle of the SWG as described in Yiou (2014) is based on the analog Markov chain method. The idea is to use the existing observations or model output that catalog the dynamical evolution of the system, *the dynamical history*, and generate synthetic series.

Algorithm 1: The algorithm generates a single synthetic trajectory by using the analogs of training set (see Table 1). Sets of K nearest analogues of each sample X_n during dynamical evolution are computed based on the metric given in equation (13). The states whose analogs we compute are constrained by the condition that their calendar days must start on May 15 and end on August 30 (there are 30 days per month in PlaSim). This results in $N_t = 105 \cdot N_{\text{years}}$. There is additional condition on the upper bound of the calendar day of the analogs of these days which is August 27 preventing SWG from accessing days outside of the summer season, since the time step is $\tau_m = 3$ days. The analogs are computed using parallelized kDTree search implemented in SciPy packages of python once per parameter α_0 (see equation (13) and discussion in Section 2.6.3) and each fold.

Define variables $(K, N_t, \mathcal{M}^{N_t \times K})$;
Input : $K \leftarrow$ the number of nearest neighbors retained
Input : $N_t \leftarrow$ the number of states in the training set
Input : Compute the matrix of nearest neighbors $\mathcal{M}^{N_t \times K}$
Input : $\mathcal{T} = \mathcal{T}(n) \leftarrow$ temperature sequence corresponding to historical states $n \in 1 \dots N_t$
Input : $s \leftarrow$ The initial state of the synthetic trajectory
Output: synthetic time series
append $\mathcal{T}(s)$ to the synthetic temperature series
while *synthetic trajectory continues* **do**
  $k \leftarrow \mathbb{U}(K)$ a uniform random number draw, i.e. $0 < k < K$
  $s \leftarrow \mathcal{M}^{N_t \times K}_{s,k}$ a random analog of the current state
  $s \leftarrow s + \tau_m$ a state is advanced according to its historical dynamical evolution.
  append s to the synthetic trajectory sequence;
  append $\mathcal{T}(s)$ to the synthetic temperature series
end
return synthetic temperature series

The method consists of constructing a *catalog of analogs*, the states that have similar circulation patterns and other thermodynamic characteristics (in our case temperature and soil moisture). Each state in the catalog comes from the realization of the dynamics (in principle, it could come not only from simulations but also reanalysis). To construct synthetic time series, one starts from any of these states and draws randomly (with uniform probability) one analog from a predefined set of nearest neighbors n. The similarity between the states is assessed via Euclidean metric in the feature space (Karlsson and Yakowitz, 1987; Davis, 2002; Yiou, 2014), as defined below (Section 2.6.3). The following step is to look-up in the dynamical history, that is, how that analog evolved dynamically after time τ_m and use the corresponding state as the next member of the synthetic time series (see Figure 2). Next, the process is repeated. The details of how the algorithm is implemented will become clear in the following sections but they are schematically represented in Figure 2 and encapsulated procedurally in Algorithm 1. The procedure is quite general although it is still possible to modify it in many aspects. For instance, the analogs are not

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

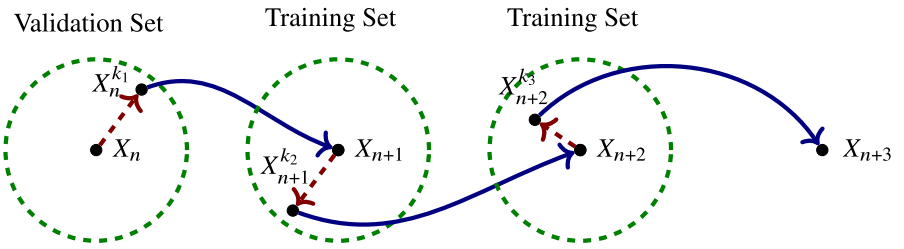


Figure 2. *The schematics of the flow of analogs. When applying the algorithm for estimation of committor (equation (3)), the first step consists of starting from the state in the validation set (VS), finding the analog in the training set (TS) and applying the time evolution operator. All subsequent transitions occur within the TS.*

always drawn from the uniform probability (Lall and Sharma, 1996; Rajagopalan and Lall, 1999; Buishand and Brandsma, 2001; Yiou, 2014), the catalog of analogs is often constrained by a *moving window* or the analogs from the same year may be excluded from possible candidates (Yiou, 2014). The common rationale behind our choices is to prioritize methods that demand less resources and use as much information as possible.

2.6.2. Time step and coarse-graining

The first question that arises is what is the appropriate choice for the time step τ_m of SWG (see schematics in Figure 2). It is natural to use synoptic time scale of cyclones on which the dynamics partially decorrelates, which corresponds to $\tau_m = 3$ days, as demonstrated by inspecting autocovariance Miloshevich et al. (2023b). This means that we are constructing a Markov chain where we are selecting a time step consistent with ignoring synoptic time-scale correlations. One may expect that the results do not depend too much on τ_m as long as it does not exceed the total correlation time (see Appendix B for further details). The coarse graining will be applied as follows, $\tilde{\mathscr{F}}(\mathbf{r}, t_0)$ will be

$$\tilde{\mathscr{F}}(\mathbf{r}, t) = \frac{1}{\tau_c}\int_t^{t+\tau_c} \mathrm{d}t_0 \mathscr{F}(\mathbf{r}, t_0). \tag{7}$$

The tilde will be suppressed without the loss of generality since the coarse-graining will be applied throughout, except in the input to the CNN. For simplicity, we perform coarse-graining in accordance with this time step $\tau_c = \tau_m$.

Note that we will also work with scalars obtained as area integrals $\mathscr{F}_{\mathbb{D}}$ over the area of France $\mathbb{D} = F$ and Scandinavia $\mathbb{D} = S$ defined as

$$\langle \mathscr{F} \rangle_{\mathbb{D}}(t) = \frac{1}{\mathbb{D}}\int_{\mathbb{D}} \mathrm{d}\mathbf{r} \mathscr{F}_{\mathbb{D}}(\mathbf{r}, t). \tag{8}$$

This allows us to rewrite the definition of heat waves (equation (2)) in a new, more concise notation

$$A(t) = \frac{1}{T}\int_t^{t+T} \mathrm{d}t' \langle \mathcal{T} - \mathbb{E}(\mathcal{T}) \rangle_{\mathbb{D}}(t'). \tag{9}$$

If we take coarse-graining into account and set $T = n\tau_c = 15$ days, we can express equation (2) as discretization

$$A(t_0) = \frac{1}{n}\sum_{i=1}^{n} \langle \mathcal{T} - \mathbb{E}(\mathcal{T}) \rangle_{\mathbb{D}}(t_0 + (i-1)\tau_c). \tag{10}$$

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

2.6.3. State space and metric

Next, we address how to combine fields of different nature (unit) in SWG. The fields can be stacked in various ways, for instance, tuples can be constructed:

$$\mathcal{X} = \left(\overline{\mathcal{Z}}(\mathbf{r},t), \langle \overline{\mathcal{T}} \rangle_{\mathbb{D}}(t), \langle \overline{\mathcal{S}} \rangle_{\mathbb{D}}(t)\right). \tag{11}$$

This approach adapted for SWG consisting of integrals over $\mathbb{D}$ is to be contrasted with the input received by CNN, whereby $\mathcal{S}$ and $\mathcal{T}$ are masked as described in Section 2.5 outside areas corresponding to the heat waves, while SWG receives only integrated temperature and soil moisture (over the same heat wave areas).

In the case of SWG, each field will be normalized to global field-wise standard deviation[5] (with the global field-wise mean subtracted) as opposed to cell-wise like in the case of CNN

$$\overline{\mathcal{F}}(\mathbf{r},t) := \frac{\mathcal{F}(\mathbf{r},t) - \mathbb{E}[\langle \mathcal{F} \rangle_{\mathbb{D}}(t)]}{\sqrt{\langle Var[\mathcal{F}] \rangle_{\mathbb{D}}(t)}}, \tag{12}$$

The overbar will be omitted in what follows and it will be implied instead.

The Euclidean distance between two points $\mathcal{X}_1$ and $\mathcal{X}_2$ is defined as

$$d(\mathcal{X}_1, \mathcal{X}_2) := \sqrt{\int d\mathbf{r} \sum_i \frac{\alpha_i}{dim(\mathcal{X}^i)} \left(\mathcal{X}_1^i - \mathcal{X}_2^i\right)^2}, \tag{13}$$

where $dim(\mathcal{X}^i)$ counts the number of grid points concerned (1 for scalars such as $\mathcal{S}$ and $\mathcal{T}$ and 1152 for $\mathcal{Z}$). When $\alpha_1 = \alpha_2 = 0$, we get definition consistent with (Yiou, 2014), except that we assign uniform probabilities for a set of nearest analogs.

Of a particular interest is the fact that geopotential contains global information on a shorter time scale, while soil moisture contains local information on a longer time scale (Miloshevich et al., 2023a). In the zeroth-order approximation, one expects that for the efficient algorithm, the fields would need to be further rescaled by their dimensionality as is done in equation (13). Thus, we choose parameters $\alpha_i = 1$ as a first guess. To optimize the performance of SWG for the conditional prediction problem (equation (3)), we perform grid search: for each τ we compute committor p and measure its skill (equation (6)) as a function of number of nearest neighbors n retained and the coupling coefficient for geopotential α_0, while the other two coefficients are set to one: $\alpha_{1,2} = 1$.

Other choices for distance function have been explored in the literature, such as Mahalanobis metric (Stephenson, 1997; Yates et al., 2003), which is equivalent to performing ZCA "whitening," that is, a procedure that is a rotation of a PCA components plus normalizing the covariance matrix. We do not explicitly compute Mahalanobis distance in this work, although we do actually compute the Euclidean distance after performing the dimensionality reduction via PCA as will be described below.

2.6.4. Dimensionality reduction

When dealing with high-dimensional fields, one often encounters the problem of "curse of dimensionality." In these cases, it is known that Euclidean distance is poorly indicative regarding the similarity of two data points. It could therefore be useful to project the dynamics onto a reduced space. Furthermore, in this way, the construction of a catalog is much more efficient from a computational point of view. Here, we decided to adopt two different dimensionality reduction techniques, namely a VAE and principal component analysis (PCA), also known as empirical orthogonal function (EOF) analysis. In a nutshell, PCA is a linear transformation of the original data, where the samples are projected onto the eigenvectors of the empirical correlation matrix (see Appendix A). Usually, the projections are made only on the eigenvectors corresponding to the largest eigenvalues, as they are the ones that contain most of the variability of the data. However, this variability may not be entirely

[5] The case of dimensionality reduction is addressed in the Appendix A, where normalization is not necessary.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

relevant to the prediction of heat waves, so this criterion is not necessarily useful. The VAE instead is a nonlinear probabilistic projections onto a latent space (usually with Gaussian measure). It consists of probabilistic *decoder* $p_\theta(x|z)$ and probabilistic (stochastic) *encoder* $q_\phi(z|x)$ and the training is performed with the aim of maximize the probability of the data $p(x)$ (Doersch, 2016) (more details can be found in Appendix A). Here, we adopt a deep fully CNN (FCNN) as an encoder followed by an upsampling FCNN decoder. Once the data have been projected into the latent space through one of the two methods, the SWG can be built exactly as explained above, simply replacing the fields in high-dimensional space with low-dimensional ones.

