Invited Articles

Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data

George Miloshevich [1,2], Bastien Cozian [1], Patrice Abry,[1] Pierre Borgnat [1], and Freddy Bouchet [1,*]

[1]*ENSL, CNRS, Laboratoire de Physique, F-69342 Lyon, France*

[2]*LSCE, UMR 8212 CEA-CNRS-UVSQ, IPSL & U Paris-Saclay, CE, Orme des Merisiers, Gif-sur-Yvette 91191, France*

(Received 30 July 2022; accepted 10 February 2023; published 4 April 2023)

Understanding extreme events and their probability is key for the study of climate change impacts, risk assessment, adaptation, and the protection of living beings. Extreme heatwaves are, and likely will be in the future, among the deadliest weather events. Forecasting their occurrence probability a few days, weeks, or months in advance is a primary challenge for risk assessment and attribution but also for fundamental studies about processes, dataset and model validation, and climate change studies. In this work we develop a methodology to build forecasting models which are based on convolutional neural networks, trained on extremely long 8000-year climate model outputs. This approach is parallel to weather model forecasting and has complementary scopes. Because the relation between extreme events is intrinsically probabilistic, we emphasize probabilistic forecast and validation. We demonstrate that neural networks have positive predictive skills, with respect to random climatological forecasts, for the occurrence of long-lasting 14-day heatwaves over France, up to 15 days ahead of time for fast dynamical drivers (500 hPa geopotential height fields), and also at much longer lead times for slow physical drivers (soil moisture). This forecast is made seamlessly in time and space, for fast hemispheric and slow local drivers. The method is easily implemented and versatile. We find that the neural network selects extreme heatwaves associated with a north hemisphere wave-number 3 pattern. We argue that this machine learning approach should be key in the future for quantitative process studies, model intercomparisons, and dataset studies. For instance, we find that the 2 meter temperature field does not contain any new useful statistical information for heatwave forecast, when added to the 500 hPa geopotential height and soil moisture fields. The main scientific message is that most of the times, training neural networks for predicting extreme heatwaves occurs in a regime of lack of data. We suggest that this is likely to be the case for most other applications to large-scale atmosphere and climate phenomena. Depending on the information to be learned, training might require dataset lengths as long as several thousands of years, or even more, for optimal forecasting skill. For instance, using 100-year-long training sets, a regime of drastic lack of data, leads to severely lower predictive skills and general inability to extract useful information available in the 500 hPa geopotential height field at a hemispheric scale in contrast to the dataset of several thousand years long. Even with several-thousand-year-long datasets, no convergence is observed in the predictive skills coming from hemispheric geopotential height fields. We discuss perspectives for dealing with the lack of data regime, for instance, rare event simulations and how transfer learning may play a role in this latter task.

DOI: 10.1103/PhysRevFluids.8.040501

*Freddy.Bouchet@cnrs.fr

I. INTRODUCTION

A. Context: The need for probabilistic forecast of extreme climate events

1. Lack of data for the most impactful climate extremes

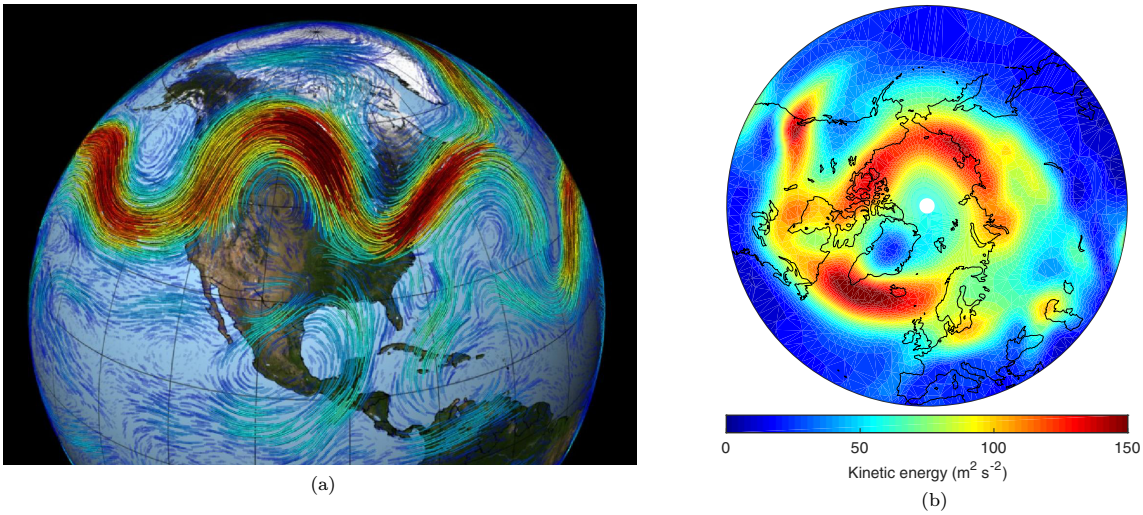
Climate change is one of the major challenges of modern societies [1] and will significantly affect humans and other living beings. Its most severe impacts are caused by rare and extreme events [2]. For instance, since 1998, most of the deaths which were caused by major related disasters have been linked to only three climate events [3]: the Western European extreme heatwave during the summer of 2003 [4], the storm surge related to cyclone Nargis in Myanmar in 2008 [5], and the extreme heatwave in Russia during the summer of 2010 [6,7], with death tolls of about 70 000, 150 000, and 100 000, respectively. Each of the physical events causing these impacts where unprecedented in the historical record, in their category, as was the case for 2021 northwestern North America heatwave [8].

These examples illustrate the need to study very rare events, most of them unprecedented. Faced with this scientific challenge, given the drastic lack of historical data, any statistical approach based solely on observation data is bound to fail. The only sensible approach is thus to use climate or weather model data, whose biases are properly characterized [9] through process studies. We choose extreme heatwaves as our topic because they will be among the most impactful climate extreme events in the future [2] and because climate models are known to reproduce better their dynamics than other extreme events, because they are large-scale phenomena less affected by small-scale physics. In the present study, we will use 8000-year-long PlaSim (planet simulator [10]) climate model simulations to devise a forecast tool for extreme heatwaves in midlatitude dynamics. We will use modern machine learning techniques, as well as tune and develop them to specifically study very rare events.

2. The compound effects of geostrophic turbulence and slow drivers for extreme heatwaves

From a fluid dynamics point of view, studying midlatitude extreme heatwaves amounts to quantifying the probabilities of rare fluctuations of the dynamics of the turbulent Earth troposphere. Midlatitude atmospheric flow is turbulent and characterized by balance between the Coriolis force and pressure gradient (geostrophic turbulence), and whose dynamics is dominated by large-scale unstable patterns. The main large-scale features are the two jet streams (one per hemisphere). These are strong and narrow eastward air currents, located at midlatitudes with maximum velocity of the order of $40\,\mathrm{m\,s^{-1}}$ close to the tropopause [see Fig. 1(a)]. The climatological position of the northern hemisphere jet stream in the PlaSim model is seen in Fig. 1(b), that represents the time average of the kinetic energy due to the horizontal component of the velocity field at 500 hPa pressure surfaces. The jet stream's meandering dynamics, due to nonlinear Rossby waves, is related to the succession of anticyclonic and cyclonic anomalies which characterize weather at midlatitudes. It is well known that midlatitude heatwaves, like the 2003 Western European heatwave or the 2010 Russian heatwaves, are due to rare and persistent anticyclonic anomalies (or fluctuations), that arise as either blocking situations [11,12] (omega shape quasi-stationary patterns) or Rossby wave breaking or shifts of the jet stream, or more complex dynamical events leading to some quasi-stationary patterns of the jet stream. Studying extreme heatwaves then amounts to studying the nonlinear and turbulent dynamics of the atmosphere, or the consequences thereof.

Studying extreme heatwaves is however not just a problem in fluid mechanics pertaining to the extremes of turbulent fluctuations. While there is a clear connection between the physical hazard (the temperature) and the fast dynamical drivers through the fluid dynamics of midlatitude troposphere (jet stream, Rossby waves dynamics and blocking situations), it is also well known that slow drivers, sometimes also called modulators, influence the frequency and the probability of the fast dynamical drivers [13,14]. For instance, deficit of soil moisture acts as a positive feedback on heatwave situations [15–27]. Indeed, in normal conditions, evaporation of soil moisture cools the ground and partition the heat flux from the heated soil to the low altitude air masses into latent and sensible heat.



(a) (b)

FIG. 1. (a) Snapshot of wind-speed velocity at the top of the troposphere, showing the jet stream over North America (from NASA). (b) Averaged horizontal kinetic energy at 500 hPa (midtroposphere) in the PlaSim model [28], showing the averaged northern hemisphere jet stream.

This effectively cools the lowest part of the atmosphere. The related produced cloud covers might also in complement affect the radiative balance. By contrast, lack or low values of soil moisture will thus favor hotter heatwaves, while persistent heatwaves themselves or precipitation deficit might favor low soil moisture values (see Refs. [13,26] for a more precise discussion of the related physical phenomena). This leads to a strong soil-moisture deficit/heatwave positive feedback. Soil moisture is called a slow driver, or modulator, because the typical timescales of its variations, from weeks to seasons, is much longer than the synoptic timescales associated with midlatitude turbulent dynamics, of a few days up to a week. Beyond these land/atmosphere couplings through soil moisture, other slow drivers for midlatitude heatwaves are classically studied in the climate literature [13,14], for instance, sea surface temperature at ocean scales, tropical or stratospheric forcing related to low modes of variability of the climate system (ENSO, QBO, and so on). It has been identified for a very long time that most extreme events are compound events of several drivers, and a recent work proposes a qualitative classification of compound types [29].

Understanding the relative effect of fast dynamical drivers and slow physical drivers is very interesting from the fundamental fluid mechanics perspective: it amounts to understanding the effect of slow varying and weak boundary condition changes on turbulent statistics. It is also critical to predict the impact of climate change on future heatwave probabilities [30,31]. The topic of physical mechanisms behind extreme heat events [32–35], and how these mechanisms may change with climate change, is an emergent area of research, with much evidence still required. For instance, the dependence between temperature and precipitation is projected to increase in many land areas, particularly in the northern hemisphere, leading to a doubling in probability of extremely hot and dry summers on top of long-term climate trends [36]. Diagnosis of heat event mechanisms is critical to understanding the potential for nonlinear responses in extreme heat beyond those expected from global mean warming alone [13].

3. The need for probabilistic forecast for rare events

The main aim of this work is to develop a setup for probabilistic forecast of extreme heatwaves, based on machine learning. Forecasting extreme events may be crucial for the sake of prevention and information dissemination for limiting risks through anticipated action, which is one of our main motivations. Another important complementary goal is the understanding of the fluid

mechanics and physical processes leading to heatwaves. As just explained, a scientific challenge is to disentangle the effect of fast dynamical drivers, for instance, atmospheric dynamics and geostrophic turbulence, and slow physical drivers, for instance, soil moisture feedbacks. There is a need for a new methodology to achieve this goal. A very common approach in the climate literature is to plot maps of dynamical or physical variables, conditioned on the outcome of the extreme events, called composite maps. While interesting, such composite maps inform only on the state of the system once it is known that the event actually occurred (*a posteriori* conditioning on the event). A much more important question is to understand which states of the systems are more likely to lead to the extreme events (*a priori* conditioning on the state of the system). While these two properties are related through Bayes law, composite maps alone are not useful to study the more important *a priori* conditioning on the state of the system. To predict the probability of a future extreme conditioned on the state of the system, the so-called committor function, one actually has to build a forecasting tool able to estimate this probability.

However, this forecast task is often considered as extremely difficult because of the very large amount of data needed and some methodological difficulties. A review of quantification methodologies to disentangle preconditions of high-impact events [37] describes regression techniques and event compositing, and stress the needed long dataset. The probability that an extreme event occurs, conditioned on the state of the system, is a function of the system state and is called a committor function [38,39]. This is the proper tool to disentangle the mechanisms that lead to extreme events in a fully nonlinear setup, beyond usual restrictive assumptions. One of the main conclusions of this work will be that machine learning and neural networks provide a way to compute committor functions, by solving a probabilistic forecast problem. This, however, requires to understand predictions using neural network in a probabilistic framework, as is further explained in the following sections.

Another reason why the forecast should be probabilistic is because for turbulent flows, like the atmosphere, the relation between between meteorological fields (predictors) and extreme events is probabilistic. This is for three reasons. First, as originally understood by Lorenz in his 1969 paper [40], for many chaotic dynamical systems with many degrees of freedom and a hierarchy of spatial and temporal scales, the memory of the initial condition is lost after a finite amount of time and the dynamics behave in an intrinsically stochastic way. In the case of Earth's atmosphere, the prediction of synoptic scales is intrinsically stochastic after a few days to a week [40]. Second, because for practical reason we have an incomplete knowledge of the initial conditions, then the initial condition should be considered as stochastic. Third, because the predictors we use do not describe the complete set of initial conditions. This means that assuming a one-to-one relation between the predictors (physical fields) and prediction (extreme heatwave after a τ-day delay) does not make sense, even in principle. The relation between predictors and prediction should be probabilistic. Our task will actually consist in predicting the occurrence of extreme heatwaves starting τ days ahead, given the knowledge of some physical fields that characterize the state of the atmosphere and soil moisture today. This is actually a classification task: given some images, or data stored in a vector, one seeks to associate a class among two: either the heatwave occurs (class one) or not (class two). From the point of view of machine learning this is very similar to image recognition. However, when recognizing the presence of a cat in an image, a one-to-one relation between the image and the class actually exists: either a cat is present on the image or not. The machine learning tool can then associate a probability to the prediction, which can be interpreted as the level of confidence of the tool due to its practical limitations, associated, for instance, with incomplete training or lack of data. By contrast, when predicting extreme events for a chaotic dynamical system, the relation between the predictors and the classes is intrinsically probabilistic. Then the probability given by the forecasting tool should be interpreted as intrinsic, and reflect both intrinsic uncertainty due to the unknown real probability, and practical uncertainty due to the limitations of the learning tool. We will see the consequence of this remark on the machine learning implementation and testing.

B. State of the art for machine learning approaches for forecasting climate extremes

Extreme event prediction and, more broadly, weather forecast, have recently attracted numerous studies which exploit machine learning techniques. This is contrasted with the mainstream approach which involves running expensive numerical weather prediction models. This dichotomy between physics-based and pattern-based prediction is well documented in the review articles [41,42]. Notably there are some studies which attempt to bridge the gap by combining the approaches [43]. Overall pattern-based techniques such as neural networks or analog method may do relatively well in seasonal/subseasonal forecasting [44], at timescales longer than Lyapunov time.

As an example, a forecast tool of the El Nino southern oscillation (ENSO) index has been built using a convolution neural network (CNN) [45]. A model pretrained on CMIP ensemble was then trained on historical reanalysis. Other developments include [46] where 500 hPa geopotential height was predicted using gridded reanalysis data.

Deep learning has been applied to the spatial and temporal detection of extreme weather events such as hurricanes [47,48], tropical cyclones [49], droughts [50,51], storm surges [52] and wind power generation [53], and heatwaves [54,55]. For further reference see Ref. [56] and the citations therein. Recently, for predicting extreme heat events globally, neural networks trained on reanalysis data [57] have given positive skill compared to the ECMWF subseasonal-to-seasonal control forecast after two weeks. Neural networks, where the 500 hPa geopotential height and surface temperature were used as predictors, were able to predict both short duration [54] or long-lasting [55] heatwaves, when trained on climate model data. In these works, the performance was evaluated and tuned to metrics related to confusion matrix such as Matthew's correlation coefficient, which are well suited for one-to-one or deterministic relation between predictors and prediction. However, making and testing probabilistic forecast is very important, as stressed in Sec. I A. Changing this paradigm requires to test probabilistic forecasts using probabilistic scores. Such probabilistic scores have been used for a long time in evaluation of weather model forecast, for instance, the logarithmic score which is an objective and proper score [58]. Another proper score, although less applicable to rare events, a Brier score, was used in a recent study [59] where subseasonal forecast was made for high temperatures in western and central Europe using random forest approach applied to ERA5 reanalysis. In general, traditional techniques such as random forest are quite competitive with neural network approaches when dealing with smaller dataset sizes. Finally, we stress the work by Ref. [60] that produces state-dependent probabilistic Madden-Julian oscillation forecast with neural networks.

One of the aims of this work will be to perform and test *probabilistic* forecasts for the first time to the best of our knowledge, using neural networks, for extreme climate events or atmospheric dynamics phenomena. Probabilistic forecast will be performed through a natural interpretation of Softmax probabilities. When working with rare events, because of learning difficulties with class imbalance, it might be useful and efficient to undersample the majority class. Using such undersampling at the same time as making a probabilistic forecast however requires an interpretation of Softmax probabilities that takes into account the undersampling rate [61,62], as will be explained.

In a recent work [55], neural networks have been used to forecast extreme heatwaves. One of the key originality of this first work was to consider for the first time the forecast of long-lasting extreme heatwaves, with durations of several weeks. This is a key point as most of the extreme heatwaves with the largest impact, for instance, the Western European one in 2003, the Russian one in 2010, or the North American Pacific coast one in 2021, lasted long, from two to five weeks. The lack of comprehensive studies of the statistics of long-lasting events has actually been stressed in the last IPCC report [1]. We refer to the Introduction and Sec. 2.1 of Ref. [55] for a thorough discussion of this crucial point about the definition of extreme heatwaves. Other key achievements of Ref. [55] were to demonstrate the efficiency of neural network to predict long-lasting events, to implement and assess the interest of large-class undersampling and transfer learning. From the point of view of machine learning methodology, this new paper builds on the previous one [55], but with several

crucial methodological improvements: probabilistic forecasts and tests, implementation of large class undersampling in a probabilistic setup, and use of both fast and slow physical and dynamical drivers. Another distinction in this new paper is that we work with fields in the physical space rather than in the Fourier space. This proves more efficient from the point of view of the forecast skills, and especially so when studying the importance of local versus global information for best performance. We also use a much longer, 8000-year dataset, which represents the climate of the decade 1990–2000 with a more realistic daily cycle, and which allows for a detailed study of the lack of data regime and a more comprehensive analysis of the various drivers.

C. Goals, contributions, and outline

Section I A discusses the importance of forecasting long-lasting extreme heatwaves because of their impact. We have also reviewed the large interest in the climate literature for understanding the respective effects of fast dynamical drivers, related to troposphere dynamics, geopotential height and temperature maps, and slow physical drivers, for instance, soil moisture. We have stressed that it is crucial that this forecast should be probabilistic and that it has to be performed in a regime of lack of data.

To build a machine learning forecast setup that will be able to address these goals, several new methodological contributions are proposed in this work. To devise and use neural networks that predict probabilistic forecasts will be our first methodological goal. The neural network output will be the probability of the extreme event, as a function of the state of the system, also called a committor function. As a second methodological goal, this probabilistic forecast will be validated using a probabilistic score. Because of the regime of lack of data and large class-imbalance, we will propose and test a large class undersampling strategy adapted to probabilistic forecast, as a third methodological goal. We will demonstrate the efficiency of these three methodological contributions for predicting long-lasting extreme heatwaves using climate model outputs.

