Applied Computing and Geosciences 23 (2024) 100185

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences

# Long-term temperature prediction with hybrid autoencoder algorithms

J. Pérez-Aracil [a,*], D. Fister [b], C.M. Marina [a,b], C. Peláez-Rodríguez [a], L. Cornejo-Bueno [a], P.A. Gutiérrez [c], M. Giuliani [d], A. Castelleti [d], S. Salcedo-Sanz [a]

[a] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, Madrid, Spain
[b] Institute of Robotics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Maribor, Slovenia
[c] Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain
[d] Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milano, Italy

ARTICLE INFO

*Keywords:*
Autoencoder
Temperature prediction
Hybrid models
Heatwave

ABSTRACT

This paper proposes two hybrid approaches based on Autoencoders (AEs) for long-term temperature prediction. The first algorithm comprises an AE trained to learn temperature patterns, which is then linked to a second AE, used to detect possible anomalies and provide a final temperature prediction. The second proposed approach involves training an AE and then using the resulting latent space as input of a neural network, which will provide the final prediction output. Both approaches are tested in long-term air temperature prediction in European cities: seven European locations where major heat waves occurred have been considered. The long-term temperature prediction for the entire year of the heatwave events has been analysed. Results show that the proposed approaches can obtain accurate long-term (up to 4 weeks) temperature prediction, improving Persistence and Climatology in the benchmark models compared. In heatwave periods, where the persistence of the temperature is extremely high, our approach beat the persistence operator in three locations and works similarly in the rest of the cases, showing the potential of this AE-based method for long-term temperature prediction.

## 1. Introduction

Seasonal and subseasonal climate prediction problems have gained momentum in recent years, mainly due to their importance in the current context of climate change (Masson-Delmotte et al., 2021). One of the most striking effects of climate change is the constant increase in long-term average temperature (global warming) (Seager et al., 2019; Change, 2018), associated, in turn, with an unforeseen increase in extreme climatic events, with fatal economic and social consequences. In this context, prediction problems related to air temperature have become an essential field of study, with applications in very different areas, such as human health (Díaz et al., 2002a,b), natural resource conservation (Bergmann et al., 2016), agriculture (Hatfield et al., 2011) or long-term energy planning (Olabi and Abdelkareem, 2022), among others.

Long-term temperature prediction can be classified as subseasonal forecast (up to 60 days) (SSF) and subseasonal-to-seasonal forecasts (S2S) (Vitart and Robertson, 2018), where many different algorithms, including DL methods, have previously been proposed. However, note that traditional state-of-the-art forecasts for the SSF and S2S prediction problems are dynamic numerical models. These models are usually based on the Navier–Stokes equations, which can consider the momentum, mass, and enthalpy of atmospheric states (Schultz et al., 2021). Dynamic models are reliable tools because they respect the underlying physics up to 10 days in advance. Beyond prediction time-horizons of 10 days in advance, the reliability of dynamic models decreases substantially, mainly due to the chaotic nature of the atmosphere (Krishnamurthy, 2019). However, alternative non-deterministic models can better capture the issue's chaotic nature in some cases. Thus, modern DL models, typically ensemble-based, can supplement or overcome/replace traditional predictions based on numerical models (Schultz et al., 2021; Bhend et al., 2012; Ren et al., 2020).

There are many previous works with different approaches to long-term air temperature prediction, many of them involving Machine Learning (ML) (Sen et al., 2023) or Deep Learning approaches (Salcedo-Sanz et al., 2023). Several previous works have discussed the application of Neural Networks to long-term air temperature prediction problems (Johnstone and Sulungu, 2021), discussing neural network results on a problem of daily mean, maximum, and minimum temperature time series in Turkey. In Yu et al. (2021), a spatio–temporal graph attention network approach was applied, and in Venkadesh et al. (2013), a genetic algorithm is used for the selection of input data in

* Corresponding author.
*E-mail address:* jorge.perezaracil@uah.es (J. Pérez-Aracil).

https://doi.org/10.1016/j.acags.2024.100185
Received 29 March 2024; Received in revised form 24 June 2024; Accepted 1 August 2024
Available online 8 August 2024

an air temperature prediction problem using artificial neural networks. In Abdel-Aal and Elhadidy (1995), different artificial neural networks were applied to a problem of daily maximum temperature prediction in Dhahran, Saudi Arabia. Other ML approaches have also been applied to long-term prediction of air temperature, such as Support Vector Regression algorithms (SVR) (Paniagua-Tineo et al., 2011; Chevalier et al., 2011; Mellit et al., 2013), multi-models based on ML (Ahmed et al., 2020), gradient boosting neural approaches (Peng et al., 2020), neuro-evolutionary algorithms (Gómez-Orellana et al., 2023), recursive networks (Huang et al., 2018) or works where many of these previous algorithms have been compared together (Oettli et al., 2022).