2.6.5. SWG committor

Due to the necessity of transitioning between TS and S (see Figure 2), we require two transition matrices $\mathscr{M}_{\text{tr}}$ and $\mathscr{M}_{\text{va}}$. The former is a straightforward application of what was discussed above, while the latter allows the trajectory starting in the VS to find the appropriate analog in TS. During this procedure (when $\tau = 0$), we retain the mean value of temperature over τ_c days that was already known in the VS (see Section 2.6.2), and reuse it when computing the synthetic label $A(t)$, equation (10). All the remaining entries in $A(t)$ are based on the temperatures corresponding to the subsequent states visited in the TS.

When $\tau_m = 3$ and $T = 15$ and $\tau = 0$ days, we need to evolve the trajectory only four times, thus each state s can have at most K^4 different values of A. Therefore, if we denote by $N_a(s)$ the number of trajectories such that A is greater than a, the committor function of the state s will be $p(s) = \frac{N_a(s)}{K^4}$. Because we are also interested in larger lead times such as $\tau = 15$ days we actually need to start the trajectories earlier, which leads to exponentially growing computational trees: with total nine steps of Markov chain necessary to be applied starting from each day in VS. To reduce the computational burden, we use Monte Carlo sampling with 10,000 trajectories at each day from VS rather than exploring the full tree. We do not observe improvements in the skill after refining with more trajectories.

Due to large number of trajectories, we had to parallelize our python code with a Numba package. SWG requires computation of a large matrix of nearest neighbors, for which we use kDTree from scikit-learn package and we run it in parallel on 16 CPU cores of Intel Xeon Gold 6142. Assuming there was no dimensionality reduction performed prior to this operation, the feature space resulting from $\mathcal{Z}$ field is 1152 dimensional (number of cells) and we run a grid search to find optimal α_0 coefficient in Eq. (2) for each of the five folds. The procedure takes approximately 18 hours when dimensionality reduction (see Appendix A) is not applied.

Finally, because of coarse-graining (see Section 2.6.2) and the definition of labels (equation (8)), special care has to be made when comparing quality of the prediction of (3) with respect to (Miloshevich et al., 2023a), where $\mathcal{X}$ was daily. Indeed, the coarse-grained field $\mathcal{X}$ retains information until τ_c days later due to forward coarse-graining (equation (7)). Thus, for a fair comparison between different coarse-grained committors, one must consistently shift the lead time τ of $\tau_{c_1} - \tau_{c_2}$ to ensure that fields $\mathcal{X}_1$ and $\mathcal{X}_2$ have no information about the future state of the system. In this paper, we consider coarse-graining time of 3 days.

2.6.6. Return time plots

The return time plots are produced using the method described in details in Lestang et al. (2018). Here, we will mostly summarize the algorithm. The idea is to estimate the return time of block-maximum (summer) $A(t)$ surpassing a particularly high threshold value a. Thus, the heat wave definition differs slightly with respect to what is used in the committor definition (Section 2.4.1), that is, we do not restrict a to the 95th percentile, but instead identify the largest value of $A(t)$ each summer which will correspond to a of that year. The interval (duration of the block) is defined so that heat wave may start in June 1 and end in August 30, so the interval depends on the duration of the event T, which for this application may take values other than just 15 days. If $T = 30$ days, for instance, the last heat wave may start no later than August 1. Next, the yearly thresholds a's are sorted in the descending order from largest to smallest: $\{a_m\}_{1 \leq m \leq M}$, where M is the total number of years. With this definition, the most extreme return time, is identical to the length of the

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

dataset (in years) under consideration (Table 1) and the thresholds are ordered $a_1 \geq a_2 \geq \ldots \geq a_M$. The simplest approach is to just to associate the rank of the thresholds a to the inverse return time.

$$r = \frac{M}{\text{rank}(a)}, \tag{14}$$

This definition is intuitively reasonable for large a but runs into problems at the other end of the a axis.

A more precise estimator uses the assumption of Poisson process $P(t) = \lambda \exp(-\lambda t)$ approximation (which largely affects the small return times), *modified block maximum estimator* in Lestang et al. (2018) which reads in our notation as

$$r = -\frac{1}{\log(1 - \text{rank}(a)/M)}, \tag{15}$$

where M is the length of the dataset in years and $\text{rank}(a)$ is the order in which the particular year appears in the ordered list of years (by threshold).

2.7. Generalized extreme value fit

We chose the scikit-extremes package of Correoso (2019) to perform the generalized extreme value (GEV) fits. The maximal yearly summer $A(t)$ is calculated and then fit using maximum likelihood method estimator. Because the initial guess for the parameters is important and can have an impact on the shape parameter of the GEV distribution, linear moments of the distribution are used to obtain those start values. The GEV function is as follows

$$f(x,c) = \exp\left(-(1-cx)^{1/c}\right)(1-cx)^{1/c-1}, \tag{16}$$

where c corresponds to the shape parameter. To estimate the confidence intervals, the delta method is employed. Some years in the simulation tend to be cooler and even with negative temperature anomaly $A_{\max}(T) < 0$. In fact, there is a handful of such $A_{\max}(T) < 0$ years in the TS of D100. To better estimate the GEV parameters, we remove such years.

3. Results

3.1. Probabilistic forecasting

The first goal is to compare SWG and CNN as tools for probabilistic heat wave forecasting. This is done in a framework depicted schematically in Figure 3.

3.1.1. Comparisons with CNN

Here, we plot the NLS (equation (6)) comparing the predictions of CNN (orange curve) versus SWG (blue curve) described in Sections 2.5 and 2.6 for France (top panels) and Scandinavia (bottom panels), respectively, in Figure 4. The comparison is made on D500 (see Table 1) but similar results are obtained considering shorter dataset (see Appendix C). The left panels are plotted as a function of lead time τ and just like in Miloshevich et al. (2023a), we see a downward trend in NLS due to chaotic nature of the atmosphere (the predictability diminishes with time). As was described in Section 2.6.3, parameters α_0 and the number of nearest neighbors of SWG were tuned based on cross-validation which allows us to provide error bars (shading around the curves). We find that using 10 nearest neighbors is almost always optimal (right panels), whereas the optimal values of α_0 (middle panels) tend to be on the order of 50–100. Later, we shall see that α_0 also may depend on the preprocessing the data (such as dimensionality reduction).

We are following the identical procedures of data processing for both SWG and CNN (Section 2.4.3). However, because of the coarse-graining in SWG, the comparisons are only appropriate, if we shift orange curve by $\tau_c - 1 = 2$ days ahead (increasing τ). The latter adjustment is explained in Section 2.6.5. Figure 4

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

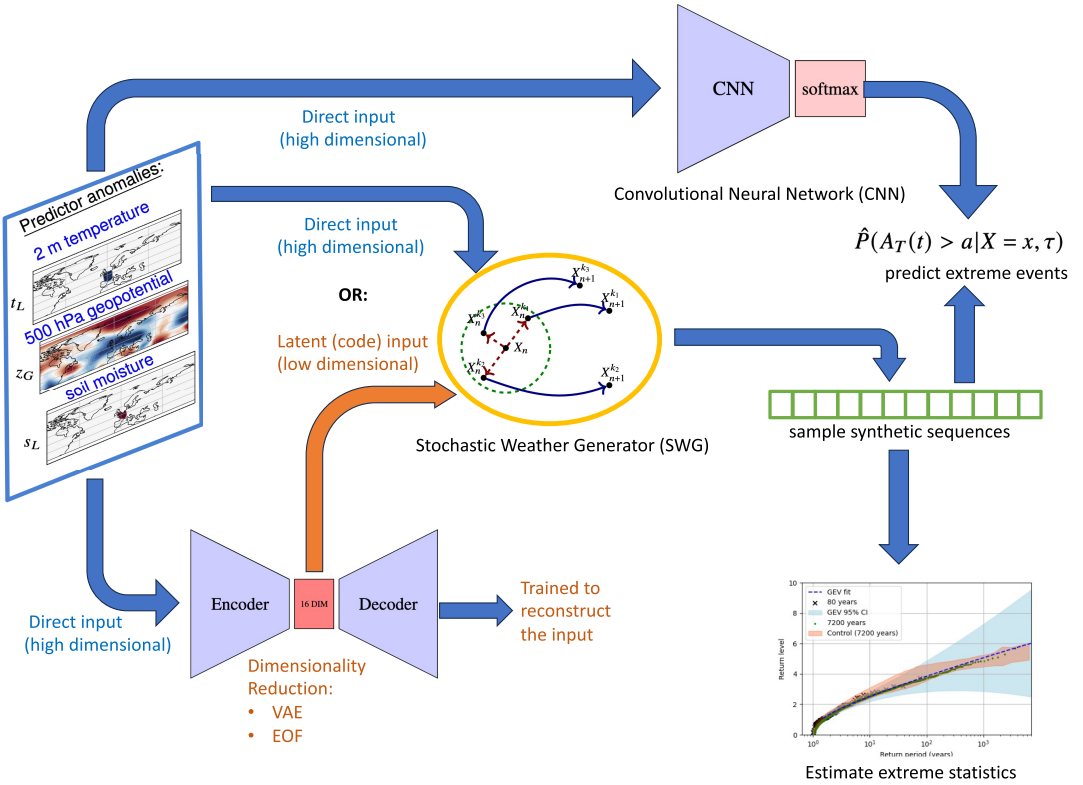


Figure 3. *Schematics of different methodologies used for probabilistic forecasting and estimates of extreme statistics. On the left, we show the three input fields which are labeled directly on the plot. On the right, the target of the inference is displayed (for instance, probabilistic prediction as described in Section 2.6.5). On top, we have the direct CNN approach. In the middle, we show the analog method (SWG), which is the main topic of this work and which is compared to CNN for this task. At the bottom, the option is presented to perform dimensionality reduction of the input fields and pass them as the input on which SWG is "trained." Subsequently, SWG can be used to generate conditional or unconditional synthetic series (green boxes). This synthetic data can be used for make probabilistic prediction or estimating tails of distribution, such as return time plots.*

shows that in France at $\tau = 2$ days NLS reaches approximately 0.40 ± 0.04 for CNN and an estimation of 0.30 ± 0.02 for SWG. In Scandinavia (bottom panel), we have similar results: approximately 0.37 ± 0.03 for CNN versus only 0.25 ± 0.02 for SWG. Large lead times τ result in lower predictive skill due to chaotic nature of the atmosphere. The behavior of the curve leads to the conclusion that CNN predicts heat waves better up to 10 days before the event, at which point the CNN and SWG skills are not statistically distinguishable. In case of France, the regime of large τ is mostly dominated by predictability due to soil moisture (Miloshevich et al., 2023a), whereas in Scandinavia, this long-term information is absent; thus, NLS converges to near-zero values at large τ. It is generally known (Felsche et al., 2023), that it is heat waves in Southern Europe that are preceded by the negative anomaly of soil moisture, rather than Northern Europe, which is consistent with what we see here. The reasoning we provide is that in France, the soil may actually be in a regime of strong lack of humidity in summer, which significantly amplifies the soil–atmosphere feedbacks.