Using this neural network technique, adapted and validated in a probabilistic framework, the following fluid mechanics and climate goals will be addressed. First the prediction capabilities of the neural network when changing the predictor fields will be investigated. We will demonstrate that the network is able to make best predictions when combining fast and slow drivers, with a relatively stronger contribution of fast drivers for shorter lead times and stronger contribution of slow drivers for longer lead times. Second, the effects of dataset lengths will be studied, a very important question in a regime of lack of data. We will actually conclude that the dataset length has to be extremely long for proper convergence of the learning, and that in such a regime, the optimal learning results in a tradeoff between the size of the physical domain and data availability. For instance, for predicting extreme heatwaves over France, it is optimal to use local data (North Atlantic and Europe) with one hundred-year long datasets, while it is optimal to use global data for few thousand-year long datasets. Finally, to make a connection with fluid-mechanics, we will study the interpretability of the learned committor function by computing composite maps conditioned on high extreme event probabilities.

Those goals will be studied using long datasets from the PlaSim climate model [10,63]. This model has a very realistic fluid dynamics component, similar to the climate models used for CMIP experiment described in IPCC reports. However, its physical parametrizations are simpler than such models, for instance, the ones used for CMIP6 experiments. This significantly reduces PlaSim's computation time by about a factor 100. It is thus suitable for methodological development and first studies, using extremely long datasets. It is ideal for studying learning convergence in the lack of data regime. Section II gives a more detailed introduction to the PlaSim model, its output fields, and the dataset size, resolution, quality and richness. It also discusses better the physical interest and limitations of this dataset. In Sec. II, heatwaves, predictor fields, and the probabilistic prediction problem studied in this work are also defined precisely.

Section III formalizes the problem of probabilistic forecast using neural networks, and discusses proper probabilistic scores and their relationship with cross-entropy and machine learning loss

functions. We also introduce a very useful normalized logarithmic score (NLS) which is positively oriented (the larger the better), and takes value zero for prediction according to the climatological frequency and one for perfect prediction. In addition, the relation of these scores to the Brier score, another classical probabilistic score, is discussed.

Section IV first describes the convolutional neural network architecture used here and its learning conditions. It further explains the promoted probabilistic strategy: to train the architecture as a classifier and to use it as a conditional probability predictor. The detailed training protocol using a classical cross-validation procedure, to assess confidence and reproducibility in performance, is presented. Finally, it explains the methodology for large class undersampling in a probabilistic framework.

Section V presents the fluid mechanics and climate science results. It first demonstrates which combination of physical predictors, among slow and fast drivers, give the best prediction skill. This stresses the potential of machine learning for dealing adequately with separating their respective effects. We then discuss the importance of the dataset length, the convergence of the learning skill when the dataset length is changed, and the tradeoff between dataset length and spatial extension of the physical fields. We verified that our neural network has a better performance than traditional approaches, for instance, logistic regression using empirical orthogonal function (EOF) decomposition (or PCA), although the details of this analysis are not reported in this paper. We also test the continuity and consistency of the committor function prediction when the time lag is varied.

Finally, Sec. VI discusses conclusions and perspectives.

II. LONG-LASTING HEATWAVES AND PLANET SIMULATOR DATA

Using weather maps, we aim at predicting the probability of occurrence of extreme long-lasting heatwave that starts τ days later. In the remainder we will refer to this parameter as *lead time* or sometimes *lag time*. We first define long-lasting heatwaves in Sec. II A, describe the possible predictors in Sec. II B. In Sec. II C, we explain that our approach is actually a way to compute the committor function, a key function in the field of rare event analysis and simulations. The actual weather maps we use are outputs of the PlaSim climate model, which is described in details in Sec. II D.

A. What are long-lasting heatwaves?

Several indices have been used in the meteorology, climate, and impact literature, to define heatwaves, for different purposes [14]. However, long-lasting heatwaves are the most detrimental to health [6] and other living beings. Moreover, many of the extreme heatwaves with the largest impact, for instance, the Western European one in 2003, the Russian one in 2010, or the North American Pacific coast one in 2021, lasted long, from two to five weeks. They were often composed of several subevents with the classical definitions [14]. We want to use a definition of heatwaves that actually involves a measure related to both the persistence and the amplitude of air temperature close to the ground.

We thus define heatwave as time and area average of daily 2 meter (2 m) temperature. Seminal studies [6,64,65] of the 2003 and 2010 heatwaves already considered averaged temperature over variable long time periods (7 days, 15 days, 1 month, 3 months). Several recent works [27,28,66–69] have studied heatwaves based on time and space average of either the 2 m temperature or of the surface temperature. This viewpoint is expected to be complementary with the classical definitions [14], and quite relevant to events with the most severe impacts. Such definitions have the advantage to define events at a synoptic scale which are geographically located and which begin at a specific date. This is well suited for a forecast perspective.

The daily 2 m temperature $T_{2\mathrm{m}}(\vec{r}, t)$ is a spatial field that depends on the location $\vec{r}$ and calendar day t (also called time). We use daily averages. Statistics of $T_{2\mathrm{m}}(\vec{r}, t)$ are affected by the seasonal cycle. We compute anomalies (i.e., fluctuations) by subtracting the statistical average $\mathbb{E}(T_{2\mathrm{m}})(\vec{r}, t)$ at

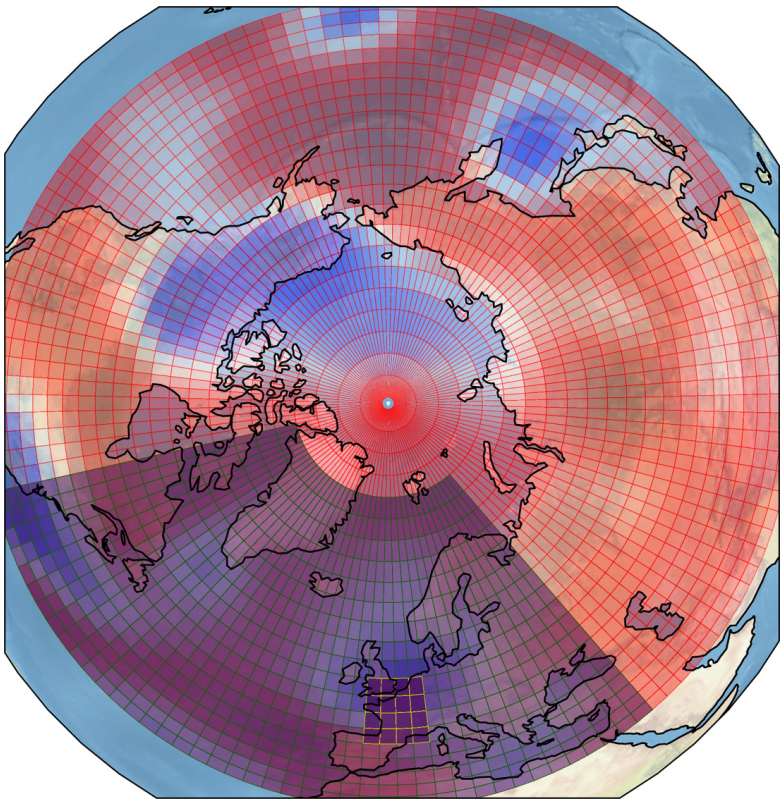


FIG. 2. For all fields, we will use gridded data on the mid and high latitude northern hemisphere as represented by red meshlines. The figure also features in purple the area $\mathcal{D}$ (France). The North Atlantic Europe sector is represented in blue.

each point $\vec{r}$ and each time t. We compute the space and time averaged 2 m temperature anomalies:

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

where T is the length of the time average, i.e., heatwave duration, and $\mathcal{D}$ is the spatial area for the heatwave of interest. The heatwave duration T and area $\mathcal{D}$ should be understood as parameters that can be changed from one study to another, depending on the kind of heatwave of interest. T ranges typically from one-day (short duration) to three months (a season). $\mathcal{D}$ is typically of the size of the synoptic scale. The synoptic scale, of about 1000 km at midlatitude, is the order of magnitude of correlation length for troposphere dynamics, and corresponds to the typical size of anticyclones, cyclones and the jet stream meanders. In the present work, mainly aimed at methodological developments, we set $T = 14$ days (two-week heatwaves), and $\mathcal{D}$ to be the France area: the set of grid-points corresponding to France area is visible on Fig. 2.

We consider summer statistics, during the months of June, July, and August. More precisely we consider $A(t)$ for t between June 1 and August 16 (inclusive), such that the time average in Eq. (1) involves only days during the months of June, July, and August. In PlaSim each month spans 30 days, thus the total length of the period of interest is 77 days. The statistics of A is considered as approximately stationary during the summer, although there is actually some very small nonstationarity. For instance, monthly breakdown for standard deviation is $\sigma_A = 1.58$ K in June, $\sigma_A = 1.49$ K in July, and $\sigma_A = 1.32$ K in August. We see that the variations of the standard deviations from one month to another are much smaller than the standard deviations themselves, and much smaller than the variations of the time averaged temperature from one month to another.

Extreme heatwaves are defined as rare large values of the time and space average $A(t)$. Following the previous works [55], we define an extreme heatwave as an event (a day) for which the time and space averaged 2 m temperature anomaly exceeds the threshold α: $A(t) > \alpha$. We introduce an indicator variable $Y(t)$ which is equal to 1 when $A(t) > \alpha$ and 0 otherwise. We have $K = 2$ classes of events: heatwaves when $Y = 1$ and no heatwave when $Y = 0$. $Y(t)$ is sampled daily. When $Y(t) = 1$, t is the day for the start of the heatwave and the heatwave lasts for T days, by definition. The threshold α can be changed depending on the heatwaves of interest. The number of classes K could also be changed. For this methodological study, we use $K = 2$ and α such that the heatwave class contains 5% of the total number of summer days (excluding the last two weeks for the reasons explained). For the PlaSim model data described below, this corresponds to $\alpha = 2.7$ K.

B. Predictors for heatwaves

Our objective is to develop a prediction tool for extreme heatwaves. From the knowledge of observed weather fields (predictors), we want to predict the probability of the event τ days later. We will vary the parameter τ to understand how the predictability potential changes with this lead time τ.

The choice of good predictors, among all possible weather fields, is a key practical and physical question. From the common knowledge among weather and climate scientists, it is known that maps of the 2 m temperature T_{2m} and of the geopotential height Z (in meters), for instance, on the 500 hPa isopressure surfaces (the 500 hPa geopotential height) are relevant variables. The geopotential height at 500 hPa (close to the middle of the troposphere) is considered an excellent representation of the dynamical state of the atmosphere. Indeed, it is closely related to pressure variations at a fixed altitude, to anticyclones (positive values), and to cyclones (negative values), in the lower troposphere. Moreover, on those surfaces the wind flows along the isolines of the geopotential height, to a good approximation. The 2 m temperature T_{2m} used as a predictor is directly related to the kernel of the integral [Eq. (1)] whose fluctuations we seek to predict, and gives further information on lower atmosphere dynamical processes compared to the 500 hPa geopotential height. We stress that $A(t + \tau)$ involves the time average over T days of $T_{2m}(t)$. The knowledge of $T_{2m}(t)$ thus provides only partial estimate of $A(t + \tau)$, just by virtue of persistence prediction or low-tropospheric advection. However, given that the correlation time is of order of a few days, to be compared to $T = 14$ days, this information gives a relatively small predictive power by itself even for $\tau = 0$, which quickly diminishes for τ larger than a few days. In the following, we never use directly the 2 m temperature or 500 hPa geopotential height fields, but rather their anomalies by subtracting their seasonal average. Representative examples of snapshots of 2 m temperature or 500 hPa geopotential height fields is shown on Fig. 3.

The 2 m temperature T_{2m} and the 500 hPa geopotential height Z are fields that evolve through the chaotic dynamics of the atmosphere with a typical time called the synoptic timescale, of order of a few days. It is known that both temperature and geopotential lose the memory of the initial condition of the atmosphere, and that their auto- and cross-correlations decay, after times of order one to two weeks. This is the predictability margin for weather. As a consequence, we expect these fast dynamical fields to lose their predictive power after times of order 15 days at most. Those fields have actually been used as predictors for machine learning approaches in past studies, either for 5-day [54] or 14-day [55] heatwaves. Those works have indeed demonstrated the predictive value of these fields for time delays τ up to about 15 days.

One of the aims of this work is to combine predictors based on fast dynamical weather fields, just discussed, with other drivers with a much slower typical evolution. As explained in the introduction, soil moisture deficits and heatwaves are coupled through positive feedback loops and reinforce each other on various timescales [15–23,25–27]. Because soil moisture is the stock of all water in the soil, it evolves on rather long timescales. Its value is correlated over weeks to months [22,70]. At a specific time, the effect of soil moisture on dynamical variables which directly cause heatwaves, is rather weak. It is basically only able to modulate the energy budget and temperature. For this reason,

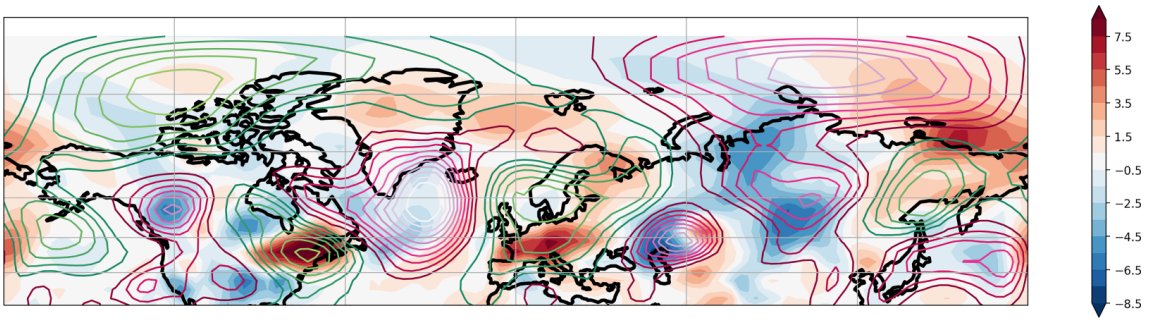


FIG. 3. Snapshot of 2 m temperature anomaly and 500 hPa geopotential height anomalies. Temperature colormap is shown on the colorbar. The geopotential height anomalies are plotted via contour lines, colored green for positive and purple for negative with a separation of 20 m between the lines. The lowest level also corresponds to 20 m. This map displays a synoptic situation (daily averaged) on the first day of the strongest heatwave in the dataset.

we expect the soil moisture predictor to have a much smaller predictive value than fast dynamical fields in the short run, but by contrast it should provide extended memory effect. One of the aims of this paper will be to test these simple qualitative ideas and to make them precise and quantitative, using the machine learning tool.

Other slow evolving drivers might be considered as good predictors for heatwaves, i.e., sea surface temperature, or slow modes of variability of the atmosphere, ice and snow cover, and so on [59]. However, as explained in the next section, the dataset we use is not suited for this, and such studies will be postponed for future works.

For simplicity, in the following, we denote Z the 500 mbar geopotential height, $T_{2\mathrm{m}}$ the 2 m temperature and S the soil moisture fields.

A second question pertaining to the predictors is to decide whether to use either all their values in the northern hemisphere mid and high latitudes, or rather to use their values on a restricted area close to the heatwave region or some intermediate scale area. Because the soil moisture feedbacks are local processes, soil moisture is expected to be relevant locally, for instance, around the area where the heatwave occurs, while geopotential height is expected to play a role on a more hemispheric scale. For instance, for many past studies of weather and climate phenomena over Europe, values of predictors over the North Atlantic and European area were typically used for the geopotential height. As examples, several very interesting studies, using analogs as a learning tool, have shown that the North Atlantic and Europe sector might be the best choice [71,72], with the interpretation that this area carries the most relevant information. For temperature, it is less straightforward to assert *a priori* whether just local or hemispheric information matter for heatwave prediction.

We will use the machine learning tool to assess the question of such optimal predictors. We will use either the northern hemisphere mid and high latitude fields, corresponding to the values of the fields above 30N, spanning a 22×128 grid points, or to a restricted area, corresponding to a 18×42 grid points, referred as the North Atlantic and European sector, or to the France area (see Fig. 2). On these restricted areas, the 500 hPa geopotential fields Z, the 2 m temperature $T_{2\mathrm{m}}$, and the soil moisture S are denoted, respectively, Z_{NH}, T_{NH}, S_{NH} for the northern hemisphere, Z_{NAE}, T_{NAE}, S_{NAE} for the North Atlantic and European sector, and Z_{F}, T_{F}, S_{F} for France (see also Sec. IV B).

These data can be stacked, as different input features, for the learning procedure. It was shown in Ref. [55] that stacking was the best approach for combination in this context, as it allows to capture interaction between learned features.

More abstractly, the set of predictors (one of the combinations of $Z_{\mathrm{NH}}, T_{\mathrm{NH}}, S_{\mathrm{NH}}$, $Z_{\mathrm{NAE}}, T_{\mathrm{NAE}}, S_{\mathrm{NAE}}$, or $Z_{\mathrm{F}}, T_{\mathrm{F}}, S_{\mathrm{F}}$), is globally called $\mathbf{X} \in \mathbb{R}^d$. From the value of the vector $X(t-\tau)$ at some time $t-\tau$, we aim to predict the probability $p(\boldsymbol{x})$, that $Y(t)$ is equal to one (to observe an heatwave at time t), given that $\mathbf{X} = \boldsymbol{x}$. This is a probabilistic classification task, conditioned on the

state $\boldsymbol{x}$. In the context of stochastic processes $p(\boldsymbol{x})$ is called a committor function, as we will explain below.

C. Committor functions for extreme heatwaves

We note that in the theory of rare events for stochastic processes, the probability $p(\boldsymbol{x})$ to observe a rare event conditioned on the state of the system $\boldsymbol{x}$, is called a committor function [73,74]. The committor function is the probability of hitting the target set $\mathcal{B}$ before the set $\mathcal{A}$: $\mathbb{P}[\tau^\star_{\mathcal{B}}(\boldsymbol{x}) < \tau^\star_{\mathcal{A}}(\boldsymbol{x})]$, where $\tau^\star_{\mathcal{A}}$ and $\tau^\star_{\mathcal{B}}$ are the first hitting times of the sets $\mathcal{A}$ and $\mathcal{B}$, given that the trajectory started at $\boldsymbol{x}$. It is possible to extend this definition to time-dependent sets $\mathcal{A}$ and $\mathcal{B}$ with an extended dynamical system; see Ref. [39]. For example in our case, the set $\mathcal{A}$ is simply the set of the model fields $(\boldsymbol{x})$ such that we have a heat wave which starts at time τ, and the set $\mathcal{B}$ is the complementary set to $\mathcal{A}$. For each value of τ, the probability p that we seek to predict is then a committor function.