In recent years, there has been a boom in the development of Deep Learning (DL) algorithms, as the next step forward in AI approaches, in many cases improving the performance of classical ML algorithms. This is also the case for long-term prediction of air temperature problems, such as Karevan and Suykens (2020), where a type of LSTM network (transductive LSTM) is applied to the prediction of temperature in Belgium and the Netherlands, or Vos et al. (2021) where a coupling of CNN and LSTM (ConvLSTM) is proposed for the long-range prediction of air temperature. In Fister et al. (2023) different DL techniques with dimensionality reduction techniques have been proposed for long-term temperature prediction problems. Several DL algorithms, such as a Convolutional Neural Network (CNN) with video-to-image translation or CNNs with preprocessing step using recurrence plots, are discussed in that work. Somewhat related to these previous DL approaches for air temperature prediction, in Taylor and Feng (2022) an integrated framework is proposed to predict sea surface temperature. The approach, called Unet-LSTM, was based on the LSTM, and showed mixed prediction skills for predicting two of the past extreme events.

We must note that prediction capabilities are limited when far-sighted out-of-sample predictions. Unfortunately, the chaotic nature of the atmosphere is not the only challenge in predicting climate-related variables. There are other important challenges related to which drivers (such as land/sea, temperature, wind, level pressure, geopotential, etc.) are optimal to carry out the best predictions, and also at which exact geographical locations, and how much time in advance should these drivers be considered. It is known that anti-cyclonic blockage on the Atlantic may heavily divert the usual way of strong winds called jetstreams, e.g., higher altitude winds that separate warmer and colder air areas, which further causes the moving of borders of weather fronts in the following days and hence causes occurrences of extreme (but luckily occasional) weather events (Stendel et al., 2021). Fortunately, both challenges mentioned above can be addressed jointly and effectively using DL techniques. However, it is known that the key climate drivers are not uniform in all extreme events (Wehrli et al., 2019), instead vary from event to event. Seasonal variations play a crucial role. For example, climate events in summer differ significantly from those in winter. Despite this, extreme events theoretically could be more or less reliably spotted within a Prediction Time-horizon (PT) of up to 7 days, while long-persistent extreme events up to 10 days (Dai et al., 2021). Following this line of reasoning, according to Weirich Benet et al. (2023), slow-moving variables such as the North Atlantic Sea Surface Temperature (SST) anomaly, Northwestern Mediterranean SST, geopotential, and precipitation are the variables most correlated with the lagged temperature. Geopotential and precipitation score significant Pearson's correlation coefficients for a time lag of 1 and 2 weeks, while SST anomalies from 4 to 6 weeks. Hence, the SST is one of the most important variables for carrying out lagged predictions 4 weeks in advance (of course, the geographical position of the SST must also be relevant). Thus, He et al. (2021) recommends that the SSF utilise the variables of atmosphere, land, and ocean, such as temperature of 2 m, soil moisture, geopotential height, sea level pressure, relative humidity, and sea surface temperature. Also, additional temporal indices, such as El Niño and North Atlantic oscillation indices, Madden-Julian oscillation indices and sudden stratospheric warning indices, may complement to achieve more reliable predictions. In Weyn et al. (2021), six different two-dimensional variables were used, with a horizontal resolution of 150 km (1.4 deg). Other recent works dealing with DL approaches, such as Grönquist et al. (2021), utilised variables like temperature, geopotential, u and v components of wind, divergence, vertical velocity, and specific humidity to derive a DL-based ensemble model for temperature prediction based on deep neural networks. In da Silva et al. (2022), a hybrid 24-h forecasting model was developed to forecast severe convective weather. This model outperformed the WRF model.