Looking at the optimality of the parameter α_0, which controls the importance of geopotential in the metric (equation (13)), we see that for France, it is of order $\alpha_0 \sim 50$, an order of magnitude higher than what we expected $\alpha_0 = 1$ based on weighting by the number of relevant grid points. In Scandinavia, the

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

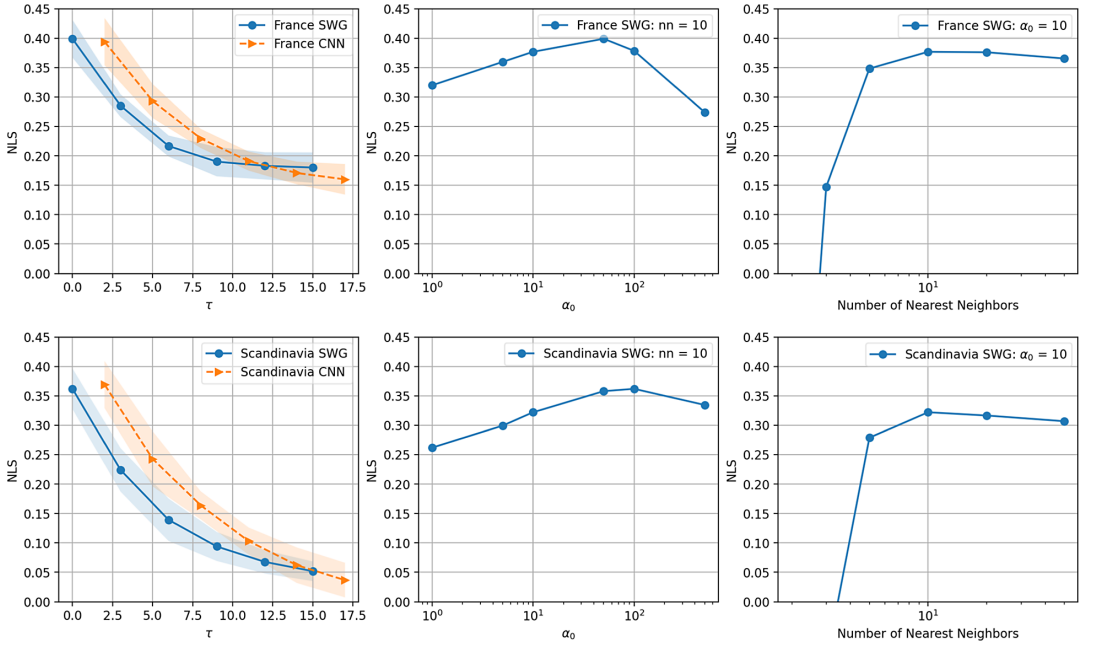


Figure 4. **Basic stochastic weather generator (SWG) (blue curve) versus convolutional neural network (CNN). (Orange curve)** *All three panels display normalized logarithmic score (NLS) (equation (6)) on the y axis. Left panel has τ on the x axis, central panel has α_0 (a hyperparameter of SWG, see equation (13)) on the x axis and right panel has n-nearest neighbors (also hyperparameter of SWG) on the x axis. On the central and the right panels, the choice for $\tau = 0$ was made. The dots show data points corresponding to the mean of the cross-validation, whereas the thickness of the shaded area represents two standard deviations. These conventions will be reused in the subsequent figures.*

optimal value is not too different, of order $\alpha_0 \sim 100$, but the dependence is weaker. The reason for this could also be the fact that soil moisture does not contribute much to heat wave predictability in that region, thus limiting its usefulness in the metric (equation (13)).

The benchmarks indicate that CNN outperforms SWG independent of the area of interest (France or Scandinavia) or even dataset length (Figure A3).

3.1.2. Dimensionality reduction

We now move on to the question of projecting the underlying space to latent dimension as described in Section 2.6.4. Three approaches are considered, one which does not involve any dimensionality reduction, one where a deep FCNN is used as an encoder followed by an upsampling FCNN decoder and one, where instead of autoencoder we use SciPy package PCA projections[6]. We plot the results in Figures 5 and 6 for France and Scandinavia, respectively. There, simple SWG trained without dimensionality reduction is represented via a blue curve, autoencoder (latent dimension $p = 16$) via orange and the other curves represent PCA for different latent dimensions p. For France, we see no statistically significant differences even in the drastic case where the latent dimension consists of the two largest principal components only[7]. This is probably due to the strong predictability power of soil moisture field (as clearly shown in Miloshevich et al. (2023a)). Indeed, the results for Scandinavia (which is less affected

[6] PCA is known as EOF analysis in the field of geosciences.

[7] Note that we speak of the two largest principal components of the geopotential, but SWG also accepts local temperature and soil moisture as input predictor variables.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

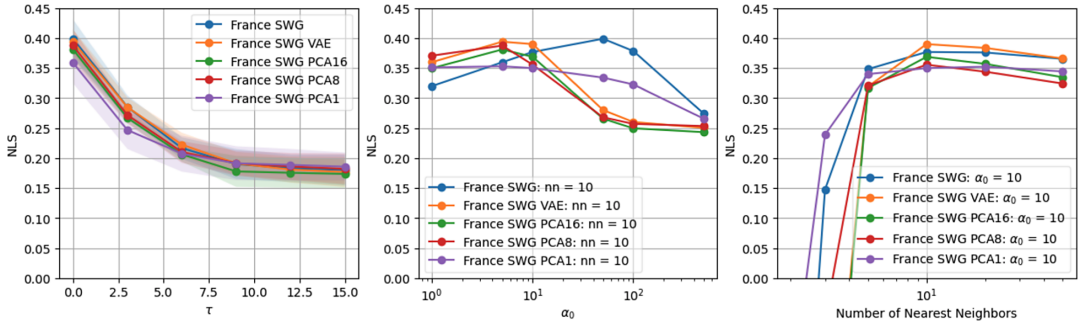


Figure 5. **Basic stochastic weather generator (SWG) versus VAESWG***: On the y axis, we have normalized logarithmic score (NLS) (equation (6)) as a function of lead time τ and hyperparameters of SWG (see the caption of Figure 4). SWG is indicated by the same (identical) blue curve as in Figure 4 while orange and green curves correspond to VAESWG where geopotential was passed through two different autoencoders (equation (A2)) (orange and green curves).*

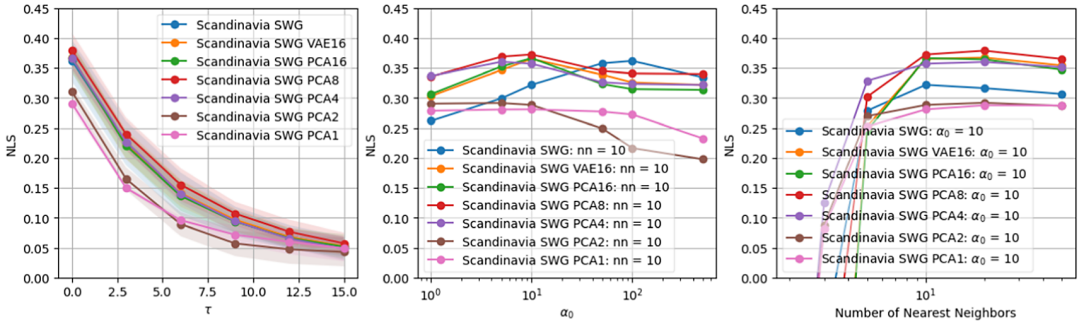


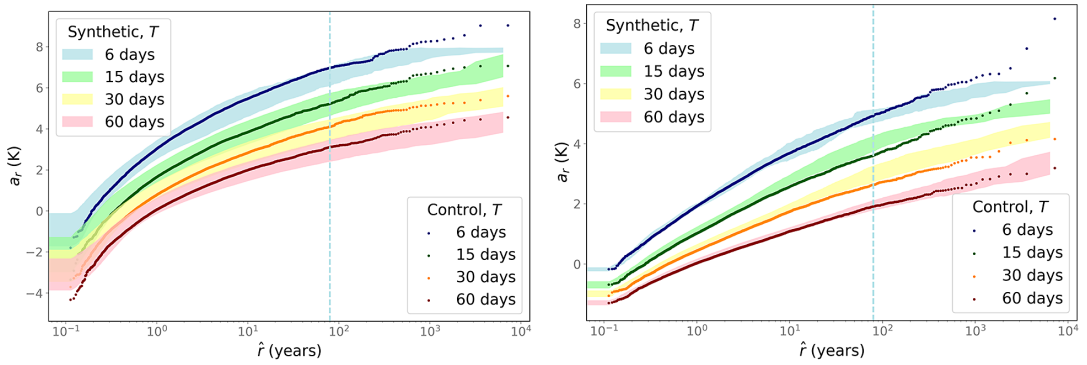
Figure 6. **Basic stochastic weather generator (SWG) versus VAESWG:** *On the y axis, we have normalized logarithmic score (NLS) (equation (6)) as a function of lead time τ and hyperparameters of SWG (see the caption of Figure 4). SWG is indicated by the same (identical) blue curve as in Figure 4 while orange and green curves correspond to VAESWG where geopotential was passed through two different autoencoders (equation (A2)) (orange and green curves).*

by soil moisture) shown in Figure 6 display a degradation when $p < 4$ while for $p \geq 4$ all methods provide essentially the same score. Interestingly, the highest score (although within the error bars) is obtained for $p = 8$, suggesting that an optimal dimension may exist.

The computation of analogs (equation (13)) is rather difficult in high dimensions and with datasets larger than years becomes unfeasible, because we also have to perform tuning of α_0 and the number of nearest neighbors. This is much more efficient if the dimension of the space has been successfully reduced, with a method such as autoencoder. However, as can be seen from Figures 5 and 6, the benefits from using complicated multilayer architecture are limited in this case, given the fact that the green curve, corresponding to PCA is almost the same (within the error bars) as the one obtained with the autoencoder. One notable difference between the cases with dimensional reduction and without is that the value for α_0 that is optimal $\alpha_0 = 5$ is an order of magnitude smaller than the one we obtain in Figure 4, corresponding to SWG learned in real space.