Committor functions are extremely useful in the simulation and prediction of rare events. Several computations of committor functions have been performed with applications in either geophysical fluid dynamics or in climate sciences [38,39,75–78], using either direct or involved approaches. However, computing or sampling a committor function is a very difficult task, especially in large dimensional spaces, because it requires to gather a very large amount of statistics about rare events.

Many interesting methods have been or are currently being devised to learn committor functions: based on direct machine learning [79], or using approximations of the stochastic dynamics, for instance, using the analog approach [77], or Galerkin approximations of the Koopman operator [80,81]. The present work, by successfully implementing a neural network that efficiently forecasts $p(\boldsymbol{x})$ the probability of extreme heatwaves conditioned on the state of the system $\boldsymbol{x}$, demonstrates that neural networks are useful and efficient to compute committor functions for extreme heatwaves, for a dynamics that takes place in a state space with about 10^6 degrees of freedom.

D. Data from the planet simulator model

In this work, we will use a very long 8000-year dataset obtained as the output of the PlaSim climate model. In this section we briefly describe the model and its specific implementation and climate for producing this dataset. We also compare it to other climate models and explain its potential limitations and why it is suited for the present study.

PlaSim climate model [10,63] solves the global dynamics of the Earth atmosphere, coupled to ocean, ice, and land surface components. Its atmosphere dynamical core solves the primitive equations for vorticity, divergence, temperature, and pressure. The governing equations are solved using a spectral method. Unresolved processes, such as radiation, interactive clouds, moist and dry convection, large-scale precipitation, boundary layer fluxes of latent and sensible heat, and vertical and horizontal diffusion, are parametrized. The land component of the model deals with the dynamics of soil moisture, which is a key physical component of the land-atmosphere feedbacks, as long as heatwaves are concerned. It is modeled by a single-layer bucket model [82]. Soil water is replenished by precipitation and snow melt and is depleted by the surface evaporation. Soil water is limited by a field capacity with prescribed geographical distribution. If the field capacity is exceeded, then the runoff is provided to the river transport scheme.

For computing this specific dataset, the model is set up to run with fixed greenhouse gases concentrations and boundary conditions (incoming solar radiation, sea surface temperature and sea ice cover distributions) cyclically repeated every year, to generate a stationary state reproducing a climate close to the one of the 1990s [10,63]. For instance, the sea surface temperature is seasonally varying along the year, a cycle which is repeated each year. The horizontal resolution is T42 in spectral space, corresponding to a spatial resolution of about 2.8 degrees in both latitude and longitude. In practice, the horizontal fields of data have a spatial size of 64×128 pixels, covering the entire globe. The vertical resolution corresponds to 10 vertical layers. Moreover, each field is

sampled in time at $\delta t = 3$ h sampling period, and daily averages are taken in the analysis stage. The 8000-year dataset is obtained by 80 runs with independent initial conditions, each 100 years long.

We have already used a similar setup in previous works [28,55]. In the new 8000-year dataset used in this work, the setup is slightly different compared to our previous 1000-year dataset. We use a diurnal cycle, which is more realistic compared to the previous studies with diurnal variation of the solar forcing and we predict 2 m temperature rather than surface temperature, which is more relevant for impacts and a bit more difficult to predict. We note that in this paper we use daily rather than 3 h average for prediction in contrast to [55]. In an ablative study, we have trained neural networks with the 1000-year datasets and concluded that the predictive skills are similar with or without daily cycles, and independent of whether 3 h long samples or daily averages were used for the prediction.

The PlaSim model has physical parametrizations that are of a lesser quality compared to up-to-date climate models which are used for CMIP experiment, analyzed in many studies documented in the IPCC reports. Its advantage, however, is that it runs about 100 times faster and is specifically suited for producing extremely long datasets. No other statistically stationary dataset with 8000 years of fixed present-day climate simulation is available using climate models for CMIP experiments. The atmosphere dynamics component of the PlaSim model is equivalent to many of the CMIP models, although the forcings and couplings are slightly degraded. The obtained large-scale fields and patterns are of an excellent quality [10,63]. In the work in preparation, we show that the composite statistics of the large-scale 500 hPa geopotential height fields, conditioned on heatwaves, leads to very similar patterns for this PlaSim model dataset, for CESM model outputs for a 1000-year similar climate, and for the ERA reanalysis dataset [83]. CESM is one of the best models used for CMIP experiments. Because of the physical setup, for instance, the lack of an active ocean in the model simulation, given PlaSim dataset is not suited for other process studies, for instance, the impact of sea surface temperature fluctuations. For the study of other slow physical drivers than soil moisture, other climate model outputs might be needed.

For all of these reasons, and because having a very long dataset was key, this PlaSim model dataset was suited for this study. This allowed us to emphasize the methodological development and the study of the training convergence with the dataset length, in the lack of data regime.

III. PREDICTION OF PROBABILITY FOR RARE EVENTS

In this section we discuss how to make a probabilistic forecast using a neural network, for a classification task, using Softmax probabilities. We also define and discuss the scores for probabilistic forecast and their relation with the neural network score functions.

A. Softmax Regression for the inference of the probabilities of rare events

1. Setting

The task we consider is the inference of probabilities of having events $Y \in \{(0,1)\}^K$ from a set of physical observables (or features) globally called $\mathbf{X} \in \mathbb{R}^d$. If one event is in the class k, then $Y_k = 1$ and for $l \neq k$, $Y_l = 0$ (the classes are exclusive, and each event belongs to one and only one class).

As explained in Sec. II A, for the specific application in this paper, $K = 2$, $Y = 1$ when a heatwave occurs at time t and $Y = 0$ otherwise, and $\mathbf{X}$ represents all the relevant dynamical and physical predictors of the state of the atmosphere at time $t - \tau$. However, the following discussion is general for any probabilistic classification and is independent on a specific dataset or network architecture.

Let us consider the pairs $(\mathbf{X}, Y)$ as random variables, having a ground truth joint probability distribution $\mathbb{P}(\mathbf{X} = \boldsymbol{x} \text{ and } Y = y) = P(\boldsymbol{x}, y) = \mathbb{P}(Y = y|\mathbf{X} = \boldsymbol{x})\mathbb{P}(\mathbf{X} = \boldsymbol{x})$. The objective is the soft assignments of features into the event types, in the sense that we want to estimate a probability that a given realized state $\boldsymbol{x}$ will conduct to the different possible classes.

Hence, the task is the inference of the probability density $\mathbf{p} = \{p_k(\boldsymbol{x})\}_{0\leqslant k\leqslant K-1}$ where $p_k(\boldsymbol{x}) = \mathbb{P}(Y_k = 1|\mathbf{X} = \boldsymbol{x})$ is the probability that $Y_k = 1$, given that $\mathbf{X} = \boldsymbol{x}$. Because the classes are exclusive, we have $\sum_k p_k(\boldsymbol{x}) = 1$.

We classically formulate the inference of $\hat{\mathbf{p}}$ (the estimated values of $\mathbf{p}$) as a soft classification problem, relying on using a softmax function at the end of the proposed learning architecture (see the detailed architecture in Sec. IV A). Softmax probability offers a convenient way to output, for each input $\boldsymbol{x}$, an output having the meaning of the probability instead of only a class of events.

2. The Softmax probabilities

Softmax parametrization is a way to output probabilities associated with a discrete variable. If taking directly the features that pass through a single (not hidden) layer as input, then it is equivalent to what is known as logistic regression. For a detailed explanation, please see Ref. [84] for a review for physicist, or Ref. [85] for a textbook in machine learning. The output probability has to be a positive function that should sum up to 1 over all the K classes. To force that, in logistic regression, the probability associated with the features $\boldsymbol{x}$ is modeled by first computing a vector in $\mathbb{R}^K$ written as $\mathbf{o}(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbf{W}\boldsymbol{x} + \mathbf{b}$, where $\boldsymbol{\theta} = (\mathbf{W}, \mathbf{b})$, and then taking the normalized exponential (also known as Softmax function) of this vector to model the probabilities:

$$P(Y_k = 1 \mid \boldsymbol{x}, \boldsymbol{\theta}) = \frac{e^{-\mathbf{o}_k(\boldsymbol{x},\boldsymbol{\theta})}}{\sum_{j=0}^{K-1} e^{-\mathbf{o}_j(\boldsymbol{x},\boldsymbol{\theta})}}. \tag{2}$$

The quantity $\mathbf{o}$, which is the nonnormalized log probability, is also called the logit. Here, $\mathbf{o}_k$ is the output associated with the discrete variable Y_k, and in this situation of logistic regression, it would be simply the linear form $\mathbf{o}_k = \mathbf{w}_k\boldsymbol{x} + \mathbf{b}$. More generally when using neural networks (see later in Sec. IV A), the logit is a nonlinear function $\mathbf{o}(\boldsymbol{x}, \boldsymbol{\theta})$, involving several layers with parameters $\boldsymbol{\theta}$), where the original features are input of the first layer. In that situation, the softmax function is the last layer of the neural network.

This softmax parametrization achieves probability regression, as will be described here. Note that it can be used also for classification if a threshold for probability counting as a positive event (heatwave) is chosen; that was the purpose of the previous work on the prediction of heatwaves in Ref. [55], and the results were only discussed in terms of categorical prediction, with a focus on recall (also called sensitivity in the binary case, i.e., the true-positive rate) and false-positive rate (1 minus the specificity, i.e., the fraction of false positive among all true negative events) of events. Here, we will study with more details the obtained probabilities on the different classes. A discussion in terms of TP and FP rates, and of the commonly used Matthews correlation coefficient (MCC) as a metrics to combine the two [86,87], as done in Ref. [55], while useful and sometimes providing qualitatively similar results, it is not properly adapted to quantify the skill of the inference of probabilities. For this purpose probabilistic score is requested.

B. Proper probabilistic score, learning loss function, and normalized logarithmic score

In the meteorology and climate fields, a huge literature has been devoted to the definition of good scores for probabilistic forecast validation. A good probabilistic score should be additive with respect to new events, proper (it should be maximum when the forecasted probability is the ground truth probability), and should not depend on unobserved events [88]. In the case of probabilistic classification, of interest in this paper, we follow the analysis of Benedetti [58]. It concludes that the only probabilistic score with these three natural properties is the logarithmic score, also often referred to as the ignorance score (see Ref. [89] and references therein). We will thus use the logarithmic score to validate the forecast skills of the neural network.

The aim of this section is to define the logarithmic score, to explain that it is nothing else than the negative of the cross-entropy loss function minimized by the neural network during the learning stage, and to define a normalized logarithmic score. The NLS is just a simple affine transformation

of the logarithmic score that has the property of being equal to 0 when the forecasted probability is the climatological frequency, and equal to 1 when the prediction is perfect.

We consider N actually observed events $Y_{(n)}$, with $1 \leqslant n \leqslant N$, associated with the observed features $\mathbf{X}_n$. We suppose that the event $Y_{(n)}$ is observed pertaining to the class k_n. This means that $Y_{(n),k_n} = 1$, and $Y_{(n),l} = 0$, for $l \neq k_n$. The couples $(\mathbf{X}_n, Y_{(n)})$ are identically distributed, and the probability that $Y_{(n);k_n} = 1$ given that $\mathbf{X}_n = \boldsymbol{x}$ is $p_k(\boldsymbol{x}) = \mathbb{P}(Y_k = 1|\mathbf{X} = \boldsymbol{x})$.

The real probabilities $p_k(\boldsymbol{x})$ are unknown. We consider a probabilistic forecast $\hat{p}_k(\boldsymbol{x})$. In our case, $\hat{p}_k(\boldsymbol{x})$ will be the output of the neural network after the learning stage. Our aim is to give a score that quantifies the quality of the approximation of p by $\hat{p}$. This score should be computed without the actual knowledge of p, and be based only on the N observations, for instance, N samples of a test dataset. The logarithmic score [58] is

$$S_N(\hat{\mathbf{p}}) = -\frac{1}{N}\sum_{n=1}^{N} \log\left[\hat{p}_{k_n}(\boldsymbol{x}_n)\right]. \tag{3}$$

We note that in the simple case considered in Ref. [58], the probabilities do not depend on the state of the system $\boldsymbol{x}$, while they do in this paper. However, all the reasoning and conclusions in Ref. [58] easily generalize to this new case. We note that with this sign convention, the minus sign in front of the logarithm, the logarithmic score is positive ($S_N > 0$) and negatively oriented (the smaller the score, the better the prediction).

We see that the logarithmic score is nothing more than the empirical cross-entropy loss function:

$$C[\hat{\mathbf{p}}] = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{K-1} \delta_{k_n,k} \log \hat{p}_k(\boldsymbol{x}_n), \tag{4}$$

where $\delta_{k_n,k}$ is the Kronecker δ. This is the loss function minimized during the learning stage of the neural network.

It is easy to check that the score is proper, by noting that according to the law of large numbers $\lim_{N\to\infty} S_N = \mathbb{E}[S_N] = L[\hat{\mathbf{p}}, \mathbf{p}]$, with

$$L[\hat{\mathbf{p}}, \mathbf{p}] = -\int d\boldsymbol{x}\, P(\boldsymbol{x}) \sum_{k=0}^{K-1} p_k(\boldsymbol{x}) \log\left[\hat{p}_k(\boldsymbol{x})\right], \tag{5}$$

and that the minimum of $L[\hat{\mathbf{p}}, \mathbf{p}]$ is obtained for $\hat{\mathbf{p}} = \mathbf{p}$.

We note that $L[\mathbf{p}, \mathbf{p}] \geqslant 0$ and that $L[\mathbf{p}, \mathbf{p}] = 0$ only in the case when for any $\boldsymbol{x}$, all the $p_k(\boldsymbol{x})$ are equal to zero except one. This is the case of a deterministic relation between $\boldsymbol{x}$ and y, when a perfect prediction is possible. In general, when the relation between $\boldsymbol{x}$ and y is stochastic, $L[\mathbf{p}, \mathbf{p}] > 0$, and it measures the level of stochasticity between $\boldsymbol{x}$ and y.

It is important to compare the obtained score with the score of a prediction based on the climatological frequency. We define the climatological frequency as the probability $\overline{p}_k$ of observing the class k, independently of the knowledge of the state of the system $\boldsymbol{x}$. The climatological forecast $\hat{\mathbf{p}} = \overline{\mathbf{p}}$ provides a baseline: any skillful forecast should be better than the climatological one. We note that $\mathbb{E}[S_N(\overline{\mathbf{p}})] = L[\overline{\mathbf{p}}, \overline{\mathbf{p}}] = -\sum_k \overline{p}_k \log \overline{p}_k$ [we have used (5), noting that the $\overline{p}_k$s do not depend on $\boldsymbol{x}$].

Generalizing the discussion in Ref. [58] to the present case, we define the *normalized logarithmic score* as

$$\mathrm{NLS}(\hat{\mathbf{p}}) = \frac{-\sum_k \overline{p}_k \log \overline{p}_k - S_N(\hat{\mathbf{p}})}{-\sum_k \overline{p}_k \log \overline{p}_k}. \tag{6}$$

We clearly see that for the climatological forecast $\mathbb{E}[\mathrm{NLS}(\overline{\mathbf{p}})] = 0$. As $-\sum_k \overline{p}_k > 0$, we see that the score is positively oriented (the larger score, the better prediction).

For a given p, the optimal value of $\mathbb{E}[\mathrm{NLS}(\hat{\mathbf{p}})]$ is $(-\sum_k \overline{p}_k \log \overline{p}_k - L[\mathbf{p}, \mathbf{p}])/(-\sum_k \overline{p}_k \log \overline{p}_k) \leqslant 1$. This optimal value is unknown, except if one would know $\mathbf{p}(\boldsymbol{x})$ and $P(\boldsymbol{x})$. Only in the case of a deterministic relation between $\boldsymbol{x}$ and y, when a perfect prediction is possible, this optimal value is equal to 1, otherwise it is strictly smaller than one.

The normalized logarithmic score is a way to quantify the predictive skill of predictive power of the models learned by the convolutional neural networks. We conclude that the *normalized logarithmic score* is positively oriented (the larger the better), its average is equal to zero for the climatological forecast, and is always smaller than 1. The unknown optimal value is strictly smaller than 1, except when a deterministic relation between the predictor $\boldsymbol{x}$ and the predicted class y exists and a perfect prediction is possible. These properties make it convenient.

In the climate and meteorology literature, other scores for probabilistic forecasts are commonly used, for instance, the Brier Score [89,90]. The Brier score can be very useful. It however depends on unobserved events (see Ref. [58]). Moreover, the relation between the logarithmic score and the cross-entropy loss function makes the learning and the validation stage fully compatible. The logarithmic score also appears to be more sensitive toward the measurements of small probabilities or probabilities close to one, and is thus better suited for the study of rare events. The information theoretic interpretation of the logarithmic score is also an appealing property.

IV. NEURAL NETWORK ARCHITECTURE AND LEARNING PROTOCOL FOR THE PREDICTION OF RARE EVENT PROBABILITIES

As explained in Sec. II, the task is to predict the probability of occurrence of a heatwave at time t, from the state of the system $\boldsymbol{x}$. These variables are sets of physical fields which are observed at time $t - \tau$. In this section, we present the neural network architecture, the training parameters and protocols.

A. Neural network architecture and learning parameters

1. Neural network architecture

For the inference of probabilities, we will use a CNN architecture as sketched in Fig. 4 [85].

It consists of a three-layer architecture, combining convolutional filters, followed by ReLu activations. For all layers, 3×3 convolution kernels are used, while the number of filters is 32 for the first and 64 for the two last ones. Two max-pool operations are inserted between the convolution layers after the activation functions. The output of the third layer is flattened and used as input of a dense layer with 64 neurons, a layer with two outputs corresponding to the heatwave labels, and a softmax function which maps the outputs to (0,1) range, as detailed for the probabilistic interpretation in Sec. III.

The probabilities obtained by softmax regression with cross-entropy loss function are not always well calibrated. For instance, it has been discussed in Ref. [91] in very deep networks (e.g., ResNET) that the calibration of the probabilities is not correct; in other words the network may be overly confident about its probabilistic predictive capability. This is a reason why we prefer to use a neural network that is not too deep. Also, this phenomenon can be worse when facing extremely rare events because of the imbalance between the classes. We discuss in Sec. IV C how to avoid biases due to overfitting.

2. Loss function

As discussed in Sec. III, we minimize the cross-entropy loss function, Eq. (4). Optimizing the cross-entropy is done as a supervised task, using both the input data fields $\boldsymbol{x}$ and heatwave classes Y.