In this paper, we deal with the problem of subseasonal long-term air temperature prediction using DL-based techniques. Specifically, we consider the prediction of air temperature from atmospheric variables with a prediction time horizon larger than one week (1, 2, 3 and 4 weeks), using hybrid ML algorithms based on autoencoders (AEs). Two DL structures are proposed and analysed: a first algorithm uses AEs to learn and detect anomalies, which may help predict temperature in the following weeks. In this case, a first AE is trained in the data acquired from the ERA5 reanalysis (Hersbach et al., 2020), to obtain the patterns associated with the temperature in the zone under study, specified in Section 2.1, and a second AE is used to detect possible anomalies and give a prediction of temperature. The second approach consists of training an AE in the data, as in the first case, and using the latent space generated as the input for a neural network, which will give the final prediction output. Both cases have been tested here in problems of long-term air temperature prediction in seven locations (European cities) where major heatwaves occurred, so we can analyse the behaviour of the prediction system including the heatwave period, where the persistence operator is usually the best option (Lorenz et al., 2010), due to the high persistence of these types of extreme events. We show that the proposed approaches based on AEs are able to beat persistence in three heatwaves and stay close in the other four cases when the prediction is focused on the heatwave period.

## 2. Data description

Climate data are obtained from the ERA5 reanalysis (Hersbach et al., 2020), a reliable source of accurate data from the past, dating back to 1950. Climate data is obtained worldwide in intervals of 0.10 or 0.25 horizontal and vertical resolution. The following variables are used: mean sea level pressure (MSL), sea surface temperature (SST), 2-m temperature (T2M), geopotential height at a constant pressure level of 500 hPa (Z500), and u- (eastward) and v- (northward), 10 m and 100 m (U10, V10, U100, V100). Together, eight different variables are obtained at hourly periods from 1950 to 2022.

The data obtained are then averaged into weekly periods due to the fact that typical heatwaves last 3 to 7 days. Therefore, the temporal resolution of a model is defined as 1 week. This is, on the one hand, a serious shrinkage of temporal resolution, but on the other, it ensures a larger consistency of climate data by removing statistical outliers. This goes in hand with the physical time periods of the phenomena we are addressing, such as rise and fall of cyclone or anticyclone systems, movements of large air masses, large air temperature gradients, etc., are more inclined towards weekly than hourly periods.

In the next step, a Geographic Area Selection (GAS) (Fister et al., 2023) is run to identify the heatwaves in it. The objective of selecting the GAS region is to limit the geographic area included in the model. We expect that common weather patterns that prevail in Europe are heavily patent within the chosen geographic area and that there is less connection with geographic zones outside this area. We describe the identification of heat waves and the selection of GAS in the next two subsections.

**Table 1**
Summary information of the heat waves considered. Note: Locations in degrees latitude and longitude rounded to 1°.

| Heat wave | Duration | Max. °C | Location |
|---|---|---|---|
| Paris 2003 | 01 Aug.–14 Aug. | 39.3 | 49°N, 2°E |
| Córdoba 1995 | 15 Jul.–24 Jul. | 42.8 | 38°N, 4°W |
| Athens 1987 | 18 Jul.–27 Jul. | 44.6 | 38°N, 24°E |
| Frankfurt 2006 | 16 Jul.–30 Jul. | 34.6 | 50°N, 9°E |
| Szczecin 1994 | 21 Jul.–10 Aug. | 34.9 | 53°N, 15°E |
| Sofia 2007 | 15 Jul.–30 Jul. | 44.9 | 43°N, 23°E |
| Smolensk 2010 | 9 Jul.–18 Aug. | 37.3 | 55°N, 32°E |

**Fig. 1.** Selected locations (cities). x-axis = longitude, y-axis = latitude.

### 2.1. Locations of the study

We have considered seven European cities where major heatwaves occurred in Europe since 1950 (Russo et al., 2015; Barriopedro et al., 2023) to illustrate the performance of the proposed approaches. Table 1 shows a detailed description of these locations, including the date of the heatwave. Note that the heatwaves affected different regions of Europe, and displayed different durations (from approximately one week to almost one month) and timings of occurrence within the high summer season (from early July to the second half of August). Table 1 also shows the spatial domains used for the predictor (pressure) and the maximum temperature registered during the heatwave ($T_{max}$). Note that the study carried out was considered as a test set for the long-term temperature prediction in these cities during all the years corresponding to the heatwave, from January to December. The geographical map in Fig. 1 shows the locations of selected cities. Figs. 2(a) and 2(b) show an example of average and maximum temperature in a year of heatwave, in this case for the heatwave of Córdoba 1995.