These conclusions should be taken in the context of the predictability of temperature using geopotential, which have rather linear Gaussian statistics. It could be that autoencoders are more useful for more complex fields such as precipitation.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press



(a) France return times

(b) Scandinavia return times

Figure 7. *Return time plot for (a) France and (b) Scandinavia heat waves using analogues of North Atlantic and Europe and the method based on equation (15). Here, we use parameters $\alpha = 1$ (default), Number of nearest neighbors $n = 10$, the analogs are initialized on June 1 of each year (using the simulation data) and then advanced according to Algorithm 1. The trajectory ends at the last day of summer. Each trajectory is sampled 800 times providing much longer synthetic series and thus estimating longer return times. Return times are computed for $T = \{6,15,30,60\}$ day heat waves (indicated on the inset legend), with dots corresponding to the statistics from the control run (D8000, see Table 1), while shaded areas correspond to the bootstrapped synthetic trajectory: the whole sequence is split into 10 portions which allows estimating the mean and variance. The shading corresponds to mean plus or minus one standard deviation.*

3.2. Sampling extremes

From now on, our goal will be to use SWG and evaluate the quality of the statistics for the extremes it can produce. In contrast to Section 3.1, we will consider heat wave extremes as large as one in 7000 years events.

3.2.1. Computing return times

We aim to estimate extreme return times from shorter sequences and compare them to the control run D8000 (Table 1). To this end, we use SWG trained on D100 (Table 1) and initialized with $\alpha_0 = 1$ and 10 nearest neighbors, parameters that we do not optimize in contrast to Section 3.1. The generated synthetic sequences are plotted for France in Figure 7a and for Scandinavia in Figure 7b. The curves were obtained using the method described in Section 2.6.6, so that the analogs are initialized in June 1 of each year (using the simulation data) and then advanced according to Algorithm 1. The trajectory lasts up to synthetic August 30.

The main conclusion from Figure 7a is that the analog method is rather well-suited for this task: not only do synthetic trajectories match return times of the reference real trajectory (80 years of analogs) but they also provide correct estimates of the returns of a much longer trajectory (8000 years) at a fairly low computational footprint, compared to running such a trajectory. For instance, for $T = 15$ day heat waves, the most extreme event with return time of 7200 years has anomaly $7.07K$ which is well within the error bars of SWG estimate $7.02 \pm 0.46K$. We can see that the generalization happens consistently, except for the extremes of 6-day heat waves, where we have sampling issue since $\tau_m = 3$ days. The procedure is repeated for Scandinavia in Figure 7b. The results are generally consistent, although we observe more deviations from the prediction of SWG. For example, heat waves of length $T = 15$ days and return times of order 2000–3000 years as well as the longest return times for $T = 30$ days tend to be systematically underestimated. Note, however, that these events are rare even in the control run and therefore their thresholds have large uncertainties. Excluding these

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

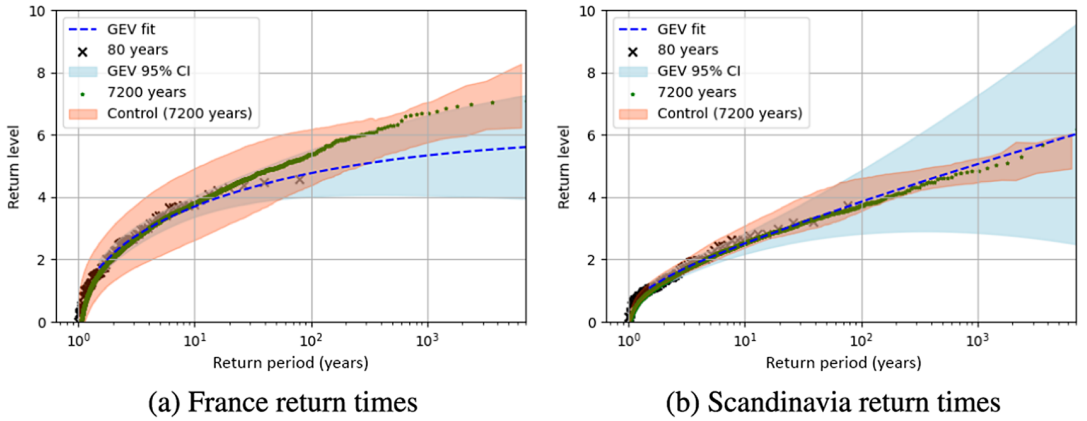


Figure 8. *Return time plot for (a) France and (b) Scandinavia $T = 15$ day heat waves using the method based on equation (14). On the y axis, a anomalies are plotted in K and on the x axis years. The black dots correspond to the most extreme heat waves in 80 years of D100. The blue dashed line corresponds to the GEV fit performed on 80 years of D100 (minus the ones which have negative A(t) maxima). The 95 percent confidence intervals are indicated via blue shade. The synthetic time series generated via SWG and identical to the green shade in Figure 7 are plotted using orange shade, except that we chose to shade two standard deviations, rather than one for consistency. The 7200 years control run (identical to green dots in Figure 7) are plotted using green dots.*

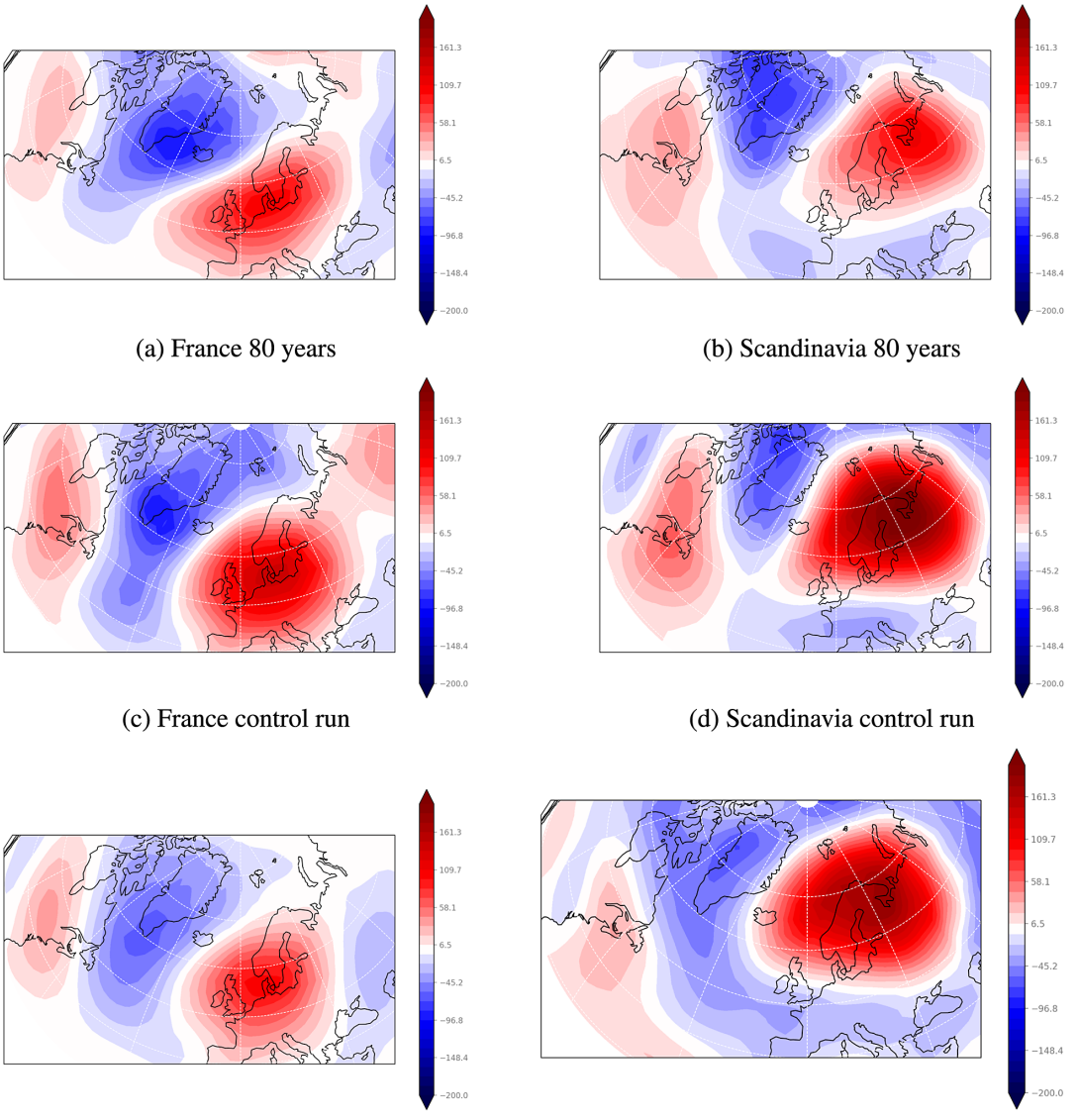
extreme events, the predictions are almost always in agreement with the return times calculated on the long run and when they are not, the relative error is smaller than 15%.

Since for this type of risk assessments EVT is often employed, we compare our approach to the GEV function fits obtained using package developed by Correoso (2019) from the 79 extreme $T = 14$ day heat waves in the 80 years of D100 (our training data) in Figure 8. For details on why only 79 heat waves are chosen, see Section 2.7.

Generally speaking, GEV fit performs worse than SWG; in France, it underestimates the severity of extremes, while in Scandinavia, it produces confidence intervals that are very large. While the GEV fit is also consistent with the extremes, we observe in the control run it provides much looser confidence intervals. On the other hand, SWG tends to shadow more closely the extremes of the long control simulation. This could be ascribed to the hypothesis that there is an upper bound for temperature extremes (e.g., Zhang and Boos (2023)). Parameters of the GEV fit yield uncertainties (especially the shape parameter) which can be very large when the available data are limited (here, 80 years). This leads to wide uncertainties for return periods that are larger than the training length, although in case of Scandinavia the return levels of the long control simulations do fall within the confidence intervals of the GEV fits, albeit not *on* the GEV fit. On the other hand, the SWG simulations are close to intrinsic properties of temperature, driven by the predictors we use. This explains why the SWG simulations follow the long control run, with a relatively narrow confidence interval, which does not increase with return periods.

We do not control for α_1 and α_2 to give us insight into which variable matters more in this aspect (i.e., temperature or soil moisture); however, since soil moisture does not have impact on heat waves on Scandinavia, one could reason that it is temperature in that case. In other words, to produce reliable return time plots the Euclidean metric should not only take circulation patterns into account but also the temperature and soil moisture in some cases. A side criticism of our approach is that we have chosen value $\alpha_0 = 1$, which seems arbitrary in view of the discussion in Section 3.1.1 on the optimality of $\alpha_0 \sim 50$ for performing the probabilistic forecast. However, it should be noted that having changed the task to be performed, there is no longer a guarantee that $\alpha_0 \sim 50$ is still an optimal value (and it is not, as shown in Appendix D).