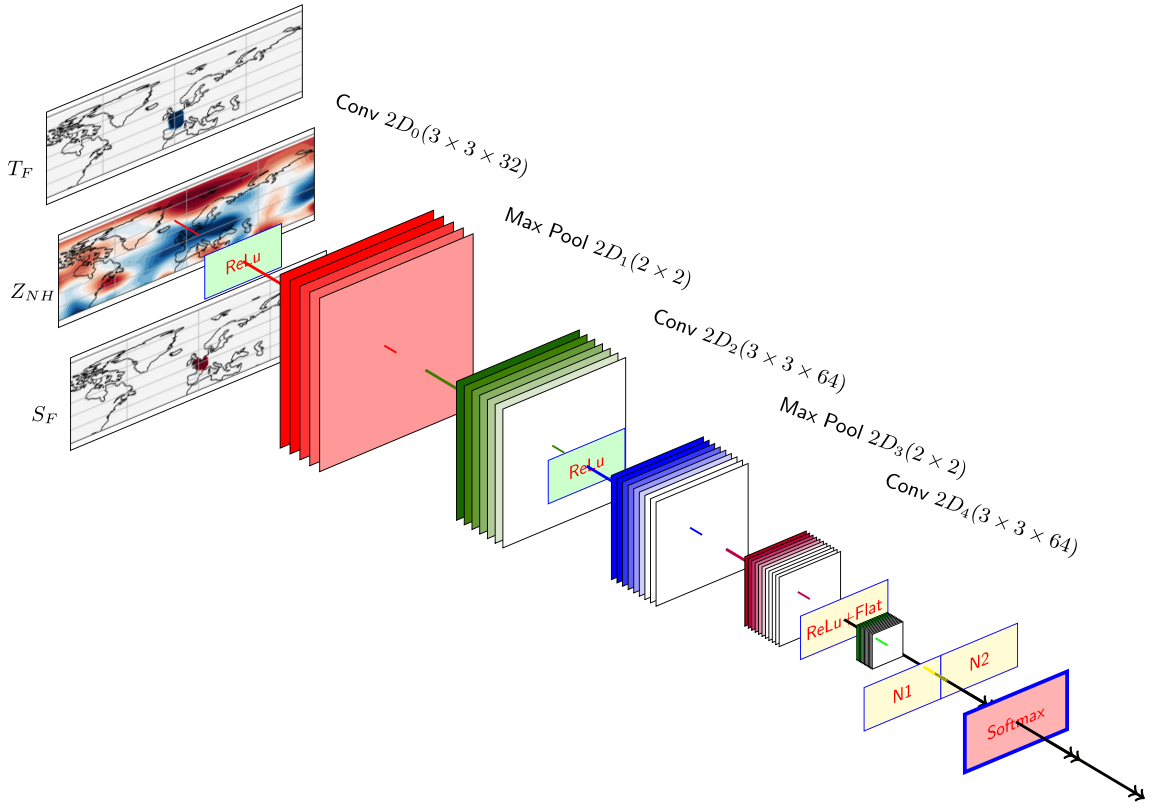


FIG. 4. Schematics of the CNN architecture. The numbers $a \times b \times c$ inside the boxes for Conv2D indicate the kernel sizes (a by b) and the number of filters c in the convolutional layers. The box "Relu" actually consists of a sequence: batch normalization-activation (ReLu), and spatial dropout. Relu stands for rectified linear unit, it provides nonlinearity and is typically used to prevent vanishing gradients instead of sigmoid for the hidden layers.

3. Learning tools and parameters

The CNN layers are implemented using Tensorflow 2 package, and CNN training is done with Adam optimizer, with learning rate set to 2×10^{-4}. Network weights are initialized using a standard Glorot distribution. The computer resources consisted of computers with dedicated graphics cards such as GV100GL [Tesla V100 PCIe 16GB], TU102 [RTX 2080 Ti Rev. A], and TU104 [RTX 2080 SUPER].

B. Training protocol

1. Data normalization

The training set consists of 8000 independent and statistically equivalent years of simulated climate (cf. Sec. II D). The season of interest is June-July-August (JJA, 90 days). When we consider 14-day time average, we have 77 days of JJA per year before the start of the heatwave (see Sec. II A). This gives 616 000 snapshots for each value of τ.

Data are normalized by grid point prior to application to the neural network: we add a constant and scale each cell of each field such that the sample mean and variance are 0 and 1, respectively.

Note that each possible input field, Z, S, or T is considered as a separate input channel and they are stacked in the CNN layers (e.g., like RGB channels in colors images). The input tensor is then of size $22 \times 128 \times$ the number of input fields (3 if using all of Z, S, and T) provided that the global field corresponding to the north hemisphere above 30N is used. In this case we use the notation

Z_{NH}, T_{NH}, S_{NH}. However, if the input consists of a smaller area corresponding to the North Atlantic region and Europe, then we use the notation Z_{NAE}, T_{NAE}, S_{NAE}, which has dimensions of 18×42 (Fig. 2). Note that in both cases we could be interested in applying additional mask to the area of France, i.e., setting to zero all values external to the area of a box around France. This operation is not applied to geopotential, so the two resulting cases are Z_{NH}, T_F, S_F and Z_{NAE}, T_F, S_F.

Sec. V will discuss in details which combination of these z, s and/or T fields, used either locally or globally give the best predictions.

2. Stratified 10-fold cross-validation

To quantify confidence in achieved prediction performance, we use a classical stratified k-fold cross-validation procedure [92]. The 8000 available years are randomly split into $k = 10$ subsets. The splitting of the data set is performed on a per-year basis to avoid that any of the years is split into a test and train set. The latter would blur validation consistency by spurious temporal correlations between validation and train sets or seasonal effects. In addition, random splitting is performed to maintain the same number of heatwaves per subset (stratification). Each resulting subset consists of 800 years.

3. Initialization, batch normalization, and dropout

The convolutional and ReLu activation layers are followed by batch normalization and subjected to a dropout. Batch normalization is expected to accelerate learning. Dropout is considered a regularization tool avoiding overfitting [91]. The dropout rate is set to 0.25.

C. Early stopping

Overfitting is a major problem of machine learning, especially in the case of deep neural networks. This means that the model is trained to reproduce the training dataset too closely and does not generalize well to the validation set or the test set. As a remedy an early-stopping strategy is typically used, which implies stopping the training at an epoch when the appropriate metric on the validation set starts getting worse. We have followed the same general outline, with a caveat. Since we are performing 10-fold cross validation, we can rely on the performance metric as provided by the NLS, Eq. (6). Each training set (fold) reaches its optimal performance on a validation set at a certain epoch, after which it starts to over-fit. This epoch generally depends on the training set although in most cases the variance is not vary large.

Nevertheless, we propose to select an epoch on which the fold-wise mean score is optimal. We refer to this as collective early stopping. This approach is more conservative about the performance of the network, and less dependent on the training/validation split.

D. Unbalanced classes and undersampling strategy for probabilistic forecast

Rare event prediction suffers by design from a severe class imbalance. To address this, we use a majority class undersampling strategy, as opposed to a minority class oversampling. This means that the training set uses all available positive events, but only a ratio $1/r$ of events of the majority events are drawn with uniform probability, with $r > 1$ (undersampling). This data reduction procedure also minimizes time and memory costs during training.

This undersampling changes the rate of positive events. We need to take account of this change of measure, otherwise the predicted probability will not be correct [61,62].

Let $p_0(x)$ and $p_1(x) = 1 - p_0(x)$ denote the probabilities that $Y = 0$ and $Y = 1$, respectively, given that $X = x$, in the original set. Let $p'_0(x)$ and $p'_1(x) = 1 - p'_0(x)$ denote the probabilities, that $Y = 0$ and $Y = 1$, respectively, given $X = x$, in the undersampled training set. These probabilities

are obviously related as

$$p'_0(x) = \frac{p_0(x)}{p_0(x) + r[1 - p_0(x)]} \quad \text{and} \quad p'_1(x) = \frac{r p_1(x)}{1 - p_1(x) + r p_1(x)}. \tag{7}$$

As an example, when $p_0 = 0.8$ and $p_1 = 0.2$ and an undersampling ratio $r = 4$ is used, fully balanced undersampled classes are obtained, with $p'_0 = 0.5$ and $p'_1 = 0.5$. During training after undersampling, the neural network actually gives an estimate $\hat{p}'_0(\boldsymbol{x})$ and $\hat{p}'_1(\boldsymbol{x})$ of the probabilities $p'_0(\boldsymbol{x})$ and $p'_1(\boldsymbol{x})$ of the event that has been seen, and not of the true ones $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x})$.

To get an estimate $\hat{p}_0(\boldsymbol{x})$ and $\hat{p}_1(\boldsymbol{x})$ of $p_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x})$, respectively, we need to invert the relation (7) between the initial probabilities and the probabilities in the undersampled set. This gives

$$\hat{p}_0(x) = \frac{r \hat{p}'_0(x)}{1 - (1 - r)\hat{p}'_0(x)} \quad \text{and} \quad \hat{p}_1(x) = \frac{\hat{p}'_1(x)}{r + (1 - r)\hat{p}'_1(x)}. \tag{8}$$

The estimated probabilities $p'_0(\boldsymbol{x})$ and $p'_1(\boldsymbol{x})$ can then be tested using Eq. (6) on a validation set or a test set, or used as the predicted probabilities for the physical discussion and for applications. Please note that we do not undersample the validation set.

In this work, we use an undersampling rate $r = 10$, consistent with [55]. This reduces the RAM memory usage approximately 10-fold and accelerates the training while not impacting the skill significantly (see Fig. 11 in Sec. V E 1).

V. PROBABILISTIC FORECAST OF EXTREME HEATWAVES

The present section aims to quantify, using the NLS, Eq. (6), the quality of the prediction of heatwave occurrence probabilities. These quantifications will be conducted as function of the lag τ between time at which data is available for prediction and heatwave occurrence. The impact of the nature of the data used for prediction (soil moisture, geopotential, and/or 2 m temperature) will be investigated together with the benefits of possible combinations of such inputs. Further, the impact of the amount of available data on prediction performed will be studied.

A. Relevant climate fields for the probabilistic forecast of extreme heatwaves

A first key question is to assess which of the predictors, among the physical and dynamical fields, have the best prediction capabilities. To this aim, we train the neural network with the large 8000-year dataset and test its skill by computing the normalized logarithmic score [cf. Eq. (6)] on a validation set. From the values of NLS, we can compare the prediction skills for extreme heatwaves, when different combinations of the fields are used: soil moisture S, geopotential height at 500 hPa Z, and 2 m temperature $T_{2\mathrm{m}}$, used either alone or combined.

In this section we always use the geopotential height at 500-hPa Z_{NH} over the northern hemisphere area, and the soil moisture S_{F} and 2 m temperature T_{F} over the France area. These choices of regional masking will prove to be the optimal ones, as we will discuss in Sec. V B. We also consider the 2 m temperature integrated over the France area, which is then a single real number denoted T_{FI}. When only one local field (FI) is used, a simple scalar logistic regression is performed. For all the other cases, we train the neural network as explained in Sec. II B.

Figure 5 reports the normalized logarithmic scores versus the lead time τ, for different combinations of fields.

1. Single-field prediction

Figure 5(a) first shows that soil moisture over France, S_{F}, conveys significant long-term ability for the prediction of heatwave probabilities. In particular, it retains predictive value at large lead time τ, hence can be considered a candidate slow physical driver. Those findings are consistent with the important role of soil moisture through its two-way coupling with heatwaves as discussed in Sec. II B. The main physical interpretation is that soil moisture deficit statistically increases the

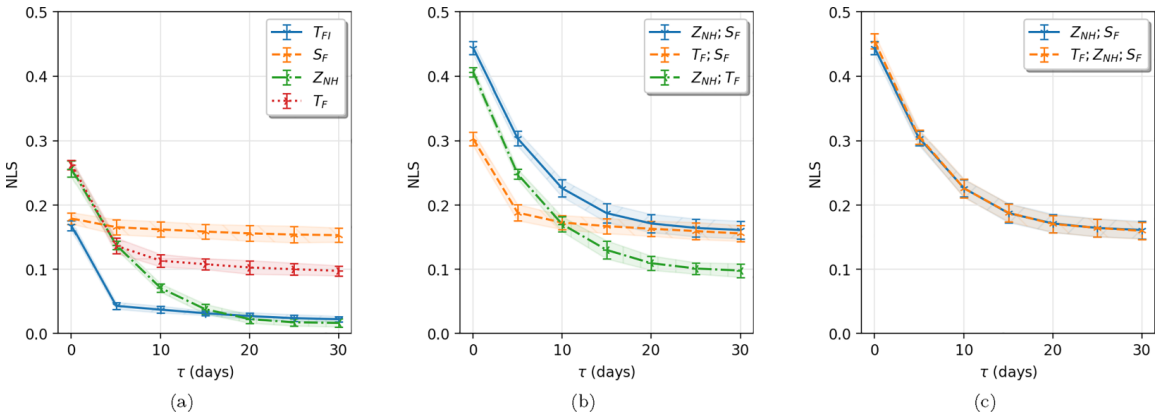


FIG. 5. Relevant climate fields for best prediction. All figures show the NLS [cf. Eq. (6)] vs the lead time τ, for different combinations of the predictor fields. Fields are either integrated over the area of France (FI), masked over the France area (F) or over the northern hemisphere (NH). From k-fold cross-validation values, we plot the averaged NLS, plus or minus one standard deviation (shaded area and error bars): (a) single-field prediction: T_{FI} (blue), S_{F} (orange), Z_{NH} (green), and T_{F} (red); (b) fields combined pairwise for prediction: (Z_{NH}, S_{F}) (blue), (T_{F}, S_{F}) (orange), (Z_{NH}, T_{F}) (green); (c) all three fields combined for prediction: (T_{F}, Z_{NH}, S_{F}) (orange).

temperature in the lower troposphere, through the impact of deficit of evapotranspiration or other water exchanges at the surface on the lower troposphere energy cycle, and on cloud cover. We also note that the predictive skill of soil moisture alone only weakly decays with lead time τ, which can be interpreted as a consequence of the long correlation time of soil moisture compared to the maximum 30 days lag-times considered here, soil moisture being a stock. The soil moisture predictability skill is nearly a constant versus τ. This constant could be directly related to the conditional probability to have a heatwave given some soil moisture field S_{F}, and could be estimated in a straightforward fashion based on a climatological prediction conditioned on the values of the soil moisture fields, regardless of any other information about dynamics.

The curve Z_{NH} in Fig. 5(a) shows that when using only the northern hemisphere 500 hPa geopotential height for training, the neural network has better prediction skill for short times. This skills decays roughly exponentially with τ with the approximate rate of decay of 0.13 per day: $\mathrm{NLS}_{z_G} \approx 0.26 \exp(-0.13\,\tau)$. This rate corresponds to a decay time of 7.7 days (after 7.7 days, the prediction skill decreased by a factor e). The 500 hPa geopotential height field is considered as one of physical fields which characterizes the best midlatitude troposphere dynamics, Rossby waves, cyclonic and anticyclonic anomalies. The decay of the skill is interpreted as the progressive loss of memory for the evolution of this dynamical field, due to the chaotic dynamics of the midlatitude troposphere, over timescales of the order of the synoptic timescale, which corresponds to few days.

By comparing the NLS from Z_{NH} and S_{F}, we confirm that in the short run soil moisture has significantly smaller predictive value than the 500 hPa geopotential height, but, in contrast, it keeps its predictive skill over much longer time lags τ, as expected based on the discussion in Sec. II B. However, such qualitative statements are now precisely quantified, thanks to the neural network we introduced here. We see, for instance, that for these PlaSim simulations, the geopotential height Z_{NH} alone has a *NLS* of 0.26 ± 0.01 compared to 0.18 ± 0.01 for soil moisture S_{F} at $\tau = 0$, and that the predictive skill of soil moisture alone becomes larger than the one of 500 hPa geopotential height for τ larger than about 4 days.

The interest of looking at the predictive power of the temperature integrated over France, T_{FI}, is to assess the predictability properties related to persistence and possibly low tropospheric advection. T_{FI} on Fig. 5(a) shows this prediction for $\tau = 0$. Although not visible on the Figure, the related skill extends for τ of order of a few days, subsequently the predictability power is lost. For larger values of τ, $\tau \geqslant 5$ days, T_{FI} reaches a weakly decreasing plateau.

The information contained in T_{FI} is the spatial average of the temperature field over France T_{F}. As a consequence, T_{F} on Fig. 5(a) has a higher NLS. The fact that the predictive skill for T_{F} is much better than the one for T_{FI} shows that the details of the spatial pattern of temperatures over France matter much for the prediction of extreme heatwaves defined globally over France. This is a very interesting result. Anticipating the following discussion, we will interpret the predictive skills of both T_{F} and T_{FI}, for $\tau \geqslant 5$, to be due to their mutual information with the soil moisture. We can then interpret the better predictive skill of the spatial field T_{F}, compared to the one of T_{FI}, as the consequence of the larger information content of the soil moisture on some specific areas. This interpretation is consistent with past studies that argued that soil moisture matters more in areas prone to its deficits, rather than areas where soil is unlikely to dry. This interpretation should be studied further in future works.

The two plateaus with strictly positive skills for both T_{F} and T_{FI} for large time lag τ are striking. We cannot expect those skills, that extend over timescales much longer than the synoptic times, to be related to persistent properties or to free troposphere dynamics, because they extend over timescales much longer than the mixing time of the uncoupled troposphere dynamics. This long-term skills might be related to some correlations between T_{FI} and some slow physical drivers. We might hypothesize that T_{F} and T_{FI} contain statistical information related to the soil moisture. To study this hypothesis, we will now study the predictive skills of combined fields.

2. Predictions using fields combined pairwise

We now study the predictability skills of the neural network when trained using combinations of the two fields. The results, reported on Fig. 5(b), show that the best combination is the couple $(Z_{\mathrm{NH}}, S_{\mathrm{F}})$. Compared to the results on Fig. 5(a), it is striking to see that the predictive skills of Z_{NH} and S_{F} seem to add up [93]. The curve can be approximated as $0.288 * \exp(-0.144\,\tau) + 0.155$, with a decay time of about 6.9 days relatively close to the one obtained for the geopotential height alone. Using the couple $(Z_{\mathrm{NH}}, S_{\mathrm{F}})$, the neural network is able to conveniently retain the useful information for prediction, from both the fast dynamical field Z_{NH} and the slow physical driver S_{F} in a seamless way.

The predictive skill of the couple $(T_{\mathrm{F}}, S_{\mathrm{F}})$ is the worst among the three couples for small lead times τ, and is not better than the skill of the field S_{F} alone for large lead times. For large lead times, this means that all the useful predictive information lies in S_{F}. This remark supports the idea that the plateau for T_{F} in Fig. 5(a) has to be interpreted as the predictive skill for the 2 m temperature, as a consequence of its mutual information with the soil moisture. Moreover, for large lead times, clearly, the flow of information is from the soil moisture to the 2 m temperature, as combining both fields do not give improvements with respect to soil moisture alone. However, the 2 m temperature actually provides new complementary information for small lead times, most probably because of the skill associated with persistence or low-tropospheric advection.