### 2.2. Geographic area selection

The GAS region considered in this paper is defined on the basis of nonlinear correlation coefficient analysis, which was performed as follows. A Spearman's correlation coefficient is calculated between (1) time series of each of the 8 variables for each geographic location, given at an interval of 1-degree resolution, and (2) time series of T2M at each location. No prediction time horizon is included here. An analysis is done for weekly data between June and July (included) for the years 1950–2020 ($N = 444$). Fig. 3 shows the correlation coefficient analysis for Paris. Dark red indicates a strong positive correlation and dark blue indicates a strong negative correlation. A red box indicates the chosen GAS region for Paris. The chosen GAS region is fixed for all variables and is selected manually as a compromise between (1) the strength of a correlation and (2) the closeness to all points in which the T2M is to be predicted. Areas within the red box are then only supplied with the first step AE; the rest of the information is disregarded.

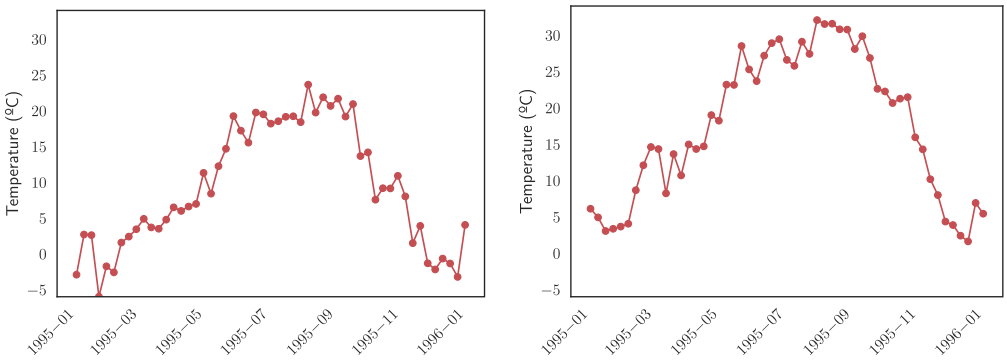
## 3. Proposed methods

This section describes the design and specifics of implemented methodologies. All implemented methods are fed by the input climate data on the weekly basis. The goal of the methods is to predict the T2M for given geographic location. Two different architectures (frameworks), each with two different configurations, are proposed in this paper and employed during the experiments carried out: AE+MLP/AE*+MLP, and AE+AE/AE*+AE*, where the "*" denotes the alternative, i.e., shallower configuration of framework (with less number of trainable weights). Both of these frameworks consist of a common first-stage architecture, i.e. AE, which allows to extract the meaningful features of the input data by reconstructing it. It consists of an encoder and a decoder. The encoder is fed with the input data, and aims to extract the meaningful features, usually reducing the dimensionality of the data. The encoded data is called *latent space*. It feeds the decoder part of the AE, which aims to decode the latent space to reconstruct the input data. A non-symmetrical architecture has been selected for the AE to avoid the extraction of spurious features. Thus, the encoder part is more complex and deeper than the decoder model. A first-stage AE is used in this work to detect anomalies between inputs and outputs, which are then inputted into the second-stage model, i.e., MLP or AE. Effectively, first-stage AE transforms inputs through a narrowing, called latent space, back to the original outputs without considering any PTs (delays) between inputs and outputs. Anomalies are then calculated as deviations between actual images and predicted images by means of statistical learning. Theoretically, the latent space preserves all the most important information necessary to reconstruct the image. Therefore, encoding the input into latent space can be seen as a feature selection/extraction algorithm. Two distinctive second-stage methods are utilised, i.e. MLP and AE (the latter with a similar architecture as the first-stage AE). The MLP is directly connected to the latent space since all important information should be there, and the output of the MLP is wired to the T2M at the desired location. The second stage AE accepts the input of the first stage AE and outputs the T2M for the whole region (continent) rather than a single location. Both the second stage architectures incorporate PTs between inputs and outputs and are therefore able to forecast into the future. In the following, both stages are described in detail.