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press



(a) France 80 years (b) Scandinavia 80 years

(c) France control run (d) Scandinavia control run

(e) France synthetic SWG (f) Scandinavia synthetic SWG

Figure 9. *Composite maps of geopotential height (meters) anomalies at 500 hPa for heat wave in France and Scandinavia ($T = 15$ days). (a) Forward 3 day running mean at $\tau = 0$, that is, the heat wave onset, a composite of the 10 most extreme France heat waves in an 80-year-long dataset with the threshold ≈ 4 K. (b) Same as (a) but for Scandinavia heat waves. (c) Composite for the control run with a collection of France heat waves above the threshold 4 K. (d) Same as (c) but for Scandinavia heat wave. (e) Composite for the synthetic run performed by stochastic weather generator (SWG) above the same threshold for France heat waves. (f) Same as (d) but for Scandinavia heat wave.*

3.2.2. Teleconnection patterns for extremes

We use synthetic trajectories generated by SWG trained on only 80 years (from D100, see Table 1) to estimate extreme teleconnection patterns that were never observed in that run. Note that we do not use autoencoder to generate new circulation patterns. Instead, SWG generates new sequences from the already existing ones which results in new values of $A(t)$ (equation (2)) since it is a running mean over

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

temperature series. This is precisely the recipe we have followed in Section 3.2.1. To validate the teleconnection patterns thus obtained, we compare them to the control run (D8000). The results are plotted in Figure 9.

First, we identify the 10 most extreme heat waves in D100 and plot the composite on their first3 days ($\tau = \{0,1,2\}$ days) since the onset in Figure 9a[8]. The threshold for this event happens to be approximately $4K$. The pattern consists of anticyclonic (in red) and cyclonic (blue) anomalies (the word "anomalies" will be suppressed in what follows for brevity).

Next, we compare this lack of data regime (essentially 10 events) to the composite map that can be computed using a 7200 years long control run in Figure 9c using the same $4K$ threshold and plotting the first 3 days ($\tau = \{0,1,2\}$ days) of the composites. In contrast to the previous situation, this leaves us with many more events satisfying the constraint. The picture changes somewhat, most notably, the cyclone and anticyclone become more pronounced, while the other patterns maintain relationship consistent with a Rossby wave pattern. This is in line with the understanding of teleconnection patterns expected in European heat waves. This pattern is also consistent with wave number 3 teleconnection for heat waves in France obtained in Ragone and Bouchet (2021) and Miloshevich et al. (2023b) computed for 1000-year-long sequence.

Finally, we compare both figures with a very long synthetic SWG run in Figure 9e. The resulting teleconnection pattern is again similar to the one in Figure 9e. Overall, the advantage offered by SWG in this case is rather small and mostly has to do with the shape of the anticyclone over Northern Europe which in the short D100 composites appears more narrow.

We repeat the exact same steps but for the heat waves in Scandinavia, starting from the 10 most extreme events in D100 in Figure 9b to the composites resulting from synthetic SWG run in Figure 9f. In this case, SWG performs better in that it is able to capture the magnitude of the anticyclone over Scandinavia more accurately (compared to Figure 9d) than D100 composite, although the other features are hard to distinguish.

4. Discussion

We have systematically compared performance of SWG and CNN (Miloshevich et al., 2023a) tasked to predict the occurrence of prolonged heat waves over two European areas using simulation data from intermediate complexity climate model PlaSim. We have also studied the ability to sample extreme return times and composite maps via the method of SWG.

In addition to the study of heat waves in France considered in Miloshevich et al. (2023a), we also apply our methodology to heat waves in Scandinavia. This area has different hydrology which impacts long-term predictability. In order to achieve better predictive ability with SWG, we took the version developed earlier (Yiou, 2014), which previously used only analogs of circulation, predictors we refer to as *global*, and upgraded it with additional predictors such as temperature and soil moisture integrated over the area of interest (where heat waves occur), which we refer to as *local.* This was done because of the ample evidence of importance of slow drivers such as soil moisture (Zeppetello et al., 2022). Consequently, we had to use appropriate weights in the definition of Euclidean distance to control the importance of global versus local predictors. The benchmarks of CNN versus SWG were obtained based on NLS which is particularly well suited for studying rare events (Benedetti, 2010). The conclusion is that CNN outperforms SWG in probabilistic forecasting, although this is more evident for particularly large (hundreds of years long) Training Sets (TSs), which would not be available in the observational record. This is in line with the statement of Miloshevich et al. (2023a) on the convergence of such methods in deep learning that requires massive datasets. These conclusions could be interesting for extreme event prediction but are also relevant for future developments of geneological rare event algorithms that aim to resample heat waves based on the forecasts provided by either CNN or SWG (in the approach similar to Lucente et al. (2022b)).

We have studied the performance of SWG in relation to the estimation of large return times and extreme teleconnections. This is achieved by generating synthetic trajectories of extensive length based on the

[8] This is in-line with $\tau_m = 3$ day coarse-graining approach, we have taken for SWG inputs as explained in Section 2.6.2.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

initial training data of a short 80-year PlaSim run. We have shown that out-of-the-box (without hyperparameter optimization) implementation of SWG, which takes the three relevant fields as input for the Euclidean metric, was capable of reproducing the return times of the much longer control run (7200 years) for a range of heat wave durations: 15, 30, and 90 days ones. These SWG calculations shadow more closely the control run (with smaller variance) than the EVT estimates, which are typically used to address similar questions. However, the version of SWG whose weights have been optimized for a different task of conditional (intermediate range to subseasonal) probabilistic prediction tends to underestimate the return times slightly. Nevertheless, its synthetic teleconnection patterns have qualitatively accurate features. This is an important result in view of works such as (Yiou et al., 2023) that rely on SWG to estimate the risks of the extremes.

In the hopes of improving the performance of SWG, we have considered two dimensionality reduction techniques to project the geopotential to the smaller latent space. These techniques involve traditional EOFs and VAEs. We found that while improving the efficiency at which the analogs are computed, the dimensionality reductions did not modify the predictive skill, which leads us to suggest EOF as a superior dimensionality technique for this task. However, it is possible that VAE could still be useful for other types of extremes, such as precipitation, where the application of traditional analog method is more controversial. We leave such studies for the future. When looking for optimal dimensionality of the latent space, we found that a few 500 hPa geopotential EOF components coupled with local temperature and soil moisture are sufficient to obtain optimal forecasting skill for SWG, which, however, as stated earlier, falls short of CNN prediction, which takes the full fields as in input.

Another possibility of extending this work would be comparing SWG to other modern tools for learning propagators, for instance, simple U-Net architectures or a corresponding Bayesian neural network (Thuerey et al., 2020, 2021), which provides uncertainty of the prediction thus a kind of propagator. Many other types of architectures are possible, however, given overall simplicity of SWG when training climate model emulators with small datasets SWG should be treated as an indispensable baseline method to justify the use of complex architectures. Fortunately, we have supported this paper with all the necessary code and documentation for straightforward implementation of SWG that can be applied to other projects.

As stated above, another interesting potential application for SWGs and other statistical methods such as deep learning is rare event algorithms (Lucente et al., 2022a). Rare event algorithms have been used in the past to sample heat waves (Ragone et al., 2018; Ragone and Bouchet, 2021) and could potentially allow to sample other extreme events (Webber et al., 2019) with expensive high-fidelity models at a cheaper cost. Data-driven approaches could be used to improve such rare event simulations. This is because knowing the probability of an event gives a prior information to the rare event algorithm (Chraibi et al., 2021). A recent work (Jacques-Dumas et al., 2023) compares SWG and CNN in computing importance functions for two simple Atlantic meridional overturning circulation models. We believe that providing similar benchmarks is important; thus, in this manuscript, we concentrate on the comparisons between deep learning and analog forecasting.

Finally, the results here were obtained based on dataset generated by intermediate complexity climate model, PlaSim with the ocean that was driven using stationary climatology and with relatively coarse resolution. As other modes of long-term predictability come from the ocean the next study should investigate the same questions based on higher fidelity model with better parameterizations. In particular, the question of transfer learning should be addressed, that is, how to properly generalize out-of-sample by pretraining on a simpler models and fine tuning on more complex datasets, such as CESM or reanalysis.

Acknowledgments. The authors acknowledge CBP IT test platform (ENS de Lyon, France) for ML facilities and GPU devices, operating the SIDUS solution (Quemener, 2014). This work was granted access to the HPC resources of CINES under the DARI allocations A0050110575, A0070110575, A0090110575, and A0110110575 made by GENCI. The authors would like to acknowledge the help of Alessandro Lovo in maintaining the GitHub page. The authors would also like to acknowledge the help of the referees in streamlining and improving the quality of this article.

Author contribution. Conceptualization: D.L., F.B., P.Y., G.M.; Data curation: D.L., G.M.; Formal analysis: D.L., G.M.; Investigation: D.L., P.Y., G.M.; Methodology: D.L., F.B., G.M.; Software: D.L., G.M.; Writing – original draft: D.L., F.B., P.Y.,

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

G.M.; Writing – review & editing: D.L., F.B., P.Y., G.M.; Funding acquisition: F.B., P.Y.; Project administration: F.B., P.Y., G.M.; Supervision: F.B., P.Y., G.M.; Validation: G.M.; Visualization: G.M.

Competing interest. The authors report no competing interests.

Data availability statement. Data for this study are available from https://zenodo.org/records/10102506.

Funding statement. This work was supported by the ANR grant SAMPRACE, project ANR-20-CE01-0008-01 and the European Union's Horizon 2020 research and innovation program under grant agreement No. 101003469 (XAIDA). This work has received funding through the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

Code availability statement. The coding resources for this work, such as the Python and Jupyter notebook files, are available on a GitHub page https://github.com/georgemilosh/Climate-Learning and are part of a larger project at LSCE/IPSL ENS de Lyon with multiple collaborators working on rare event algorithms.

Abbreviations

CESM	Community Earth System Model
CNN	Convolutional Neural Network
EOF	Empirical Orthogonal Function
EVT	Extreme Value Theory
GCM	General Circulation Model
GEV	Generalized Extreme Value
MCC	Matthew's Correlation Coefficient
NLS	Normalized Logarithmic Score
NWP	Numerical Weather Prediction
PCA	Principal Component Analysis
PlaSim	Planet Simulator
S2S	Subseasonal-to-Seasonal
SWG	Stochastic Weather Generator
TS	Training Set
VAE	Variational Autoencoder
VS	Validation Set

Balaji V (2021) Climbing down charney's ladder: Machine learning and the post-dennard era of computational climate science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379*(2194), 20200085.

Barriopedro D, **Fischer EM**, **Luterbacher J**, **Trigo RM and García-Herrera R** (2011) The hot summer of 2010: Redrawing the temperature record map of europe. *Science 332*(6026), 220–224.

Benedetti R (2010) Scoring rules for forecast verification. *Monthly Weather Review 138*(1), 203–211.

Benson DO and Dirmeyer PA (2021) Characterizing the relationship between temperature and soil moisture extremes and their role in the exacerbation of heat waves over the contiguous United States. *Journal of Climate 34*(6), 2175–2187.

Berg A, **Lintner BR**, **Findell K**, **Seneviratne SI**, **van den Hurk B**, **Ducharne A**, **Chéruy F**, **Hagemann S**, **Lawrence DM**, **Malyshev S**, **Meier A and Gentine P** (2015) Interannual coupling between summertime surface temperature and precipitation over land: Processes and implications for climate change. *Journal of Climate 28*(3), 1308–1328.