It is interesting to note that the couple $(Z_{\mathrm{NH}}, T_{\mathrm{F}})$ performs rather well, and that the information of Z_{NH} and T_{F} seems to add up too, just like the one for Z_{NH} and S_{F} do. This is particularly striking for small lead times τ. Indeed, one might have expected that the information about the 2 m temperature might have been contained in the dynamical field Z_{NH}, for short lead times. This is however not the case, the better skill when combining the two fields clearly proves that the temperature field value contains relevant predictive information that was not included in the dynamical field Z_{NH}, even for small lead times. Where does this information come from? Is it related to slow physical drivers? To answer this question, we will need to compare the result for $(Z_{\mathrm{NH}}, S_{\mathrm{F}})$ with the neural network skill when all three fields are combined together for the training.

3. Predictions combining all three fields together

Figure 5(c) displays the prediction performance when using the three fields $(T_{\mathrm{F}}, Z_{\mathrm{NH}}, S_{\mathrm{F}})$ together, and compares it to the results obtained with the best pair $(Z_{\mathrm{NH}}, S_{\mathrm{F}})$. There is no improvement for the predictive skill when adding the 2 m temperature field T_{F} to the geopotential height and soil

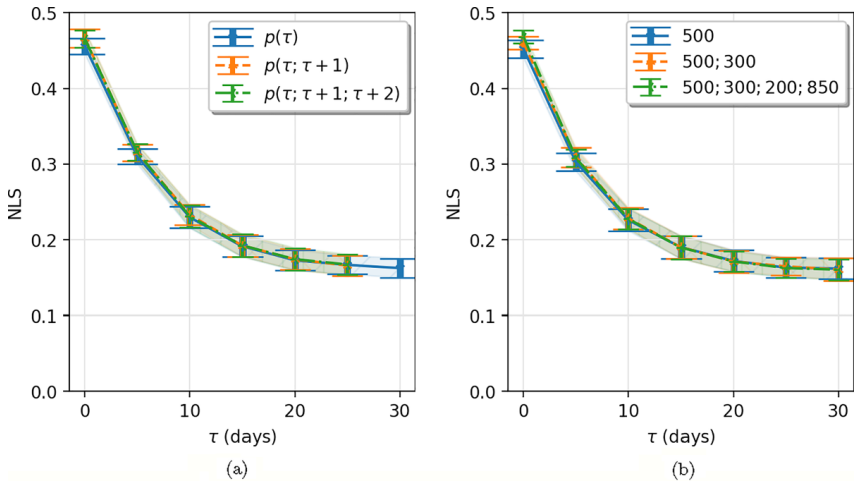


FIG. 6. Prediction skills when training the network with extra fields. The normalized logarithmic score *NLS* is on the y axis. (a) Addition of stacked fields at extra time steps: $(T_F, Z_{NH}, S_F)(t-\tau)$ (blue), $[(T_F, Z_{NH}, S_F)(t-\tau), (T_F, Z_{NH}, S_F)(t-\tau-1)]$ (orange), $[(T_F, Z_{NH}, S_F)(t-\tau), (T_F, Z_{NH}, S_F)(t-\tau-1), (T_F, Z_{NH}, S_F)(t-\tau-2)]$ (green). The NLS features a very small improvement when adding the previous day information, but none if we add two previous days. (b) Addition of extra levels of geopotential height $[T_F, Z_{NH}\,(500\text{ hPa}), S_F]$ (blue), $[T_F, Z_{NH}\,(500\text{ hPa}), Z_{NH}\,(300\text{ hPa}), S_F]$ $[T_F, Z_{NH}\,(850\text{ mbar}), Z_{NH}\,(500\text{ hPa}), Z_{NH}\,(300\text{ hPa}), Z_{NH}\,(200\text{ hPa}), S_F]$ (green). The NLS features a very small improvement, although not statistically significant, when adding more geopotential height fields.

moisture fields (Z_{NH}, S_F), except maybe for very short lead time $\tau < 5$. The improvement for very short lead time is not statistically significant as it is within error bars. If this was not the case, then we could interpret it as the effect of properties of persistence or of low-tropospheric advection of the 2 m temperature field over France. This lack of improvement means that all useful information for prediction in the 2 m field, is actually already contained in the (Z_{NH}, S_F) fields.

Our first conclusion is that the best prediction is obtained when the neural network is trained using the combined information from the northern hemisphere 500 hPa geopotential height field and the soil moisture over the France area. The neural network is able to seamlessly combine the information of the fast dynamical driver, the 500 hPa geopotential height field, and the slow physical one, the soil moisture. The temperature field over France does not seem to convey complementary information to these two fields, except perhaps at $\tau = 0$. But even for $\tau = 0$ the improvement is not statistically significant given the dataset: 0.455 ± 0.012 versus 0.445 ± 0.010.

It is customary in other prediction studies for extreme heatwaves to use the local 2 m temperature field. Given what we have observed this makes sense when the information about soil moisture is not available.

4. Is it useful to consider more predictor fields?

We now ask wether it might be useful to consider more predictor fields. We will train the neural network, first using the same fields but observed at more than one timestep, and second considering the value of the geopotential height on other pressure isosurfaces. The two sets of results are visible on Fig. 6.

We first train the neural network with the optimal set of fields (T_F, Z_{NH}, S_F), as in Fig. 5(c). But during the training stage, rather than using the field values only at lead time τ (at time $t-\tau$) we also use the field values at lead time $\tau+1$ (previous day, at time $t-\tau-1$) and $\tau+2$ (second previous day). Those previous day fields are stacked with the fields at lead time τ. Figure 6(a) shows the skill of the trained network adding the previous day fields (in orange), or the two previous day fields (in green). Adding the fields at previous timesteps is similar to delay-embedding in dynamical system

theory [94]: the information lost in taking only part of the initial conditions of the deterministic dynamics can be recovered using fields at previous timesteps, in principle. The question we address here is more a practical one: can a given neural network learn this missing information from the fields at previous timesteps, given the dataset length and its other practical limitations.

The result in Fig. 6(a) shows a small statistically insignificant improvement when one adds the field values at lead time $\tau + 1$, but no further improvement when one adds both the field values at lead time $\tau + 1$ and $\tau + 2$. This is a very interesting result. One can interpret this incapacity of given neural network to use the information at previous lead times in three different ways. The first possible interpretation, intrinsic to heatwave dynamics, would be that the gain in information content in the fields at previous lead times, to predict extreme heatwaves, is actually very small and within the error bars of our experiments. The second possible interpretation, practical in nature, would be that we have not found a network structure that could reliably recover this information. The third interpretation, would be that the 8000-year-long dataset is too small for the neural network to practically learn such detailed information. Although we cannot support precisely this claim with the present dataset, the analysis in the next section, of a lack of data regime, makes the third interpretation plausible.

Rather than complementing the predictor fields with the ones at previous lead times, we now add other relevant dynamical fields at the same lead time τ. Atmospheric and climate scientists know that the 500-hPa geopotential height field is rather relevant for dynamics, but that geopotential height at other altitudes or pressure isosurfaces are also useful and provide complementary information, for many phenomena. We now train the neural network with several sets of these fields, in addition to the optimal set of fields (T_{F}, Z_{NH}, S_{F}). The obtained skills are shown in Fig. 6(b).

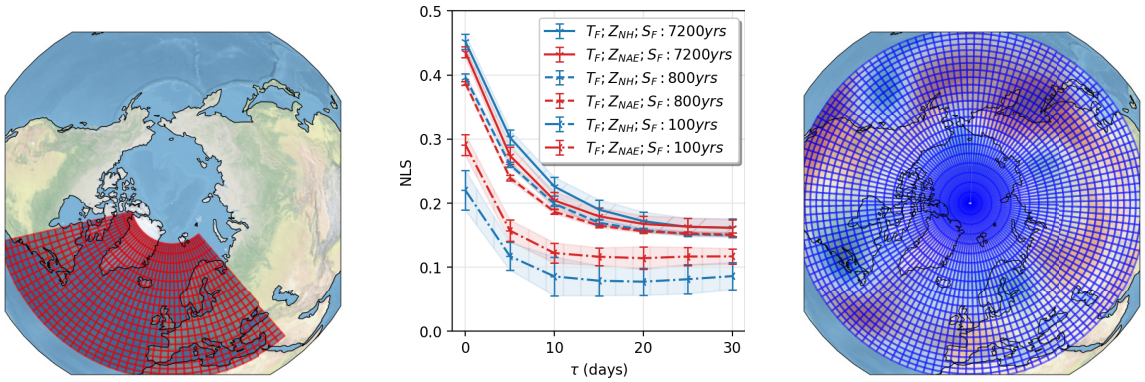
The conclusion is that adding the geopotential height at 300 hPa (upper troposphere), orange curve, slightly improves the network prediction skill compared to the reference blue curve. However, this improvement is marginal, visible only for $\tau = 0$, and even for $\tau = 0$ it is within the error bar and thus not statistically significant. Similarly, we observe a minute increase in the normalized logarithmic score when adding further the geopotential height at 850 hPa (lower troposphere), green curve. As for the case of delay embedding, the incapacity to improve the neural network prediction by adding more fields can be interpreted as being either intrinsic, or due to improper network architecture, or due to a lack of data for training. We suppose that the lack of data is the most plausible explanation.

Our second conclusion is that the set (T_{F}, Z_{NH}, S_{F}) or (Z_{NH}, S_{F}) are the optimal ones, with marginal difference in their predictive skills, for a dataset length of 8000 years. The addition of any other extra fields including different lag-times prove to be superfluous.

B. Convergence of prediction skills with training dataset length and optimal areas for predictors: A regime of lack of data

The reanalysis datasets [83] assimilate all available observations with the laws of physics embedded in the weather models and thus offer the most precise available approximation of the real state of the atmosphere. They are, however, only available during the last 70 years, at most. Is such a short dataset long enough to make reliable prediction using neural networks? One of the key goals of this work is to understand the effect of dataset lengths on the probabilistic predictions that can be issued by neural networks. This is an important question, because many practical applications of neural networks in atmosphere and climate sciences currently use reanalysis datasets for both training and validation.

Within our PlaSim model, we now study the effect of reduction of the training set on the prediction skill. We use the prediction skill with a neural network trained on 7200 years of data presented in Sec. V A as a benchmark. From the PlaSim dataset, we extract two training subsets of shorter year span: 100 years and 800 years. For both cases, we estimate the skill on a validation set that contains the complement of the full 8000-year-long dataset [95]. For these experiments,



(a) (b) (c)

FIG. 7. Prediction skills versus dataset lengths and optimal geographical area. Panels (a) and (c) show PlaSim grid for the North Atlantic and Europe sector (NAE, a), and northern hemisphere mid and high latitude sector (NH, c), respectively. (b) Normalized logarithmic score versus lead time τ, for neural networks trained with (T_F, Z_{NH}, S_F) predictors (blue) and (T_F, Z_{NAE}, S_F) predictors (red), and with datasets of length 7200 years (plain lines), 800 years (dashed lines) and 100 years (dashed-dotted lines). The results illustrate the lack of data regime, with very slow convergence of the prediction skill with the dataset length, and with a clear tradeoff between dataset length and size of optimal geographical area for best prediction.

we chose the predictors (T_F, Z_{NH}, S_F) which have been proven optimal in Sec. V A, when using a 7200-year training dataset.

The results are shown by the blue curves on Fig. 7(b). The conclusion is that reducing the dataset up to 100 years has severe consequences for the prediction skill, with a NLS [see Eq. (6)] nearly halved compared to the benchmark obtained with a 7200-year training set. We stress that for the 100-year training set, even the plateau skill, corresponding to the effect of soil moisture only, is not correctly predicted. When using a 800-year training dataset, the prediction skill is still quite significantly lower than when using a 7200-year one. However, the difference of NLS is now of the order of about 10% at most. This suggests that the convergence of the skill with the dataset length probably occurs on the order of a few thousands to a few tens of thousands of years, if one uses only the three predictor fields (T_F, Z_{NH}, S_F). We thus conclude that as long as the source for training of neural network contains only few centuries or even millennia this results in the regime of lack of data, which consequently implies a regime of drastic lack of data when using reanalysis datasets.

In such a regime it is customary for machine learning applications that there exists a tradeoff between the dataset length and the complexity of the predictors. Indeed, the amount of requested data for optimal training tends to increase when more features are included, in other words greater variety of predictors may lead to overfitting. We now study this tradeoff, as another and complementary manifestation of the regime of lack of data for neural networks applied to extreme heatwaves.

To this end we train a neural network with the predictor set (T_F, Z_{NAE}, S_F), where the 500 hPa geopotential height information is used only on the North Atlantic and European area: Z_{NAE}. We will compare its skill to the benchmark one (T_F, Z_{NH}, S_F) that uses the 500 hPa geopotential height on the whole northern hemisphere mid and high latitude. Dynamically, the information on the North Atlantic and European sector is more important for France heatwave than the information on the rest of the northern hemisphere (see, for instance, Refs. [71,72]). However, we have recently demonstrated that extreme heatwaves are associated with hemispheric teleconnection patterns [28]. It is thus likely that the rest of the northern hemisphere should contain useful complementary information which might be more difficult to learn. We then expect that if we have sufficiently long training datasets, the neural network should have a better skill with the complete field (T_F, Z_{NH}, S_F). Otherwise, the predictor will turn out to be too complex which would result in the degradation of the normalized logarithmic score.

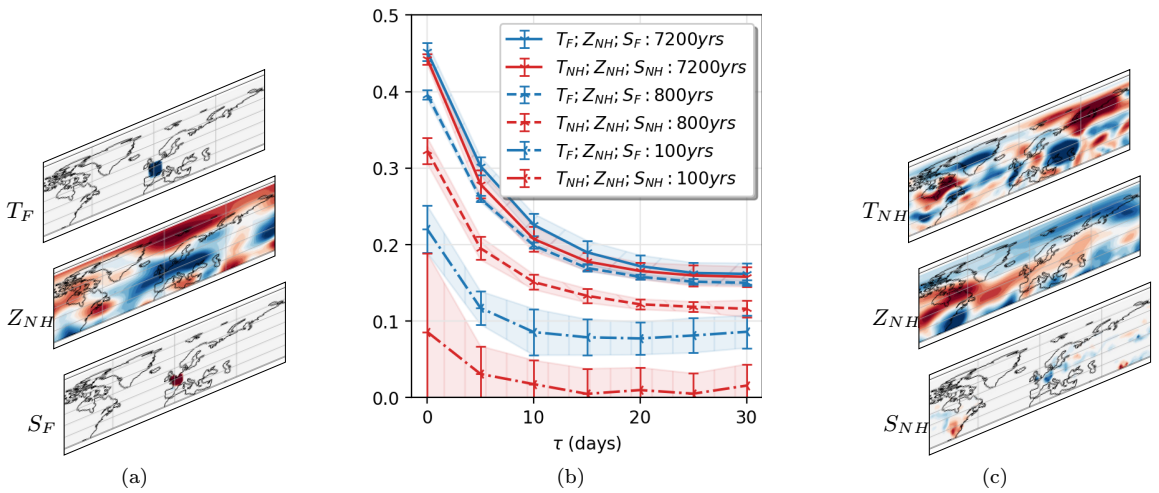


FIG. 8. Prediction skills versus dataset lengths and optimal geographical area for the 2 m temperature and soil moisture. Panels (a) and (c) show typical predictor fields, either masked over some restricted area [panel (a)], or over the whole mid and high latitude northern hemisphere [panel (c)]. Panel (b) Normalized logarithmic score versus lead time τ, for neural networks trained with ($T_{\rm F}, Z_{\rm NH}, S_{\rm F}$) predictors (blue) and ($T_{\rm NH}, Z_{\rm NH}, S_{\rm NH}$) predictors (red), and with datasets of length 7200 years (plain lines), 800 years (dashed lines), and 100 years (dashed-dotted lines). The results illustrate the lack of data regime, with very slow convergence of the prediction skill with the dataset length, and with a clear tradeoff between dataset length and size of optimal geographical area for best prediction. The optimal area for 2 m temperature and soil moisture is the local one (France area).

The red curve on Fig. 7(b) presents the result for the predictor set ($T_{\rm F}, Z_{\rm NAE}, S_{\rm F}$), to be compared with the benchmark curve with ($T_{\rm F}, Z_{\rm NH}, S_{\rm F}$). Comparing the plain red and blue curve, we see that with a 7200-year training dataset, the ($T_{\rm F}, Z_{\rm NH}, S_{\rm F}$) predictor set is indeed the optimal one. With a 7200-year training dataset the neural network is actually able to extract the supplementary information beyond the one which is contained in the North Atlantic and European area. The improvement is significant with increase of the normalized logarithmic score up to 10%, which is important.

Comparing now the dashed red and blue curves, for the case with 800-year-long training dataset, one clearly sees the same pattern, although with a smaller improvement when comparing the predictor for the complete field ($T_{\rm F}, Z_{\rm NH}, S_{\rm F}$) and the one with the incomplete one ($T_{\rm F}, Z_{\rm NAE}, S_{\rm F}$). However, using only 100 years for training, the dashed-dotted line features an opposite conclusion. The training experiment with the incomplete field, on the North-Atlantic Europe sector, gives a better skill than the training with the complete field. The interpretation we give is that a 100-year-long training set is not complete enough to deal with the complexity of the predictor defined on a larger area. This is a manifestation of the tradeoff between dataset length and predictor complexity in a regime of lack of data. This confirms our qualitative prediction and makes it quantitative.

We complement this study of the tradeoff between predictor complexity and dataset length in a regime of lack of data, by discussing the cases of soil moisture and 2 m temperature. For those fields, the situation is different because it might be clear on physical grounds that these two fields are relevant mainly locally, close to the heatwave area. Figure 8(b) features the same benchmark blue curve as the one on Fig. 7(b): the prediction skill for a neural network trained with the optimal predictors ($T_{\rm F}, Z_{\rm NH}, S_{\rm F}$). It also shows the prediction skill for a neural network trained with the predictors ($T_{\rm NH}, Z_{\rm NH}, S_{\rm NH}$), where now both the soil moisture and temperature fields are used on the full northern hemisphere mid- and high-latitude sector. The results clearly show that this prediction with hemispheric temperature and soil moisture is systematically worse than the one with local predictors. This confirms that the optimal area is the local France one, for these two fields. Thus, complexifying fields with parts that contain no relevant information and provide essentially noise

might be manageable with neural networks trained on huge datasets, but it is a problem in a regime of lack of data. Figure 7(b) also shows that the degradation of the score, in relative terms, is lower when the amount of data is increased, in agreement with our interpretation.