### 3.1. AE+MLP architecture

First-stage AE architecture is organised to transform multichannel input into the same number of channel outputs, considering no PT between the two. AE transformation consists of two vital elements: (1) reduction of the input images in the latent space (encoding) and (2) extension of the latent space into the output images (decoding), preserving the original size dimension. The obtained latent space is then wired to the MLP network (number of MLP inputs equal to latent space dimension). An MLP is organised to subsequently reduce the number of inputs into a single output neuron, which effectively represents the temperature in a given city with considered PT (PT1 = PT of 1 week, ..., PT4 = PT of 4 weeks). Table 2 summarises the AE architecture, while Fig. 4 shows the AE+MLP architecture graphically. Fig. 5 shows the specific (detailed) AE version used in this paper. Note that only the encoder part is used for forecasting in the case of the AE+MLP.

Another feature we used is the randomised masking of the inputs (the first layer of the AE). The purpose of randomised masking was twofold, i.e. (1) to increase the sample size and (2) to enter the white noise into the model. The sample size was increased due to the effect of obtaining different images with different information (kind of data augmentation DeVries and Taylor, 2017). Injecting white noise on




(a) Weekly average. (b) Weekly maximum.

**Fig. 2.** Cordoba temperature in 1995.

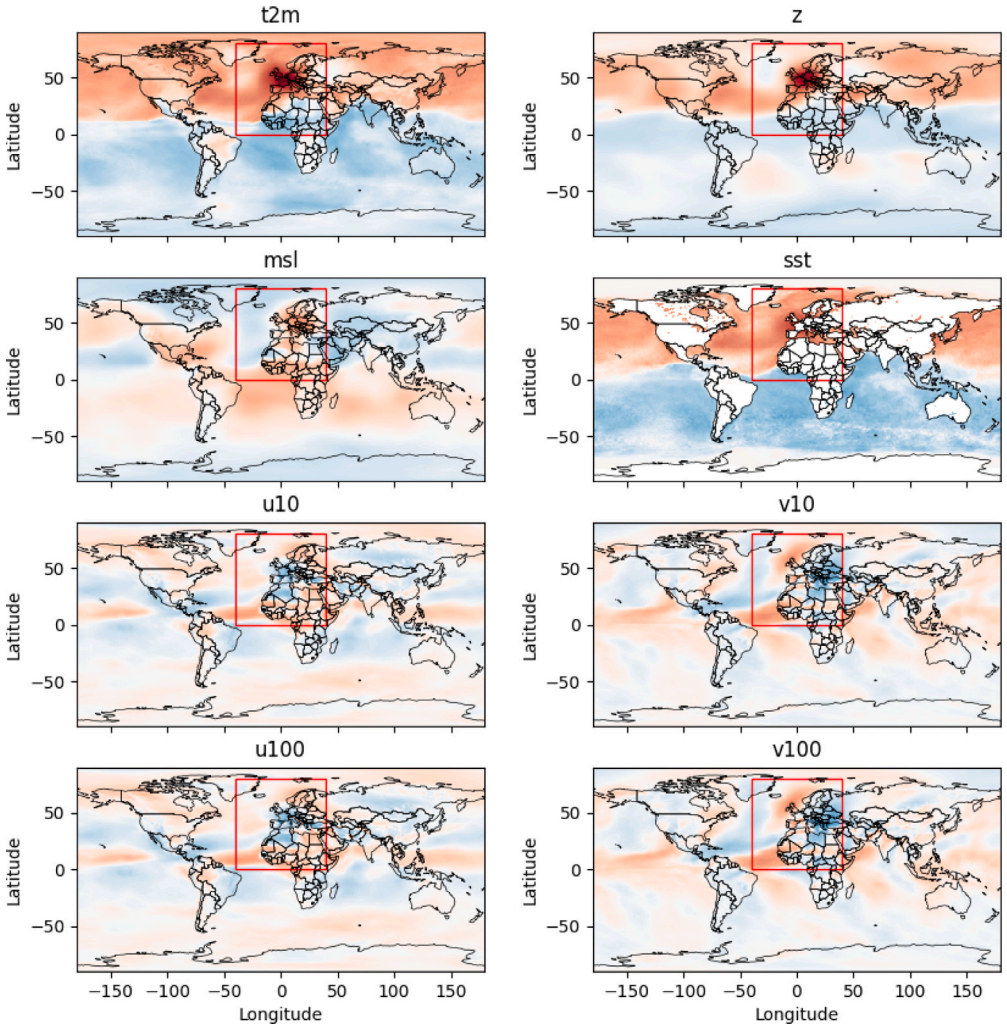



**Fig. 3.** Correlation coefficient results of 8 different variables to T2M for Paris, years = 1950–2000, months = 6–7. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the other improved the learning performance. The broad literature reports that intentionally adding noise to the model inputs smooths the input space, hence suppressing overfitting (Du et al., 2022; Li and Liu, 2016), improving the generalisation skill of the model by regularising it (Bishop, 1995b), and enhancing the robustness of the model (Bishop, 1995a). Overall, the problem becomes easier to