Besombes C, **Pannekoucke O**, **Lapeyre C**, **Sanderson B and Thual O** (2021) Producing realistic climate data with generative adversarial networks. *Nonlinear Processes in Geophysics 28*(3), 347–370.

Beyer K, **Goldstein J**, **Ramakrishnan R and Shaft U** (1999) When is "nearest neighbor" meaningful? In Beeri C and Buneman P (eds.), *International Conference on Database Theory*. Berlin, Heidelberg: Springer, pp. 217–235.

Bhatia S, **Jain A and Hooi B** (2021) Exgan: Adversarial generation of extreme samples. *Proceedings of the AAAI Conference on Artificial Intelligence 35*(8), 6750–6758.

Bjerknes V (1921) The meteorology of the temperate zone and the general atmospheric circulation. *Monthly Weather Review 49*(1), 1–3.

Boulaguiem Y, **Zscheischler J**, **Vignotto E**, **van der Wiel K and Engelke S** (2022) Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environmental Data Science 1*, e5.

Buishand TA and Brandsma T (2001) Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resources Research 37*(11), 2761–2776.

Chattopadhyay A, **Nabizadeh E and Hassanzadeh P** (2020) Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems 12*(2), e2019MS001958. https://doi.org/10.1029/2019MS001958.

Chen K, **Kuang C**, **Wang L**, **Chen K**, **Han X and Fan J** (2022) Storm surge prediction based on long short-term memory neural network in the East China Sea. *Applied Sciences 12*(1), 181.

Choi W, **Ho C-H**, **Jung J**, **Chang M and Ha K-J** (2021) Synoptic conditions controlling the seasonal onset and days of heatwaves over Korea. *Climate Dynamics 57*(11), 3045–3053.

Chraibi H, **Dutfoy A**, **Galtier T and Garnier J** (2021) Optimal potential functions for the interacting particle system method. *Monte Carlo Methods and Applications 27*(2), 137–152.

Christidis N, **McCarthy M and Stott PA** (2020) The increasing likelihood of temperatures above 30 to 40 °C in the United Kingdom. *Nature Communications 11*(1), 3093.

Cohen J, **Coumou D**, **Hwang J**, **Mackey L**, **Orenstein P**, **Totz S and Tziperman E** (2019) S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *WIREs Climate Change 10*(2), e00567.

Correoso K (2019) scikit-extremes. Available at https://scikit-extremes.readthedocs.io/en/latest/.

D'Andrea F, **Provenzale A**, **Vautard R and De Noblet-Decoudré N** (2006) Hot and cool summers: Multiple equilibria of the continental water cycle. *Geophysical Research Letters 33*(24).

Davis JC (2002) *Statistics and Data Analysis in Geology*, 3rd Edn. Wiley.

Ding H, **Newman M**, **Alexander MA and Wittenberg AT** (2019) Diagnosing secular variations in retrospective enso seasonal forecast skill using cmip5 model-analogs. *Geophysical Research Letters 46*(3), 1721–1730.

Doersch C (2016) Tutorial on variational autoencoders. Preprint, arXiv:1606.05908.

Felsche E, **Böhnisch A and Ludwig R** (2023) Inter-seasonal connection of typical European heatwave patterns to soil moisture. *npj Climate and Atmospheric Science 6*(1), 1.

Field CB, **Barros V**, **Stocker TF and Dahe Q** (2012) *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.

Finkel J, **Gerber EP**, **Abbot DS and Weare J** (2023) Revealing the statistics of extreme events hidden in short weather forecast data. *AGU Advances 4*(2), e2023AV000881.

Fischer EM, **Seneviratne SI**, **Vidale PL**, **Lüthi D and Schär C** (2007) Soil moisture–atmosphere interactions during the 2003 European summer heat wave. *Journal of Climate 20*(20), 5081–5099.

Fraedrich K, **Jansen H**, **Kirk E**, **Luksch U and Lunkeit F** (2005) The planet simulator: Towards a user friendly model. *Meteorologische Zeitschrift 14*(3), 299–304.

Fraedrich K, **Kirk E and Lunkeit F** (1998) Puma: Portable university model of the atmosphere. *Deutsches Klimarechenzentrum*, page 38.

García-Herrera R, **Díaz J**, **Trigo RM**, **Luterbacher J and Fischer EM** (2010) A review of the European summer heat wave of 2003. *Critical Reviews in Environmental Science and Technology 40*(4), 267–306.

Gessner C, **Fischer EM**, **Beyerle U and Knutti R** (2021) Very rare heat extremes: Quantifying and understanding using ensemble reinitialization. *Journal of Climate 34*(16), 6619–6634.

Grönquist P, **Yao C**, **Ben-Nun T**, **Dryden N**, **Dueben P**, **Li S and Hoefler T** (2021) Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379* (2194), 20200092.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

Ham Y-G, **Kim J-H and Luo J-J** (2019) Deep learning for multi-year ENSO forecasts. *Nature 573*(7775), 568–572.
Hirschi M, **Seneviratne SI**, **Alexandrov V**, **Boberg F**, **Boroneant C**, **Christensen OB**, **Formayer H**, **Orlowsky B and Stepanek P** (2011) Observational evidence for soil-moisture impact on hot extremes in Southeastern Europe. *Nature Geoscience 4*(1), 17–21.
Horowitz RL, **McKinnon KA and Simpson IR** (2022) Circulation and soil moisture contributions to heatwaves in the United States. *Journal of Climate 35*(24), 4431–4448.
Horton RM, **Mankin JS**, **Lesk C**, **Coffel E and Raymond C** (2016) A review of recent advances in research on extreme heat events. *Current Climate Change Reports 2*(4), 242–259.
Jacques-Dumas V, **Ragone F**, **Borgnat P**, **Abry P and Bouchet F** (2022) Deep learning-based extreme heatwave forecast. *Frontiers in Climate 4*, 789641.
Jacques-Dumas V, **van Westen RM**, **Bouchet F and Dijkstra HA** (2023) Data-driven methods to estimate the committor function in conceptual ocean models. *Nonlinear Processes in Geophysics 30*(2), 195–216.
Jézéquel A, **Yiou P and Radanovics S** (2018) Role of circulation in European heatwaves using flow analogues. *Climate Dynamics 50*(3), 1145–1159.
Karlsson M and Yakowitz S (1987) Nearest-neighbor methods for nonparametric rainfall-runoff forecasting. *Water Resources Research 23*(7), 1300–1308.
Kingma DP and Welling M (2013) Auto-encoding variational bayes. Preprint, arXiv:1312.6114.
Koster RD and Suarez MJ (2001) Soil moisture memory in climate models. *Journal of Hydrometeorology 2*(6), 558–570.
Krouma M, **Silini R and Yiou P** (2023) Ensemble forecast of an index of the Madden–Julian oscillation using a stochastic weather generator based on circulation analogs. *Earth System Dynamics 14*(1), 273–290.
Lall U and Sharma A (1996) A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research 32*(3), 679–693.
Lestang T, **Ragone F**, **Bréhier C-E**, **Herbert C and Bouchet F** (2018) Computing return times or return periods with rare event algorithms. *Journal of Statistical Mechanics: Theory and Experiment 2018*(4), 043213.
Lguensat R, **Tandeo P**, **Ailliot P**, **Pulido M and Fablet R** (2017) The analog data assimilation. *Monthly Weather Review 145*(10), 4093–4107.
Loaiza-Ganem G and Cunningham JP (2019) The continuous Bernoulli: Fixing a pervasive error in variational autoencoders. *Advances in Neural Information Processing Systems 32*, 1–11.
Lopez-Gomez I, **McGovern A**, **Agrawal S and Hickey J** (2022) Global Extreme Heat Forecasting Using Neural Weather Models. *Artif. Intell. Earth Syst. 2*, e220035, https://doi.org/10.1175/AIES-D-22-0035.1.
Lorenz EN (1969) Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences 26*(4), 636–646.
Lorenz R, **Jaeger EB and Seneviratne SI** (2010) Persistence of heat waves and its link to soil moisture memory. *Geophysical Research Letters 37*(9).
Lucente D, **Herbert C and Bouchet F** (2022a) Committor functions for climate phenomena at the predictability margin: The example of El Niño Southern oscillation in the Jin and Timmermann model. *Journal of the Atmospheric Sciences 79*, 2387–2400.
Lucente D, **Rolland J**, **Herbert C and Bouchet F** (2022b) Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain. *Journal of Statistical Mechanics: Theory and Experiment 2022*(8), 083201.
Manabe S (1969) Climate and the ocean circulation: I. The atmosphere circulation and the hydrology of the earth's surface. *Monthly Weather Review 97*(11), 739–774.
Matthews B (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure 405*(2), 442–451.
Mehta P, **Bukov M**, **Wang C-H**, **Day AG**, **Richardson C**, **Fisher CK and Schwab DJ** (2019) A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports 810*, 1–124.
Miloshevich G, **Cozian B**, **Abry P**, **Borgnat P and Bouchet F** (2023a) Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. *Physical Review Fluids 8*, 040501.
Miloshevich G, **Rouby-Poizat P**, **Ragone F and Bouchet F** (2023b) Robust intra-model teleconnection patterns for extreme heatwaves. *Frontiers in Earth Science 11*, 1235579.
Min S-K, **Kim Y-H**, **Lee S-M**, **Sparrow S**, **Li S**, **Lott FC and Stott PA** (2020) Quantifying human impact on the 2018 summer longest heat wave in South Korea. *Bulletin of the American Meteorological Society 101*(1), S103–S108.
Miralles DG, **Gentine P**, **Seneviratne SI and Teuling AJ** (2019) Land–atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Annals of the New York Academy of Sciences 1436*(1), 19–35.
Miralles DG, **Teuling AJ**, **van Heerwaarden CC and Vilà-Guerau de Arellano J** (2014) Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nature Geoscience 7*(5), 345–349.
National Academies of Sciences Engineering and Medicine (2016) *Attribution of Extreme Weather Events in the Context of Climate Change*. Washington, DC: The National Academies Press.
Quemener E (2014) *"SIDUS", the solution for extreme deduplication of an operating system. The Linux Journal.*
Ragone F and Bouchet F (2021) Rare event algorithm study of extreme warm summers and heatwaves over Europe. *Geophysical Research Letters 48*(12), e2020GL091197.
Ragone F, **Wouters J and Bouchet F** (2018) Computation of extreme heat waves in climate models using a large deviation algorithm. *Proceedings of the National Academy of Sciences 115*(1), 24–29.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

Rajagopalan B and Lall U (1999) A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research 35*(10), 3089–3101.

Reichstein M, **Camps-Valls G**, **Stevens B**, **Jung M**, **Denzler J**, **Carvalhais N and Prabhat** (2019) Deep learning and process understanding for data-driven earth system science. *Nature 566*(7743), 195–204.