Our third conclusion is that learning the probabilities of extreme heatwaves with neural networks clearly takes place in a regime of drastic lack of data. Several hundreds or even thousands of years would be needed for optimal prediction, even when using only two or three representative fields as predictors. Trying to use more fields, for instance, more information about the vertical structure of the geopotential height, or for example information related to the temporal development of the dynamical fields, most probably requires even longer training datasets. Our results clearly show that with a 7200-year training dataset, the skill is at best only very marginally improved when using more fields. Moreover, in this regime of lack of data, there is a tradeoff between dataset length and predictor complexity. For instance, for predicting extreme heatwaves over France, benefiting from hemispheric information beyond the North Atlantic and Europe sector requires at least several hundreds of years of learning datasets.

C. Physical insight, interpretability of neural network predictions, and committor function composite maps

What has the neural network actually learned? The interpretability of neural network predictions is a pervasive question when they are applied in physical sciences. In this section we propose a basic approach for visualizing committor function. With this aim, we plot composite maps of the 500 hPa geopotential height, conditioned on very large values of the committor function.

The neural network output is the probability $p(\boldsymbol{x})$ to observe a heatwave τ days from now, given that we observe today the predictor field $\boldsymbol{x}$. $p(\boldsymbol{x})$ is a function over the set of all the possible predictor states. This function, called a committor function, is therefore extremely complex and impossible to visualize for high dimensional spaces, by contrast with committor functions for simple dynamical systems [38]. To get insights about some very specific behaviors of this function, we will try to look at a single property: how do the fields that give very large values of $p(\boldsymbol{x})$ look like? Equivalently, we will consider the fields that the neural network ranks as the most likely to produce a heatwave, and look at the average of their corresponding 500 hPa geopotential height field.

To simplify the discussion, we consider a neural network trained on the hemispheric 500 hPa geopotential height Z_{NH} only. Once it has been optimized using the training set, the neural network can associate to any other field, for instance, in the validation set, the estimated probability p that this field will lead to a heatwave. We select all the events in each of the validation sets which are above the 99.9 percentile in the distribution of the committor values. For example, for $\tau = 0$ those are all the events with $p > 0.68$. Notice that according to our protocol we have 10 folds of train-validation split (Sec. IV B). This means that we can associate a committor value with each day of the full 8000-year-long set and no occasion are we evaluating the committor in the training set. We then compute the average of the 500 hPa geopotential height maps, conditioned on having p values above the 99.9 percentile, for the entire 8000-year-long set. The resulting composite maps displayed on Fig. 9 reflect averaged properties of the fields which are most likely to lead to a heatwave, according to the neural network prediction. This operation can be repeated for different values of lead times τ, which may have smaller values of p threshold, simply because the neural network becomes less certain for larger values of τ

For $\tau = 0$, on Fig. 9(a), we observe a tripole structure for the geopotential height anomalies. First of all, we see an anticyclonic anomaly over Europe, which is expected, given that heatwaves in summer are associated with anticyclonic anomalies. We also see a cyclonic anomaly over Greenland and the Arctic area and two other anticyclones over Eastern North America and Northern Siberia. For this composite, the anticyclonic anomaly over Europe is extremely strong: the maximum value of the composite average of the 500 hPa geopotential height anomaly has a maximum value of 151 m [see Fig. 9(a)], to be compared with a typical maximal 500 hPa geopotential height anomaly over Europe of order of 120 m (see, for instance, the snapshot on Fig. 3) and a variance for

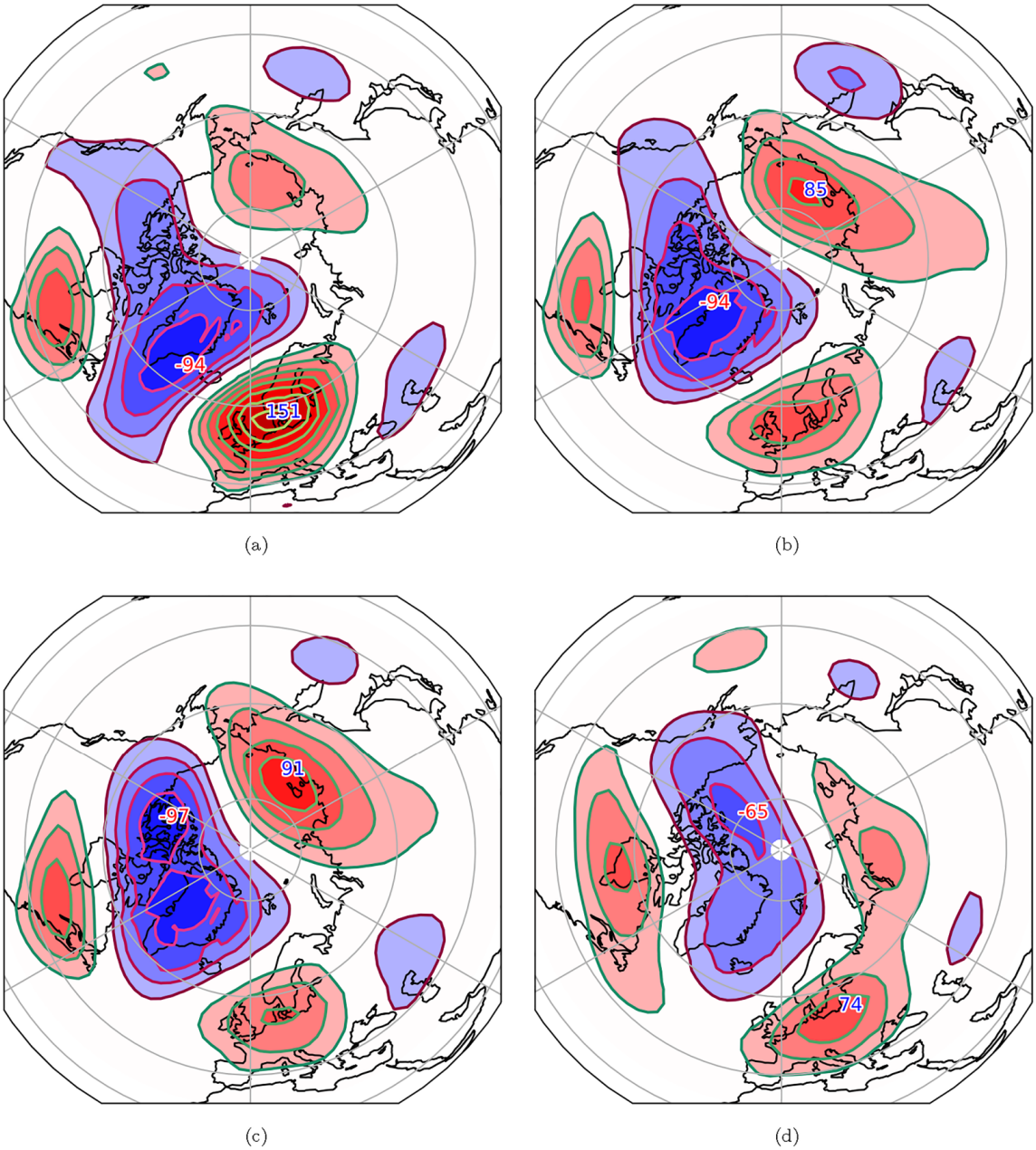


FIG. 9. Composites of the 500 hPa geopotential height maps Z_{NH} (in meters), conditioned on committor values p above the 99.9 percentile, at different lead time τ. Panels (a), (b), (c), and (d) with values of $\tau = 0,\ 5,$ 10, and 15, respectively. The regions of positive anomalies are indicated in red, while the negative anomalies are indicated in blue (we are using seismic colormap). The isolines are separated by a value of 20 m. The maximum geopotential height anomaly is indicated with a number colored in blue, while the minimum value is colored in red.

the climatology of the 500 hPa geopotential height anomalies at midlatitude of order of 60 m. Obtaining such a large value for a composite average, means that all the fields in the composite have a systematic stronger than usual anomaly over Europe with a coherent pattern. Similarly the cyclonic anomaly over Greenland is very strong, with a minimum value of the averaged 500 hPa geopotential height anomaly of -94 m, to be compared with typical minimal 500 hPa geopotential height anomalies over the Greenland-Arctic area of order of -200 m (see, for instance, the snapshot

on Fig. 3) and a variance for typical 500 hPa geopotential height anomalies of order of 90 m at high latitudes. This also points to a very coherent and systematic pattern over Greenland. The two other anticyclonic anomalies over Eastern North America and Northern Siberia have weaker values, of order of 40 to 60 m, which are still comparable to the variance of the 500 hPa geopotential height, thus showing a relatively strong coherence. The coherence of the overall pattern can also be assessed by comparing those values to the standard deviations within the composite set itself, which are of order of 40 m in midlatitudes and 60 to 90 m in the Arctic area. All those comparisons point to fairly coherent and robust patterns superimposed with fluctuations of the order of the standard deviation.

The overall pattern is a clear mode 3 pattern, with an overall shift of the cyclonic anomalies poleward and of the anticyclonic anomalies equatorward. This structure is much reminiscent of the wave-number 3 extreme teleconnection observed for European heatwaves [28] and has been interpreted as related to Rossby waves with wave-number 3, and phase speed close to zero, leading to a long-lasting quasistationary pattern. This result suggests that the recognition of this wave-number quasistationary pattern might be key for the neural network prediction skill.

Let us now look at different values of the time lag τ. The three other panels of Fig. 9 show the composite 500 hPa geopotential height maps, conditioned on p values above the 99.9 percentile, for $\tau = 5$, 10, and 15, respectively. The four patterns look surprisingly similar, whatever the value of τ, with a consistent wave-number 3 pattern, poleward shift of the anticyclonic anomalies and equatorward shift of the cyclonic ones. This result suggests that the long-term prediction skill of the neural network might also be associated with this quasistationary pattern. Following this remark a natural hypothesis would be that the long-term skill of the neural network might be related to the probability of this pattern to stay quasistationary for a long enough period. Testing this very interesting hypothesis is however beyond the capabilities of the approach described in this paper and will be considered in future works.

In addition to the strong analogies, we note that the four patterns are slightly modified when changing τ. For $\tau = 5$ the anticyclonic anomaly over Northern Asia is stronger and larger, and the anticyclone over Europe is less intense. This tendency is even more pronounced for $\tau = 10$. For $\tau = 15$, the wave-number 3 pattern turns to a tripolar structure.

Plotting composite 500 hPa geopotential height maps, conditioned on very large values of the committor function, gives most probably only a limited view of what the neural network might have learned. Trying to interpret the neural network results with an averaged quantity (composite) only, is very limited from the point of view of a stochastic interpretation of the prediction. However, these composite averages already clearly show very interesting teleconnections associated with previously discussed wave-number 3 patterns. This opens questions for more detailed future analysis of the dynamical features which are important for prediction, their probabilities, and the capabilities of the neural network to identify them. Those future analyses will also consider tools for physical interpretability of machine learning forecasting.

D. How to ensure continuity of the committor function when the lead time is changed and how to accelerate the training stage

The proposed neural network predicts the committor function: the probability $p(\boldsymbol{x}, \tau)$ to observe a heatwave τ days from now, given that we observe today the predictor field $\boldsymbol{x}$. First, we display in Fig. 10(a) a trace of this committor function (in green) at $\tau = 0$, for the summer of a randomly chosen year, compared to the actual realization of temperature anomaly A. One sees that committor function and actual events are correlated, yet not identical, as expected since the committor is only a probability of having a heatwave and the neural network is only trained on discrete labels and is not provided the full information $A(t)$.

Up until now, we have studied some aspects of how the total prediction skill evaluated on validation sets depended on τ. In this subsection, we are rather interested in how committor function varies along the trajectory (passage of t physical time in the simulation) while fixing a specific event

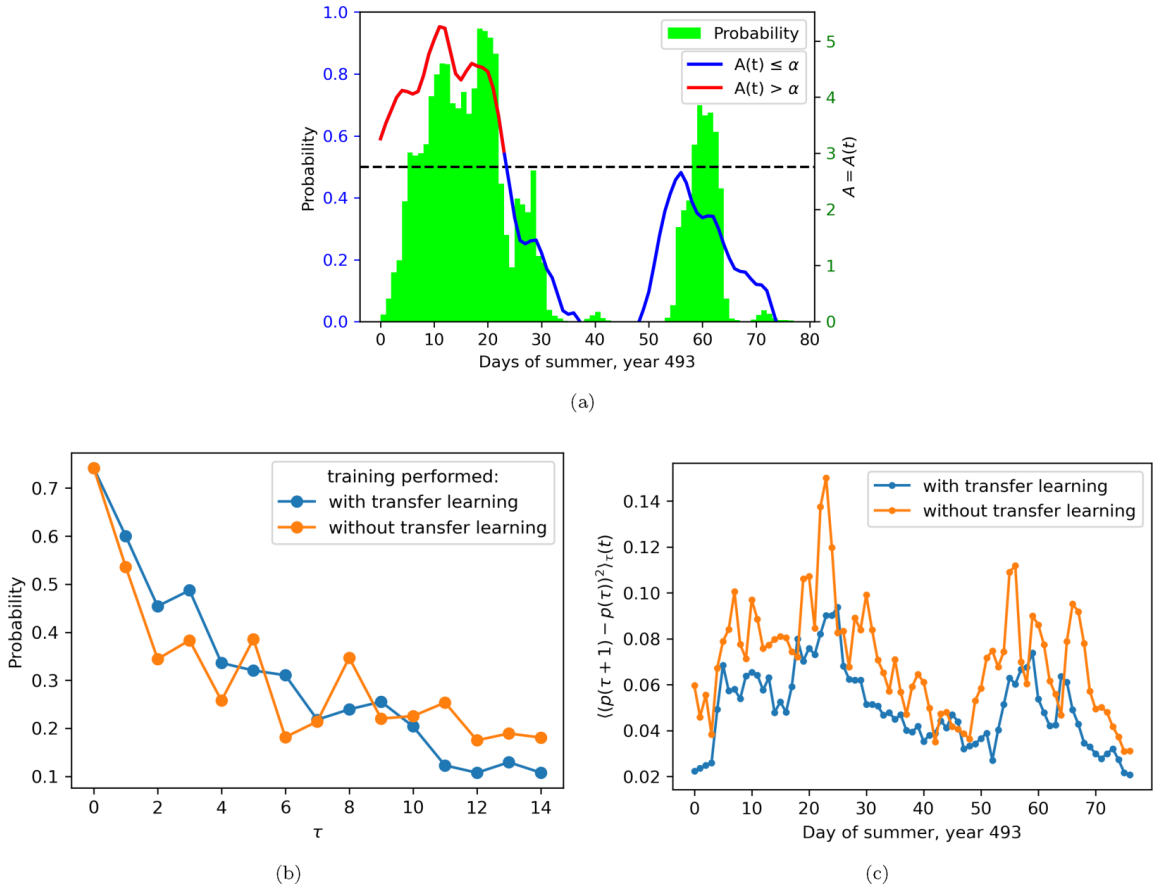


FIG. 10. (a) Illustration of the $A(t)$(blue/red line) and $p[X(t), 0]$ (in green) trajectories during summer of a specific year; red line corresponds to moments where the actual $A = A(t) > \alpha$ [see Eq. (1)], above 95 percentile threshold, while blue line denotes that $A = A(t) \leqslant \alpha$. The filled green segments show the probability predicted by the neural network trained on T_F, Z_{NH}, S_F. This plot illustrates the general correlation between $A = A(t) \leqslant \alpha$ and T_F, Z_{NH}, S_F. (b) Evolution of the committor function along a trajectory $p[X(t^\star - \tau), \tau]$ as a function of lead time τ, for a prediction of a potential heatwave at a prescribed physical time $t^\star$. The orange points show the committor learned independently for each τ, while the blue ones are obtained with transfer learning from one τ to the next. (c) Display of $\sigma_\tau(t)$, as defined in Eq. (10), for the same year as in panel (b), for the method with transfer learning (in blue) and without (in orange). We see that the latter is almost always above. This illustrates that variance between subsequent points is larger when transfer learning is not applied.

at time $t^\star$. Thus, we must vary simultaneously $\tau = t^\star - t$. Of particular interest is a smoothness property of the committor as we vary t while fixing $t^\star$.

This smoothness property could be understood as a consistency of the prediction through time. A lack of smoothness, in a risk prevention context, would mean that the prediction would highly fluctuate from one day to another. Besides the fact that this would probably be the sign of some deficiency in the prediction, this might also be detrimental from a communication point of view, and create concerns and disbelief among the users of the information. On more scientific grounds, if the prediction p is used as an input for another computation or algorithm, then the consistency and smoothness properties might also be very important, both theoretically and practically.

We demonstrate in this section that transfer learning can be used to address this issue. Transfer learning is used extensively in deep learning, where it allows dramatic reduction of the training time, and improvement of the skill, by using networks pretrained on more general large datasets [96,97]. It has also been used for several climate and weather applications, for instance, for ENSO prediction

[45], where the authors have pretrained the network on CMIP model outputs, prior to applying it to reanalysis datasets. When looking for smoothness properties with respect to τ, an alternative could be to train on several lag times at the same time, see, for instance, Ref. [98] and references therein.

To study the τ dependence of p, one could either fix the state $\boldsymbol{x}$ or rather follow the evolution of state $\boldsymbol{x}$ with time. In this section, we make the second choice. We thus fix a time $t^\star$ corresponding to the potential start of a heatwave, and we study how $p[X(t^\star - \tau), \tau]$ depends on τ. When τ decreases, the prediction is made closer in time to the start of a potential heatwave. We then expect the event to be more predictable, as τ decreases. There is no reason to expect a monotonic evolution. However, in general, we expect $p[X(t^\star - \tau), \tau]$ to be a smooth function of τ. Since we are working with discrete daily data the highest resolution we can achieve is obviously $\Delta\tau = 1$ day. Thus, smoothness must be understood in colloquial terms rather than a strict mathematical definition. Below we give a specific example.

We first train the network in an independent way for different values of τ, with the reference predictors (T_{F}, Z_{NH}, S_{F}). For illustration purposes, we choose a specific year and a value of $t^\star$ such that it corresponds to the strong heatwave at $t^\star$. The orange curve on Fig. 10(b) shows the evolution of the prediction $p[X(t^\star - \tau), \tau]$ with τ. One sees that the prediction is relatively consistent over time, when τ decreases. However, we observe some fluctuations, from one day to the next, of the order of 10% to 30% of the predicted probability. The level of these fluctuations is higher for intermediate values of τ, between 2 to 10 days.

To reduce these fluctuations and to improve time consistency and smoothness of the prediction, we adopt a transfer learning strategy. The main idea is to initialize the weights of the neural network for the model at a given lead time τ, based on the trained model at a previous lead time $\tau - 1$. The heuristic idea is that the corresponding change in $X(\tau)$ is not so large and already contains very good information for the prediction at the next time step. Note that this also allows to drastically reduce the training time: early stopping of the training is typically necessary only after 5 epochs, as opposed to 40 or more when starting from random initialization. The reference is the blue curve on Fig. 10(c). In addition, we tested the effect of this transfer learning strategy on the overall prediction skill but we have seen no significant improvement or degradation. This suggests a hypothesis that we have reached the capacity of the network to learn the extreme events.