Rezende DJ, **Mohamed S and Wierstra D** (2014) Stochastic backpropagation and approximate inference in deep generative models. In Xing EP and Jebara T (eds.), *Proceedings of the 31st International Conference on Machine Learning*, Volume *32* of Proceedings of Machine Learning Research. Bejing, China: PMLR, pp. 1278–1286.

Rowntree P and Bolton J (1983) Simulation of the atmospheric response to soil moisture anomalies over Europe. *Quarterly Journal of the Royal Meteorological Society 109*(461), 501–526.

Russo S, **Sillmann J and Fischer EM** (2015) Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters 10*(12), 124003.

Schubert SD, **Wang H**, **Koster RD**, **Suarez MJ and Groisman PY** (2014) Northern Eurasian heat waves and droughts. *Journal of Climate 27*(9), 3169–3207.

Schulz B and Lerch S (2022) Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Monthly Weather Review 150*(1), 235–257.

Seneviratne S, **Nicholls N**, **Easterling D**, **Goodess C**, **Kanae S**, **Kossin J**, **Luo Y**, **Marengo J**, **McInnes K**, **Rahimi M**, **Reichstein M**, **Sorteberg A**, **Vera C**, **Zhang X**, **Rusticucci M**, **Semenov V**, **Alexander LV**, **Allen S**, **Benito G**, **Cavazos T**, **Clague J**, **Conway D**, **Della-Marta PM**, **Gerber M**, **Gong S**, **Goswami BN**, **Hemer M**, **Huggel C**, **van den Hurk B**, **Kharin VV**, **Kitoh A**, **Klein Tank AMG**, **Li G**, **Mason S**, **McGuire W**, **van Oldenborgh GJ**, **Orlowsky B**, **Smith S**, **Thiaw W**, **Velegrakis A**, **Yiou P**, **Zhang T**, **Zhou T and Zwiers FW** (2012) Changes in climate extremes and their impacts on the natural physical environment. In Field CB, Barros V, Stocker TF and Dahe Q (eds.), *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the IPCC*. Cambridge: Cambridge University Press, pp. 109–230.

Seneviratne S, **Zhang X**, **Adnan M**, **Badi W**, **Dereczynski C**, **Di Luca A**, **Ghosh S**, **Iskandar I**, **Kossin J**, **Lewis S**, **Otto F**, **Pinto I**, **Satoh M**, **Vicente-Serrano S**, **Wehner M and Zhou B** (2021) *Weather and Climate Extreme Events in a Changing Climate*. Cambridge, UK and New York, NY, USA: Cambridge University Press, pp. 1513–1766.

Seneviratne SI, **Corti T**, **Davin EL**, **Hirschi M**, **Jaeger EB**, **Lehner I**, **Orlowsky B and Teuling AJ** (2010) Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews 99*(3–4), 125–161.

Seneviratne SI, **Koster RD**, **Guo Z**, **Dirmeyer PA**, **Kowalczyk E**, **Lawrence D**, **Liu P**, **Mocko D**, **Lu C-H**, **Oleson KW and Verseghy D** (2006) Soil moisture memory in AGCM simulations: Analysis of global land–atmosphere coupling experiment (GLACE) data. *Journal of Hydrometeorology 7*(5), 1090–1112.

Shukla J and Mintz Y (1982) Influence of land-surface evapotranspiration on the earth's climate. *Science 215*(4539), 1498–1501.

Sohn K, **Lee H and Yan X** (2015) Learning structured output representation using deep conditional generative models. In Cortes C, Lawrence N, Lee D, Sugiyama M and Garnett R (eds), *Advances in Neural Information Processing Systems*, Vol. *28*. Curran Associates, Inc.

Stéfanon M, **D'Andrea F and Drobinski P** (2012) Heatwave classification over europe and the mediterranean region. *Environmental Research Letters 7*, 014023.

Stephenson D (1997) Correlation of spatial climate/weather maps and the advantages of using the Mahalanobis metric in predictions. *Tellus A 49*(5), 513–527.

Thuerey N, **Holl P**, **Mueller M**, **Schnell P**, **Trost F and Um K** (2021) *Physics-based Deep Learning*. WWW.

Thuerey N, **Weißenow K**, **Prantl L and Hu X** (2020) Deep learning methods for Reynolds-averaged Navier–Stokes simulations of airfoil flows. *AIAA Journal 58*(1), 25–36.

van den Dool H (2007) *Empirical Methods in Short-Term Climate Prediction*. USA: Oxford University Press.

Van den Dool H, **Huang J and Fan Y** (2003) Performance and analysis of the constructed analogue method applied to us soil moisture over 1981–2001. *Journal of Geophysical Research: Atmospheres 108*(D16), 8617.

van Straaten C, **Whan K**, **Coumou D**, **van den Hurk B and Schmeits M** (2022) Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in Western and Central Europe. *Monthly Weather Review 150*(5), 1115–1134.

Vargas Zeppetello L and Battisti D (2020) Projected increases in monthly midlatitude summertime temperature variance over land are driven by local thermodynamics. *Geophysical Research Letters 47*(19), e2020GL090197.

Vautard R, **Yiou P**, **D'Andrea F**, **de Noblet N**, **Viovy N**, **Cassou C**, **Polcher J**, **Ciais P**, **Kageyama M and Fan Y** (2007) Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit. *Geophysical Research Letters 34*(7), L07711.

Wang X, **Slawinska J and Giannakis D** (2020) Extended-range statistical enso prediction through operator-theoretic techniques for nonlinear dynamics. *Scientific Reports 10*(1), 1–15.

Watson PAG (2022) Machine learning applications for weather and climate need greater focus on extremes. *Environmental Research Letters 17*(11), 111004.

Webber RJ, **Plotkin DA**, **O'Neill ME**, **Abbot DS and Weare J** (2019) Practical rare event sampling for extreme mesoscale weather. *Chaos: An Interdisciplinary Journal of Nonlinear Science 29*(5), 053109.

Wilks DS (1992) Adapting stochastic weather generation algorithms for climate change studies. *Climatic Change 22*(1), 67–84.

Wilks DS, editor (2019) *Statistical Methods in the Atmospheric Sciences*, 4th Edn. Elsevier.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

Yates D, **Gangopadhyay S**, **Rajagopalan B and Strzepek K** (2003) A technique for generating regional climate scenarios using a nearest-neighbor algorithm. *Water Resources Research 39*(7).

Yiou P (2014) Anawege: A weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development 7* (2), 531–543.

Yiou P, **Cadiou C**, **Faranda D**, **Jézéquel A**, **Malhomme N**, **Miloshevich G**, **Noyelle R**, **Pons F**, **Robin Y and Vrac M** (2023) Worst case climate simulations show that heatwaves could cause major disruptions in the Paris 2024 Olympics. Working paper or preprint.

Zeppetello LRV, **Battisti DS and Baker MB** (2022) The physics of heat waves: What causes extremely high summertime temperatures? *Journal of Climate 35*(7), 2231–2251.

Zhang Y and Boos WR (2023) An upper bound for extreme temperatures over midlatitude land. *Proceedings of the National Academy of Sciences 120*(12), e2215278120.

Zhou S, **Williams AP**, **Berg AM**, **Cook BI**, **Zhang Y**, **Hagemann S**, **Lorenz R**, **Seneviratne SI and Gentine P** (2019) Land–atmosphere feedbacks exacerbate concurrent soil drought and atmospheric aridity. *Proceedings of the National Academy of Sciences 116*(38), 18848–18853.

Appendix

A. Principal component analysis and variational autoencoder

In PlaSim, we encounter high-dimensional fields, compared to what is done in toy models such as Lorenz 63. Generally speaking, one has to deal with problems of "curse of dimensionality" in such cases and the need to somehow project the dynamics $x \to z$. To be more specific, let $\mathcal{D} \in \mathbb{R}^{n,d}$ be a dataset consisting of n samples $x \in \mathbb{R}^d$ and let $z = f(x) \in \mathbb{R}^p$ a lower dimensional projection of the data ($p \ll d$). In what follows we will give two brief descriptions of two widely used projection techniques involving the use of both linear and nonlinear functions f, namely principal component analysis (PCA) and variational autoencoder (VAE). PCA (also known as empirical orthogonal functions) consists of a linear transformation of the data points x, which aims at preserving as much variance as possible of the original dataset. To do so, x is projected onto the low-dimensional space spanned by the first p-eigenvectors of the correlation matrix $\Sigma = \mathcal{D}^T\mathcal{D}$[9]. The dimensionality p of the latent space is determined by the percentage of data variability that is desired to be preserved by the transformation. Here, however, dimensionality reduction is used as a preprocessing step in order to optimize a prediction task. From this perspective, not all information regarding the variability of observations is necessarily relevant to the forecasting task.

The second projection technique we want to discuss consists of a nonlinear dimensionality reduction via VAE (Kingma and Welling, 2013). The idea is to project fields such as geopotential x on the latent low-dimensional space z, from where realistic looking states can be sampled using, for instance, a Gaussian measure. VAE consists of probabilistic *decoder* $p_\theta(x|z)$ and probabilistic (stochastic) *encoder* $q_\phi(z|x)$. The goal is to maximize the probability of the data $p(x)$ in the training set (TS) (Doersch, 2016)

$$p(x) = \int dz p_\theta(x|z)p(z), \tag{A1}$$

where traditionally the prior $p(z) = \mathcal{N}(0, I)$. The choice for the likelihood given by the decoder is often chosen to be Gaussian $p(x|z) = \mathcal{N}(x|f_\theta(z), \sigma^2 I)$ or Bernoulli (which is particularly suitable when the underlying data are binary (Loaiza-Ganem and Cunningham, 2019) and $f_\theta(z)$ can be some neural network with weights θ. Formally, posterior $p_\theta(z|x)$ could be written using Bayes's rule but in most practical cases this leads to intractable expression and methods such as importance sampling may lead to the variance in the estimate that can be very high if the proposal distribution is poor so instead evidence lowest bound is used (Mehta et al., 2019). This leads to approximate posterior which is also parameterized using neural networks $q_\phi(z|x) = \mathcal{N}\left(z|\mu_\phi(x), \Sigma_\phi(x)\right)$. For details on how this is achieved, such as "reparametrization trick" (see, e.g., Rezende et al., 2014). Other variants of VAE have been developed such as conditional VAE (Sohn et al., 2015) but since we are looking for a space on which distance can be computed to allow simulating unconditional trajectories we believe the latent space should be also unconditional. VAEs are not the only probabilistic generative models, and indeed generative adversarial networks (GANs) are often preferred choice since VAEs often produce smoothened images. On the other hand, GANs may suffer from technical difficulties such as mode collapse and are generally more difficult to train.