For quantifying the reduction in the fluctuations of p, we introduce a smoothness metric. From the discrete series of p, we compute the forward difference of the committor at successive τ:

$$\Delta_\tau(t) := p[X(t - \tau - 1), \tau + 1] - p[X(t - \tau), \tau], \tag{9}$$

which is a function t. The smoothness metric consists in computing standard deviation of $\Delta_\tau(t)$ for all $\tau \in \{0, \ldots, 14\}$ days:

$$\sigma_\tau^2(t) := \left\langle \Delta_\tau^2(t) \right\rangle_\tau - \langle \Delta_\tau(t) \rangle_\tau^2, \tag{10}$$

where the brackets $\langle \cdots \rangle_\tau$ denote the average over the subscript parameter τ. Figure 10(c) compares $\sigma_\tau(t)$ in cases with and without transfer learning; it clearly demonstrates that the former tends to have smoother committor function w.r.t. τ. The level of fluctuations from one lead time to the next one has been reduced by a large factor, when compared to the orange curve without transfer learning. We can also apply this quantitative measure of smoothness to the sequence of all days t in each validation set of which we have 10 folds, as described in Sec. IV B. Since the sequence now also consists of all times t, we average on all t and τ, denoted as $\sigma_{\tau,t}$ which is a scalar. This results in the following values with transfer learning $\sigma_{\tau,t} = 2.59 \pm 0.07 \times 10^{-2}$ and without $\sigma'_{\tau,t} = 3.82 \pm 0.09 \times 10^{-2}$. The difference is almost 50%. This quantitative measure allows us to conclude that transfer learning improves dramatically the time consistency and the smoothness of the extreme heatwave prediction, while, independently, reducing the training computational time.

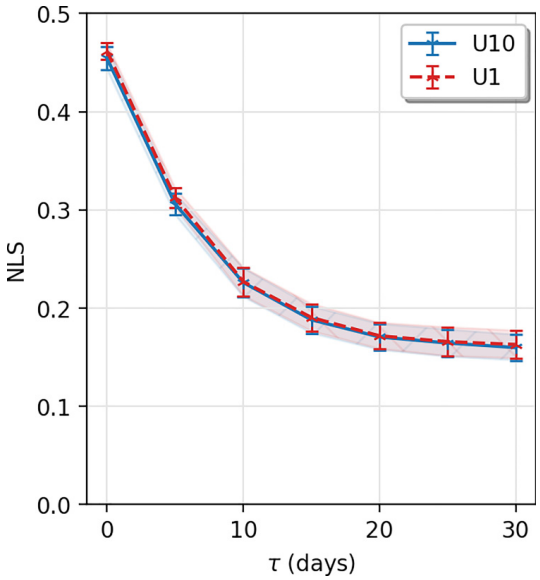


FIG. 11. Normalized logarithmic score versus lead time τ for a neural network trained with the (T_F, Z_{NH}, S_F) predictors, with undersampling with $r = 10$ (U10) (blue) and without undersampling (U1) (red). The two skill curves are within error bars of the experiment.

E. Robustness of the learning protocol with respect to the undersampling strategy, the level of rarity, and the neural network architecture

1. How does majority class undersampling affect the prediction skill?

In this section we discuss the effect of majority class undersampling strategies on the neural network prediction skills. We train the neural network with the reference set of predictors (T_F, Z_{NH}, S_F), which was proven optimal in Sec. V A. We either train the network without undersampling, or with the majority class undersampling taking into account the change of probability measure, as discussed in Sec. IV D and using Eq. (8). For this second case, the undersampling rate is $r = 10$. Figure 11 shows that the prediction skills are the same, within the error bars. This leads to a conclusion that the prediction skill of the extreme heatwaves considered here is not influenced by the undersampling strategy.

Majority class undersampling is however useful, to reduce the memory request and the training time by about a factor 10. One may wonder if the similar conclusion can be reached for other types of extremes, which could be the subject of future work.

2. Prediction skills for more extreme events

On Fig. 12 we present the comparison between predictions of 95 percentile heatwaves (consistent with all the previous analysis) and 99 percentile heatwaves. This corresponds to definitions of heatwaves for large deviations $A(t) > 2.75\,\mathrm{K}$ and much more extreme ones $A(t) > 3.91\,\mathrm{K}$ in the latter case. In other words, the objective field used to make the prediction X is exactly the same (T_F, Z_{NH}, S_F) as well as the architecture presented in Fig. 4 but the labels are defined based on two different criteria discussed above. Undersampling rate was chosen as 10 for 95 percentile heatwaves as usual and 20 for the 99 percentile case. The resulting scores are plotted on Fig. 12.

At a qualitative level, the normalized logarithmic scores behave similarly for the two cases, with the same decrease of the skill over synoptic timescales, up to a plateau corresponding to the effect of soil moisture. At a more quantitative level, we stress that the nearly equal values of the scores for the two cases is an accident. There is no logical reason to compare directly the quantitative values of the skills for the two experiments. The first reason why they are not comparable is that the

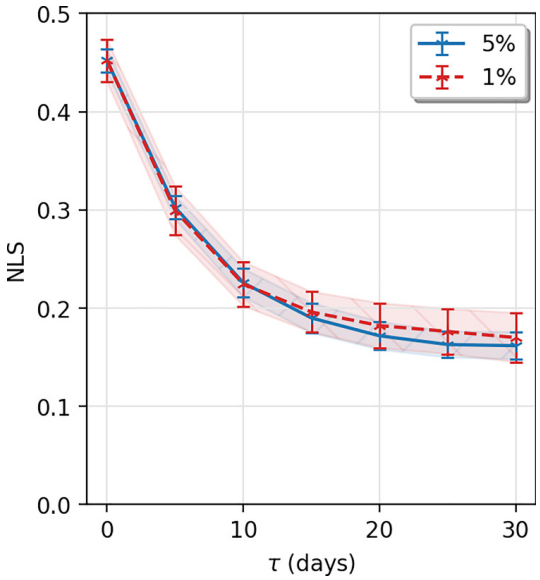


FIG. 12. Benchmarks for the optimal normalized logarithmic score obtained by the proposed neural network for 5 percent heatwaves (blue) and 1 percent heatwaves (red).

normalized logarithmic scores are normalized differently in each case, because of the different base climatological probability. The second reason is that even with the same climatological probability, there would be no reason to expect two events of different classes to have the same exact real committor value, which is the intrinsic probability we would learn if the learning would be perfect.

It is interesting to note that in our previous study [55], majority class undersampling or transfer learning among classes was improving the categorial (0 or 1) prediction of extreme heatwaves, when we were assuming a deterministic relation between predictors and heatwaves. In this new paper, neither majority class undersampling nor transfer learning affect the probabilistic prediction skill, neither positively nor negatively, when we now actually consider the probabilistic nature of the relation between predictors and the heat waves. One might be surprised by these different behaviors. One possibility would be that our learning was not perfect, in one or the other case, either because of lack of data or suboptimal learning protocol. However, we stress that even with a prefect learning, there is no disagreement nor contradiction between these seemingly different results, as what is actually tested for the prediction is of a different nature.

This rises a very interesting general question. In several previous studies, including ours [55], a categorial test, for instance, the Matthews correlation coefficient, was used to test a relation between predictors and events which is actually intrinsically probabilistic, and not deterministic. This was logically problematic and should be avoided. Beyond the logical problem, might it be possible that testing a probabilistic relation with a categorial test lead to some practical inconsistencies and divergent conclusions? As we are interested by probabilistic forecast we do not consider further this question.

3. Robustness of the results with respect to the neural network architecture

Throughout the article, the architecture displayed in Fig. 4 was consistently used when we refer to the CNN (or neural network) methodology. We have tried other architectures, changing the amount of filters, using additional layers, and other changes for a better optimized network. None were convincing in the present framework. Specifically, we found that deeper CNNs had slightly lower skill, and this is the reason why they are not considered here.

While we exclusively show results pertaining to stacking the fields, we have also considered combining the fields into separate CNNs which are then concatenated on a single dense layer.

The latter approach does not work so well, as already reported in Ref. [55], and is more difficult to implement. This suggests that stacked architecture is potentially benefiting from local cross-correlations between temperature, soil moisture, and geopotential.

VI. CONCLUSIONS AND PERSPECTIVES

A. Probabilistic forecast with machine learning and other methodological contributions

In this paper we have advocated a probabilistic approach for the forecast of weather and climate related problems, in particular extreme events, because for chaotic dynamical systems the relation between predictors and the predicted phenomena is intrinsically probabilistic. For forecast validation, logarithmic or ignorance score, occasionally used in weather forecast and climate, is directly linked to the cross-entropy skill. The latter is used in many machine learning problems as opposed to, say, Brier score. Through an affine transformation of the logarithmic score we defined the normalized logarithmic score, which has convenient properties to be equal to zero for a forecast based on the climatological frequency, to be positively oriented and to be always lower than one.

We have demonstrated the efficiency of this approach for forecasting long-lasting extreme heatwaves, within a dataset consisting of PlaSim climate model outputs. Using geopotential height, temperature, and soil moisture fields as predictors, we have trained a convolutional neural network to forecasts extreme heatwaves on a validation set. Methodologically, this probabilistic approach extends previous work using machine learning for categorical deterministic prediction of daily [54] or long-lasting heatwaves [55].

At a methodological level, we have also demonstrated the interest of transfer learning to improve the temporal consistency and smoothness with time of the prediction. This is a key issue for practical applications related to risk forecast. We have also demonstrated the interest of majority class undersampling and of transfer learning to lower the RAM, CPU, and computational time usage during the learning stage of the network.

B. Key general scientific conclusions

1. The lack of data regime for machine learning in weather and climate studies

The main scientific message of this work is that training neural networks for predicting large-scale features of weather or climate phenomena will most of the time operate in a regime of lack of data. We have demonstrated this clearly in the case of extreme heatwaves. Using subsets of 8000-year climate output with a data reduction protocol leads to a significant drop in the prediction skill. using three important fields, one at hemispheric scale (500 hPa geopotential height) and two at a local scale (soil moisture and 2 m temperature). This points to the need of thousands or tens of thousands of years of data for proper convergence, perhaps more if one would like to benefit from the information available in more complementary fields.

The climate model output has some known biases with respect to real fields, but its structure and complexity is most probably the same as the one for reanalysis datasets or real fields. It is likely that obtaining a converged statistical model based on real or reanalysis dataset would require a length of the same order of magnitude, although this cannot be tested directly. This is a drastic constraint given the definitive limitation of historical data. A similar lack of data problem exists for many applications of machine learning for physical and natural sciences, but the lack of observed or reanalysis data is rather severe for studying large-scale weather and climate phenomena. To circumvent this problem, one will have to find ways to combine model data and reanalysis datasets as discussed in Sec. VI D 1.

This problem is exacerbated when studying extreme events because of their rarity. This is indeed a very important remark. The heatwaves we have studied in this paper, which were defined as the 5% percentile of summer data with a correlation time of a few days, are events with typical return time of a few years in the studied climate. Many climate and weather phenomena have return times of a few weeks to a few years, they will equally fall in this lack of data regime. From a point of view

of extreme event impact, a return time of a few years is actually not so rare and risk management specialists are interested in much rarer events.

We have clearly demonstrated that there exists a tradeoff between the length of the dataset and the complexity of the used predictors. For instance, for forecasting extreme heatwaves over France, we showed that the 500 hPa fields contain useful information for improving the prediction skill at the hemispheric scale. However, to properly learn part of this information, the neural network needs at least a few hundred years of data. To obtain a larger improvement using hemispheric fields, compared to fields at the scale of North-Atlantic and Europe, actually require thousands of years of data. This tradeoff predictor complexity/dataset length is very natural for machine learning in a context of lack of data, and should be present in most applications of neural networks when studying large-scale features of climate or weather data.

2. Neural networks seamlessly use the predictive power of fast dynamical fields and slow physical drivers

In many predictive statistical approaches aimed at studying weather and climate phenomena, researchers discuss separately the effects of fast dynamical fields and slow physical drivers. For extreme heatwaves, see, for instance, the interesting works using the analog method for understanding the effect of fast dynamical drivers [71,72] and some complementary works on slow drivers [59]. This dichotomy makes perfect sense given the timescale separation and the complexity of the different approaches. There is however a need for methods that combine both at the same time. For instance, if one wants to quantity the respective impacts of these two types of drivers, then one needs a method able to compute predictability skills by dealing with the two types of fields together.

In this work, we have demonstrated that neural networks handle, without any practical or methodological difficulty, the 500 hPa geopotential height (fast dynamical field) and soil moisture (slow physical driver). This is in contrast to a method that would explicitly build the effect of averaging over fast drivers, conditioned on the slow drivers, which would require a lot of tricky computations. Moreover, the predictive approach provides actual numbers that quantify the respective role of the two types of fields.

3. Probabilistic forecast as a tool for physical analysis of drivers

Current weather models can also handle seamlessly the prediction of the effects of fast dynamical fields and of slow physical drivers. They are actually certainly the most precise way to make such studies. However, the objective of using neural network is different and complementary. Neural networks with probabilistic forecast provide a statistical model, which associates to each set of drivers a predictive skill. Alternatively suppressing the different predictors, is a way to estimate the causal relation between any set of fields and the event of interest. This can be used for *a posteriori* statistical studies, to perform fast and efficient process studies, and to analyze the impact of different drivers. Making similar studies with weather or climate model would be extremely difficult in practice and would require huge computations. We warn however, that any inference about information content from machine learning experiments, assumes that the learning is of a good quality.

We have analyzed, quantitatively, the relative potential of soil moisture and 500-hPa geopotential height, in triggering extreme heatwaves. By adding or removing different fields, and comparing the prediction skills, we can see which are the main drivers. For instance, we have demonstrated that the 2 m temperature carries part of the predictive information of both the soil moisture or the 500 hPa geopotential height, however we conclude that it carries no new significant information by itself. We have also demonstrated that geopotential height at other altitudes or isopressure levels, carry nearly no new information that can be tapped with a 8000-year-long dataset with the given neural network.

Those examples illustrate the potential use of neural network for other process studies in weather and climate dynamics. The key point is the quantitative nature of the analysis.

C. Conclusions for extreme heatwave drivers

The main conclusions for extreme heatwave prediction are as follows.The 500 hPa geopotential height combined with soil moist contains the most useful information in the short run, with only a very small improvement of the skill provided by adding the 2 m temperature. The prediction skill associated to the 500 hPa geopotential height decays approximately exponentially with a decay time of about 7 days. Soil moisture contains very important complementary information, with a plateau skill that does not decay much on timescales of order of 15 days to a month. This corresponds to the conditional probability to observe extreme heatwaves for some given soil moisture, independently of the dynamics. These two sets of information seem to add up when the two fields are used together, to make the best possible prediction.

We have also concluded that the set of 500-hPa geopotential fields which are selected by the neural network as having a large probability to lead to long-lasting heatwaves, are consistently distributed around a characteristic hemispheric pattern dominated by wave-number 3 Rossby waves with a shift poleward of the cyclonic anomalies and a shift equatorward of the anticyclonic anomalies. This pattern is also seen in composite maps, conditioned on extreme heatwaves, which are plotted independently of the neural network. An analogous wave-number 3 pattern has already been observed for European and Scandinavian long-lasting heatwaves [28]. This consistency shows that the neural network is either able to recognize this pattern, or is able to recognize other characteristic features of extreme heatwaves which correlate with this pattern. Understanding further those very interesting observations, and the dynamical nature of this pattern, will require developing the interpretability of machine learning approaches as further discussed in Sec. VI D 1.

D. Key perspectives

1. Perspectives for the lack of data problem for machine learning for climate studies

We have concluded in the previous sections that the requested dataset length for convergence of the training of statistical models based on deep architectures is much longer than reanalysis datasets. Since the latter are so short, the use of model data is necessary. However, climate models are more biased compared to reality than reanalysis datasets are. There is thus a need to couple the use of climate model and reanalysis datasets to make the best of their complementary potential. A natural way is to use transfer learning: first learning from extremely long climate model datasets and then reusing the weights of the learned model as an initial condition for a new training for the reanalysis dataset. Such a transfer learning approach has already been used in several past works in atmospheric sciences and climate studies [45].

However, this approach (transfer learning) might not be sufficient, for instance, if the climate model dataset itself does not contain enough characteristic events. This is probably the case for studying rare or extremely rare events. To improve the prediction skill, it is natural to assume that the rare extreme event samples are requested rather than data corresponding to the typical states of the system. Testing this assumption motivates the case for importance sampling algorithms. Regarding the difficulty of sampling exceptionally rare extreme events, e.g., unprecedented heatwaves, we have recently developed rare event simulation techniques that are able to multiply by several orders of magnitude the number of observed heatwaves with PlaSim model [28] and with CESM (the NCAR model used for CMIP experiments) [69]. We are currently working on coupling these rare event simulations with the machine learning forecast developed in this paper. The point is to improve both rare event simulations using machine learning forecast, and machine learning forecast using the unprecedented heatwave statistics obtained with rare event simulations. We have already coupled machine learning simulations with rare event algorithms, for simple academic models [77]. Coupling the rare event simulations with neural networks is a very interesting albeit complex perspective to solve the key fundamental issue of lack of data in the science of climate extremes.

Another important perspective is to develop a new neural framework that would be suited for a better physical interpretability of the prediction skills, to increase dynamical and physical understanding.

In weather and climate dynamics, ensemble forecast by a weather system (for instance, ECMWF) is considered as the reference probabilistic medium range forecast. The ensemble members are used to make probabilistic predictions, for instance, at medium range or subseasonal timescales. An important question for the future will be to compare the skill of machine learning probabilistic forecasts compared to ensemble forecast by weather systems, for instance, for predicting extreme heatwaves.

A key point though is that the two methods, weather systems based on the equations of physics and data assimilation, and learned statistical models, have completely different uses and perspectives, and are highly complementary. On the one hand, machine learning alone is unable to learn the current state of the atmosphere precisely and to incorporate the wealth of available observation data dealt with by weather systems. A weather system does not make just a single prediction on a specific event, but compute the full state of the system. On the other hand, a weather system needs dedicated infrastructures and millions of computation hours, while an already trained statistical model usually makes a forecast in less than a second on a laptop. A statistical model can be used to assess the probability from any field, not just the ones in the historical records or the forecasted ones. Then it is extremely likely that weather systems will remain the reference for actual real time medium range forecasts, while statistical model can be used for process studies, driving rare event simulations, statistical analysis, or cheap forecast for targeted scope, or possibly subseasonal to seasonal forecasts.

2. Perspectives for extreme heatwaves

We have argued in Sec. VI B 3 probabilistic forecast issued by neural networks provide a way to estimate the relations between any sets of predictor fields and the event of interest. This is then a tool to quantitatively study the role of each process, very efficiently and practically. Moreover, the quantitative nature of the relation between the predictors and the prediction gives also the opportunity to compare those relations for different models, different datasets, and different climates.

Using this tool opens the door to hundreds of process studies. It could be used, for instance, to further ascertain the impact of other slow drivers [59] on extreme heatwaves and other extreme events, and how they combine with fast dynamical drivers to produce them. One could also use this tool for the purpose of assessing model biases, to make climate change studies by comparing different datasets with different climates, and finally to make much more precise impact studies of extreme events. As an important example, it has been demonstrated that local thermodynamics drives monthly midlatitude summertime temperature variance [24], and that CMIP model might have some bias in reproducing this effect [24]. Then, training neural networks on different CMIP models could be used for intercomparison, specifically studying the effect of local thermodynamics on extreme events.

Given its very easy implementation and its scientific potential, we hope that the deep learning methodology we developed will be used to address many key questions related to extreme events, and other large-scale atmosphere and climate phenomena.

The coding resources for this work, such as the python and jupyter notebook files, are available on a GitHub page [99] branch “noxarray” and “main” and is part of a larger project at ENS de Lyon with multiple collaborators in the branch “subm2.” We do not have the infrastructure to make the 8000 year PlaSim dataset available online at this time, but it might be shared to interested colleagues, whenever feasible in practice.

ACKNOWLEDGMENTS

This work was supported by the ANR grant SAMPRACE, Project No. ANR-20-CE01-0008-01 (F.B.). 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. We acknowledge CBP IT test

platform (ENS de Lyon, France) for ML facilities and GPU devices. The platform operates the SIDUS solution [100] developed by Emmanuel Quemener. This work was granted access to the HPC/AI resources of CINES under the allocations 2018-A0050110575, 2019-A0070110575, 2020-A0090110575 and 2021-A0110110575 made by GENCI. We acknowledge the help of Alessandro Lovo in maintaining the GitHub page.

[1] IPCC, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, UK/New York, NY, 2021).

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

[3] United Nations Office for Disaster Risk Reduction (UNISDR), Review of disaster events (Centre for Research on the Epidemiology of Disasters (CRED), Brussels, Belgium, 2018).

[4] R. García-Herrera, J. Díaz, R. M. Trigo, J. Luterbacher, and E. M. Fischer, A review of the European summer heat wave of 2003, Crit. Rev. Environ. Sci. Technol. **40**, 267 (2010).

[5] H. M. Fritz, C. D. Blount, S. Thwin, M. K. Thu, and N. Chan, Cyclone Nargis storm surge in Myanmar, Nat. Geosci. **2**, 448 (2009).

[6] D. Barriopedro, E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera, The hot summer of 2010: Redrawing the temperature record map of Europe, Science **332**, 220 (2011).

[7] F. Otto, N. Massey, G. J. Van Oldenborgh, R. Jones, and M. Allen, Reconciling two approaches to attribution of the 2010 Russian heat wave, Geophys. Res. Lett. **39**, L04702 (2012).

[8] S. Y. Philip, S. F. Kew, G. J. van Oldenborgh, F. S. Anslow, S. I. Seneviratne, R. Vautard, D. Coumou, K. L. Ebi, J. Arrighi, R. Singh *et al.*, Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the U.S. and Canada June 2021, in *Earth System Dynamics Discussions* (Copernicus Publications, Gottingen, Germany, 2021), pp. 1–34.

[9] T. Woollings, D. Barriopedro, J. Methven, S.-W. Son, O. Martius, B. Harvey, J. Sillmann, A. R. Lupo, and S. Seneviratne, Blocking and its response to climate change, Curr. Clim. Change **4**, 287 (2018).

[10] K. Fraedrich, H. Jansen, E. Kirk, U. Luksch, and F. Lunkeit, The planet simulator: Toward a user friendly model, Meteorologische Zeitschrift **14**, 299 (2005).

[11] N. Nakamura and C. S. Huang, Atmospheric blocking as a traffic jam in the jet stream, Science **361**, 42 (2018).

[12] L. Wang and Z. Kuang, Evidence against a general positive eddy feedback in atmospheric blocking, arXiv:1907.00999.

[13] R. M. Horton, J. S. Mankin, C. Lesk, E. Coffel, and C. Raymond, A review of recent advances in research on extreme heat events, Curr. Climate Change Rep. **2**, 242 (2016).

[14] S. E. Perkins, A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at the global scale, Atmos. Res. **164-165**, 242 (2015).

[15] D. O. Benson and P. A. Dirmeyer, Characterizing the relationship between temperature and soil moisture extremes and their role in the exacerbation of heat waves over the contiguous united states, J. Climate **34**, 2175 (2021).

[16] F. D'Andrea, A. Provenzale, R. Vautard, and N. De Noblet-Decoudré, Hot and cool summers: Multiple equilibria of the continental water cycle, Geophys. Res. Lett. **33**, L24807 (2006).

[17] E. M. Fischer, S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär, Soil moisture–atmosphere interactions during the 2003 European summer heat wave, J. Climate **20**, 5081 (2007).

[18] M. Hirschi, S. I. Seneviratne, V. Alexandrov, F. Boberg, C. Boroneant, O. B. Christensen, H. Formayer, B. Orlowsky, and P. Stepanek, Observational evidence for soil-moisture impact on hot extremes in southeastern Europe, Nat. Geosci. **4**, 17 (2011).

[19] R. Lorenz, E. B. Jaeger, and S. I. Seneviratne, Persistence of heat waves and its link to soil moisture memory, Geophys. Res. Lett. **37**, L09703 (2010).
[20] P. Rowntree and J. Bolton, Simulation of the atmospheric response to soil moisture anomalies over Europe, Quart. J. Roy. Meteorol. Soc. **109**, 501 (1983).
[21] S. D. Schubert, H. Wang, R. D. Koster, M. J. Suarez, and P. Y. Groisman, Northern eurasian heat waves and droughts, J. Climate **27**, 3169 (2014).
[22] J. Shukla and Y. Mintz, Influence of land-surface evapotranspiration on the Earth's climate, Science **215**, 1498 (1982).
[23] M. Stéfanon, F. D'Andrea, and P. Drobinski, Heatwave classification over Europe and the Mediterranean region, Environ. Res. Lett. **7**, 014023 (2012).
[24] L. Vargas Zeppetello and D. Battisti, Projected increases in monthly midlatitude summertime temperature variance over land are driven by local thermodynamics, Geophys. Res. Lett. **47**, e2020GL090197 (2020).
[25] R. Vautard, P. Yiou, F. D'Andrea, N. de Noblet, N. Viovy, C. Cassou, J. Polcher, P. Ciais, M. Kageyama, and Y. Fan, Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit, Geophys. Res. Lett. **34**, L07711 (2007).
[26] L. R. V. Zeppetello, D. S. Battisti, and M. B. Baker, The physics of heat waves: What causes extremely high summertime temperatures? J. Climate **35**, 2231 (2022).
[27] S. Zhou, A. P. Williams, A. M. Berg, B. I. Cook, Y. Zhang, S. Hagemann, R. Lorenz, S. I. Seneviratne, and P. Gentine, Land–atmosphere feedbacks exacerbate concurrent soil drought and atmospheric aridity, Proc. Natl. Acad. Sci. USA **116**, 18848 (2019).
[28] F. Ragone, J. Wouters, and F. Bouchet, Computation of extreme heat waves in climate models using a large deviation algorithm, Proc. Natl. Acad. Sci. USA **115**, 24 (2018).
[29] J. Zscheischler, O. Martius, S. Westra, E. Bevacqua, C. Raymond, R. M. Horton, B. van den Hurk, A. AghaKouchak, A. Jézéquel, and M. D. Mahecha *et al.*, A typology of compound weather and climate events, Nat. Rev. Earth Environ. **1**, 333 (2020).
[30] A. Berg, B. R. Lintner, K. Findell, S. I. Seneviratne, B. van den Hurk, A. Ducharne, F. Chéruy, S. Hagemann, D. M. Lawrence, S. Malyshev, A. Meier, and P. Gentine, Interannual coupling between summertime surface temperature and precipitation over land: Processes and implications for climate change, J. Climate **28**, 1308 (2015).
[31] C. B. Field, V. Barros, T. F. Stocker, and Q. Dahe, *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, UK, 2012).
[32] G. Branstator and H. Teng, Tropospheric waveguide teleconnections and their seasonality, J. Atmos. Sci. **74**, 1513 (2017).
[33] K. Kornhuber, V. Petoukhov, D. Karoly, S. Petri, S. Rahmstorf, and D. Coumou, Summertime planetary wave resonance in the northern and southern hemispheres, J. Climate **30**, 6133 (2017).
[34] H. Teng and G. Branstator, A zonal wave-number 3 pattern of northern hemisphere wintertime planetary wave variability at high latitudes, J. Climate **25**, 6756 (2012).
[35] H. Teng, G. Branstator, H. Wang, G. A. Meehl, and W. M. Washington, Probability of U.S. heat waves affected by a subseasonal planetary wave pattern, Nat. Geosci. **6**, 1056 (2013).
[36] J. Zscheischler and S. I. Seneviratne, Dependence of drivers affects risks associated with compound events, Sci. Adv. **3**, e1700263 (2017).
[37] A. Tilloy, B. D. Malamud, H. Winter, and A. Joly-Laugel, A review of quantification methodologies for multi-hazard interrelationships, Earth-Sci. Rev. **196**, 102881 (2019).
[38] D. Lucente, F. Bouchet, and C. Herbert, Machine learning of committor functions for predicting high impact climate events, Technical report, Meetings (2020).
[39] D. Lucente, C. Herbert, and F. Bouchet, Committor functions for climate phenomena at the predictability margin: The example of el niño southern oscillation in the Jin and Timmermann model, J. Atmos. Sci. (2022).
[40] E. N. Lorenz, The predictability of a flow which possesses many scales of motion, Tellus **21**, 289 (1969).

[41] V. Balaji, Climbing down Charney's ladder: Machine learning and the post-Dennard era of computational climate science, Philos. Trans. R. Soc. A. **379**, 20200085 (2021).
[42] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, Deep learning and process understanding for data-driven Earth system science, Nature (London) **566**, 195 (2019).
[43] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Trans. Knowl. Data Eng. **29**, 2318 (2017).
[44] J. Cohen, D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, S2S reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts, WIREs Clim. Change **10**, e00567 (2019).
[45] Y.-G. Ham, J.-H. Kim, and J.-J. Luo, Deep learning for multi-year ENSO forecasts, Nature (London) **573**, 568 (2019).
[46] J. A. Weyn, D. R. Durran, and R. Caruana, Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data, J. Adv. Model. Earth Syst. **11**, 2680 (2019).
[47] Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins *et al.*, Application of deep convolutional neural networks for detecting extreme weather in climate datasets, arXiv:1605.01156.
[48] E. Racah, C. Beckham, T. Maharaj, S. Ebrahimi Kahou, M. Prabhat, and C. Pal, Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Red Hook, NY, 2017), Vol. 30.
[49] S. Giffard-Roisin, M. Yang, G. Charpiat, C. Kumler Bonfanti, B. Kégl, and C. Monteleoni, Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data, Front. Big Data **3**, 1 (2020).
[50] N. A. Agana and A. Homaifar, A deep learning based approach for long-term drought prediction, in *Proceedings of SoutheastCon'17* (2017).
[51] A. Dikshit, B. Pradhan, and A. M. Alamri, Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model, Sci. Total Environ. **755**, 142638 (2021).
[52] K. Chen, C. Kuang, L. Wang, K. Chen, X. Han, and J. Fan, Storm surge prediction based on long short-term memory neural network in the East China Sea, Appl. Sci. **12**, 181 (2022).
[53] X. Peng, H. Wang, J. Lang, W. Li, Q. Xu, Z. Zhang, T. Cai, S. Duan, F. Liu, and C. Li, Ealstm-qr: Interval wind-power prediction model based on numerical weather prediction and deep learning, Energy **220**, 119692 (2021).
[54] A. Chattopadhyay, E. Nabizadeh, and P. Hassanzadeh, Analog forecasting of extreme-causing weather patterns using deep learning, J. Adv. Model. Earth Syst. **12**, e2019MS001958 (2020).
[55] V. Jacques-Dumas, F. Ragone, P. Borgnat, P. Abry, and F. Bouchet, Deep learning-based extreme heatwave forecast, Front. Climate 4 (2022).
[56] M. Mudigonda, P. Ram, K. Kashinath, E. Racah, A. Mahesh, Y. Liu, C. Beckham, J. Biard, T. Kurth, S. Kim, S. Kahou, T. Maharaj, B. Loring, C. Pal, T. O'Brien, K. E. Kunkel, M. F. Wehner, and W. D. Collins, in *Deep Learning for Detecting Extreme Weather Patterns* (John Wiley and Sons, New York, NY, 2021), Chap. 12, pp. 161–185.
[57] I. Lopez-Gomez, A. McGovern, S. Agrawal, and J. Hickey, Global extreme heat forecasting using neural weather models, arXiv:2205.10972.
[58] R. Benedetti, Scoring rules for forecast verification, Monthly Weather Rev. **138**, 203 (2010).
[59] C. van Straaten, K. Whan, D. Coumou, B. van den Hurk, and M. Schmeits, Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in Western and Central Europe, Monthly Weather Rev. **150**, 1115 (2022).
[60] A. Delaunay and H. M. Christensen, Interpretable deep learning for probabilistic MJO prediction, Geophys. Res. Lett. **49**, e2022GL098566 (2022).

[61] A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets* (Springer International Publishing, Cham, 2018).
[62] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, Calibrating probability with undersampling for unbalanced classification, in *Proceedings of the IEEE Symposium Series on Computational Intelligence* (IEEE, Piscataway, NJ, 2015), pp. 159–166.
[63] K. Fraedrich, E. Kirk, and F. Lunkeit, Puma: Portable university model of the atmosphere, Deutsches Klimarechenzentrum 38 (1998).
[64] D. Coumou and S. Rahmstorf, A decade of weather extremes, Nat. Climate Change **2**, 491 (2012).
[65] C. Schär, P. L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. A. Liniger, and C. Appenzeller, The role of increasing temperature variability in European summer heatwaves, Nature (London) **427**, 332 (2004).
[66] V. M. Gálfi and V. Lucarini, Fingerprinting Heatwaves and Cold Spells and Assessing their Response to Climate Change Using Large Deviation Theory, Phys. Rev. Lett. **127**, 058701 (2021).
[67] V. M. Gálfi, V. Lucarini, and J. Wouters, A large deviation theory-based analysis of heat waves and cold spells in a simplified model of the general circulation of the atmosphere, J. Stat. Mech. (2019) 033404.
[68] F. Ragone and F. Bouchet, Computation of extreme values of time averaged observables in climate models with large deviation techniques, J. Stat. Phys. **179**, 1637 (2020).
[69] F. Ragone and F. Bouchet, Rare event algorithm study of extreme warm summers and heatwaves over Europe, Geophys. Res. Lett. **48**, e2020GL091197 (2021).
[70] J. Huang and H. M. van den Dool, Monthly precipitation-temperature relations and temperature prediction over the united states, J. Climate **6**, 1111 (1993).
[71] P. Yiou, Anawege: A weather generator based on analogues of atmospheric circulation, Geosci. Model Dev. **7**, 531 (2014).
[72] P. Yiou and C. Déandréis, Stochastic ensemble climate forecast with an analogue model, Geosci. Model Dev. **12**, 723 (2019).
[73] W. E, and E. Vanden-Eijnden, Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes, Chem. Phys. Lett. **413**, 242 (2005).
[74] L. Onsager, Initial recombination of ions, Phys. Rev. **54**, 554 (1938).
[75] J. Finkel, D. S. Abbot, and J. Weare, Path properties of Atmospheric Transitions: Illustration with a Low-Order Sudden Stratospheric Warming Model, J. Atmos. Sci. **77**, 2327 (2020).
[76] J. Finkel, R. J. Webber, E. P. Gerber, D. S. Abbot, and J. Weare, Learning forecasts of rare stratospheric transitions from short simulations, Monthly Weather Rev. **149**, 3647 (2021).
[77] D. Lucente, J. Rolland, C. Herbert, and F. Bouchet, Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain, J. Stat. Mech. (2022) 083201.
[78] P. Miron, F. Beron-Vera, L. Helfmann, and P. Koltai, Transition paths of marine debris and the stability of the garbage patches, Chaos **31**, 033101 (2021).
[79] Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, and G. Henkelman, Optimizing transition states via kernel-based machine learning, J. Chem. Phys. **136**, 174101 (2012).
[80] J. Strahan, A. Antoszewski, C. Lorpaiboon, B. P. Vani, J. Weare, and A. R. Dinner, Long-time-scale predictions from short-trajectory data: A benchmark analysis of the trp-cage miniprotein, J. Chem. Theory Comput. **17**, 2948 (2021).
[81] E. H. Thiede, D. Giannakis, A. R. Dinner, and J. Weare, Galerkin approximation of dynamical quantities using trajectory data, J. Chem. Phys. **150**, 244111 (2019).
[82] S. Manabe, Climate and the ocean circulation: I. The atmosphere circulation and the hydrology of the Earth's surface, Monthly Weather Rev. **97**, 739 (1969).
[83] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, The ERA5 global reanalysis, Quart. J. Roy. Meteorol. Soc. **146**, 1999 (2020).
[84] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, Phys. Rep. **810**, 1 (2019).

[85] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
[86] D. Chicco and G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. **21**, 1 (2020).
[87] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta Protein Struct. **405**, 442 (1975).
[88] Note1. e.g., Brier depends on unobserved events in multi-class classificaiton problem.
[89] Statistical forecasting, in *Statistical Methods in the Atmospheric Sciences*, 4th ed., edited by D. S. Wilks (Elsevier, Amsterdam, 2019), Chap. 7, pp. 235–312.
[90] G. W. Brier, Verification of forecasts expressed in terms of probability, Monthly Weather Rev. **78**, 1 (1950).
[91] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, On calibration of modern neural networks, in *Proceedings of the 34th International Conference on Machine Learning—Volume 70 (ICML'17)* (JMLR.org, 2017), pp. 1321–1330.
[92] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer Series in Statistics (Springer, New York, NY, 2009).
[93] Note2. "Add up" is used here qualitatively, there is no mathematical reason why skills should actually add up arithmetically.
[94] F. Takens, Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence* (Springer, Berlin, 1981), pp. 366–381.
[95] Note3. In case of 800 years we actually invert validation and training sets that were taken for the benchmark case and perform 10-fold cross validation. In case of 100 years we only sample 10 representative training sets of 100 years and validate on the remainder.
[96] S. J. Pan and Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. **22**1345 (2010).
[97] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput. **29**, 2352 (2017).
[98] J. A. Weyn, D. R. Durran, R. Caruana, and N. Cresswell-Clay, Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models, J. Adv. Model. Earth Syst. **13**, e2021MS002502 (2021).
[99] https://github.com/georgemilosh/Climate-Learning.
[100] M. E. Quemener, "SIDUS", the solution for extreme deduplication of an operating system, Linux J., January (2014).