Once the training is performed (on the TS), we may project the fields in both training and validation set as $\mathcal{Z}(\mathbf{r}, t) \to \mathcal{V}_{\text{tr}}\overline{\mathcal{Z}}(\mathbf{r}, t)$ to obtain latent variables $\mathbf{z}(t) \coloneqq \{z_m(t), m \in 1, \ldots, M\} \sim \mathcal{N}(0, 1)$ that are stacked along with the area integrals

$$\mathcal{X} = \left(\mathbf{z}(t), \langle\overline{T}\rangle_{\mathcal{D}}(t), \langle\overline{S}\rangle_{\mathcal{D}}(t)\right), \tag{A2}$$

For this aim, we use eight-layer encoder–decoder type network with residual connections within encoder and decoder, respectively, which allows projecting $\mathcal{Z}$ to the 16-dimensional latent space. In this case, the input corresponding to $\mathcal{Z}$ field

[9] The first p-eigenvectors are those corresponding to the largest p eigenvalues.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

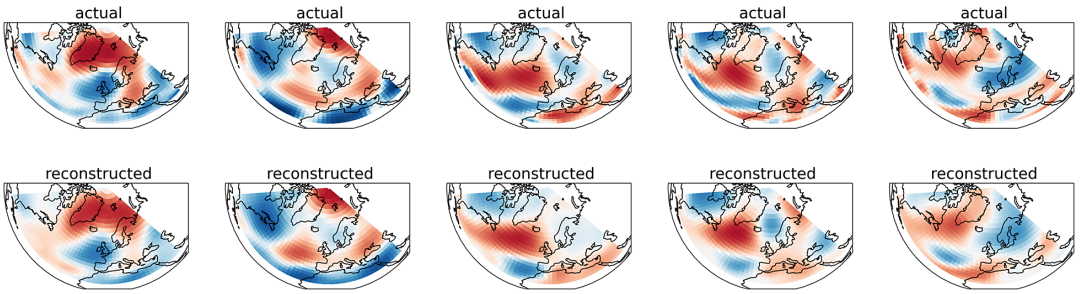


Figure A1. *Top panels: Original geopotential anomalies. Bottom panels: A realization of reconstructed geopotential anomalies after passing them through variational autoencoder (VAE).*

contribution in Eq. (13) is provided from the latent space. The VAE is trained only on the TS. The training takes approximately 50 epochs for each fold on the corresponding TS using TU102 (RTX 2080 Ti Rev. A) GPU which results in total training time of approximately 1 hour. Such large degree of freedom reduction naturally leads to significant gains (two orders of magnitude) in subsequent kDTree computation speed. The forecast skill of this VAE SWG, however, remains the same compared to regular SWG. The original and VAE reconstructed geopotential anomalies are plotted for few samples weather situations in Figure A1.

B. Daily steps

Throughout this paper, we have used $\tau_m = \tau_c = 3$ day step for our Markov chain and coarse graining. While the motivation for this was synoptic decorrelation time, one naturally wonders how optimal is that choice for predicting extremes. Moreover, the CNN was designed to learn from daily fields, which meant we had to shift the NLS curves by 2 days for proper comparisons. Here, we will show what happens when $\tau_m = \tau_c = 1$ day. Given that the temperature autocorrelation time, one would expect subsequent days to be strongly correlated, which would violate Markov's condition. Nevertheless, comparing the blue and orange curves in Figure A2, we hardly see any difference.

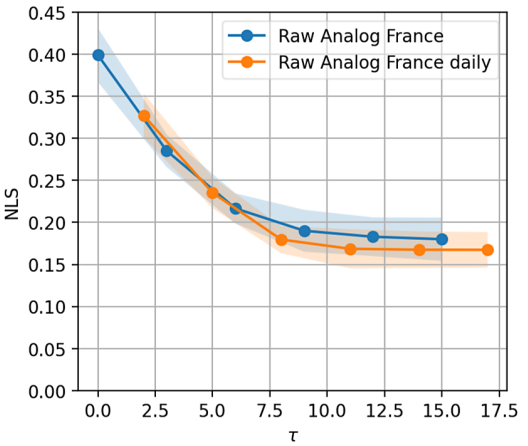


Figure A2. Basic SWG daily increments versus 3 day. Here, we show NLS (equation (6)) as a function of lead time τ for optimal hyperparameters obtained for usual analog Markov chain based on representation (equation (11)) and the procedure described in Section 2.6.1 like in Figure 6. Both figures share identical blue curves (corresponding to the same setup. Orange curve corresponds to a Markov chain whose increment τ_m and coarse-graining time τ_c are set to *1,* that is, no coarse-graining. Note that for fairness of comparison, we have shifted the orange and green curves by 2 days (see discussion in Section 2.4.2). For details on the interpretation of different panels, see the caption of Figure 4.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

C. Committors for 100 years

Results of Figure 4 show that CNN outperforms SWG when using D500 (see Table 1), in other words, 400 years of training data. Climate models allow working with large datasets; however, in practice, we are often limited by the paucity of the observational record. Thus, we provide benchmarks for shorter dataset D100, thus 80 years of training, the same that is used in Section 3.2. The results are plotted in Figure A3, which shows that, while CNN still gives better results on average, the spread of the NLS tends to be quite large. The trends for α_0 and number of neighbors remain the same.

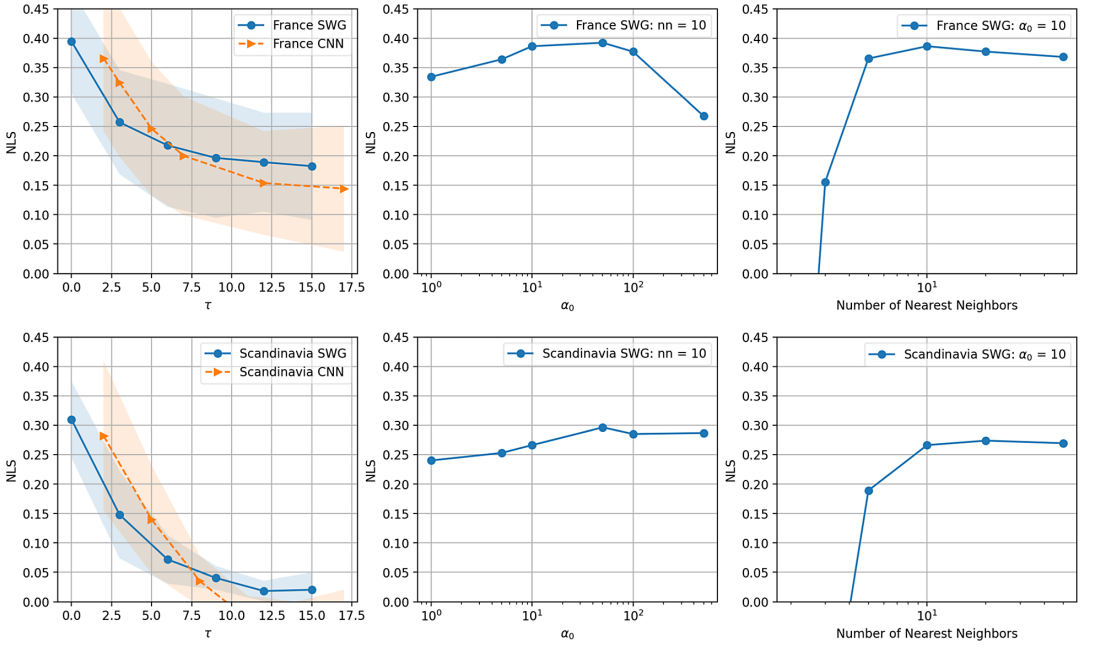


Figure A3. **Basic stochastic weather generator (SWG) (blue curve) versus convolutional neural network (CNN). (Orange curve)** This figure is identical to Figure 4 except for the number of years used to train/validate the algorithm. We have relied on D100 here (Table 1).

D. Return times for different values of the hyperparameter

Here, we attach Figure A4 that is equivalent to Figure 7a in all but the value of $\alpha_0 = 50$. This parameter was chosen optimal for this area in Section 3.1. As we can see, it is not optimal for generating synthetic return times (underestimation). This underestimation is related to the reduction of variance problem. We do not include the case of Scandinavia, but the results look very similar.

https://doi.org/10.1017/eds.2024.7 Published online by Cambridge University Press

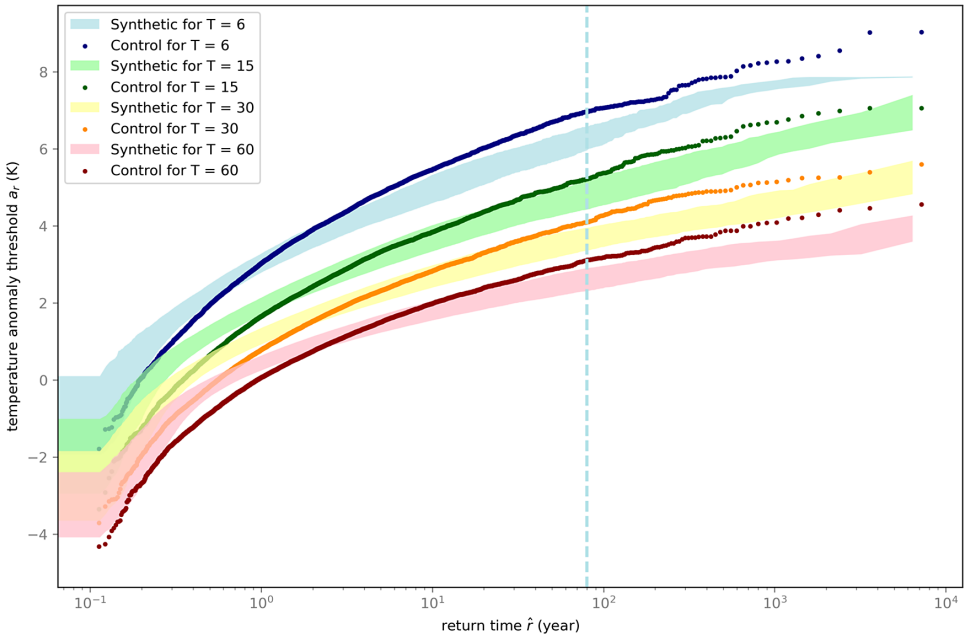


Figure A4. Return time plot for France heat waves using analogues of North Atlantic and Europe. Here, we use parameters $\alpha = 50$ (default), number of nearest neighbors $n = 10$, the analogs are initialized on June 1 of each year (using the simulation data) and then advanced according to Algorithm 1. The trajectory ends at the last day of summer. Each trajectory is sampled 800 times providing much longer synthetic series and thus estimating longer return times. Return times are computed for $T = \{6,15,30,60\}$ day heat waves (indicated on the inset legend), with dots corresponding to the statistics from the control run (D8000, see Table 1), while shaded areas correspond to the bootstrapped synthetic trajectory: the whole sequence is split into 10 portions which allows estimating the mean and variance. The shading corresponds to mean plus or minus one standard deviation.

Cite this article: Miloshevich G, Lucente D, Yiou P and Bouchet F (2024). Extreme heat wave sampling and prediction with analog Markov chain and comparisons with deep learning. *Environmental Data Science*, 3: e9. doi:10.1017/eds.2024.7