learn. Here, a *mask* of 80 × 80 values was generated for each input, with either 0s or 1s, for each dimension separately (see Eq. (1)). The *mask* was then multiplied element-wise with the inputs. Randomised masking was employed for each of the AE architectures separately (for AE+AE,

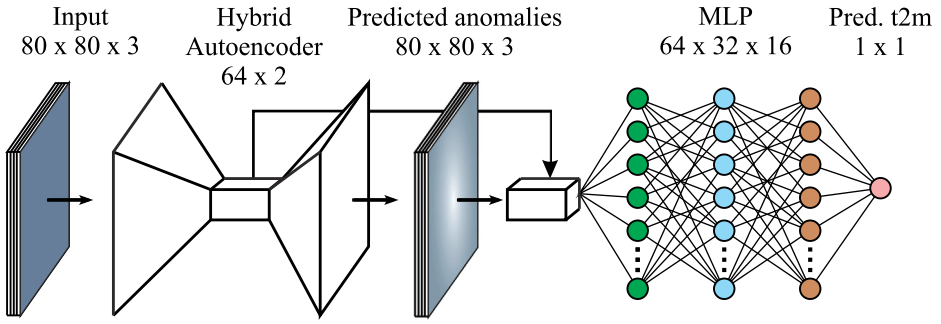



**Fig. 4.** Scheme of the general AE+MLP architecture.

**Table 2**
The hybrid (deep) AE architecture used during experiments. The latent space transformation resembles the VAE architecture, but because utilising the simple loss function it does not qualify as a VAE. Therefore, we call such an architecture hybrid AE. LeakyReLU's $alpha = 0.3$. $x$ = number of output channels control parameter, for the first-stage AE equal to 3, for the second-stage equal to 1. The layer "randomised mask" generates a mask of 80 × 80 either 0s or 1s for each of the 3 dimensions, if $rand(0,1) < 0.5$ and then multiplies this mask element-wise with the inputs. Padding set to "same" for all Conv2D layers.

| Block type | Description | Kernel size | Size of feature maps | Stride |
|---|---|---|---|---|
| Input | | | 80 × 80 × 3 | |
| Randomise mask | | | 80 × 80 × 3 | |
| Down/1 | Conv2D/LeakyReLU | 3 × 3 | 80 × 80 × 32 | 1 |
| Down/2 | Conv2D/LeakyReLU | 3 × 3 | 80 × 80 × 32 | 1 |
| Down/3 | Conv2D/LeakyReLU | 3 × 3 | 80 × 80 × 32 | 1 |
| Down/4 | Conv2D/LeakyReLU | 3 × 3 | 80 × 80 × 32 | 1 |
| Down/5 | Conv2D/LeakyReLU | 3 × 3 | 80 × 80 × 32 | 1 |
| Down/6 | Conv2D/LeakyReLU | 3 × 3 | 40 × 40 × 32 | 2 |
| Down/7 | Conv2D/LeakyReLU | 3 × 3 | 40 × 40 × 64 | 1 |
| Down/8 | Conv2D/LeakyReLU | 3 × 3 | 40 × 40 × 64 | 1 |
| Down/9 | Conv2D/LeakyReLU | 3 × 3 | 20 × 20 × 64 | 2 |
| Down/10 | Conv2D/LeakyReLU | 3 × 3 | 20 × 20 × 128 | 1 |
| Down/11 | Conv2D/LeakyReLU | 3 × 3 | 20 × 20 × 128 | 1 |
| Down/12 | Conv2D/LeakyReLU | 3 × 3 | 10 × 10 × 128 | 2 |
| Flatten | | | 12.800 | |
| Dense/1 | | | 256 | |
| Dropout/1 | Rate = 0.3 | | 256 | |
| Dense/2 | | | 64 | |
| Dense/3 | | | 64 | |
| Dense/4 | | | 64 | |
| Lambda | Dense/3 and dense/4 inputs | | 64 | |
| z_sampling | | | 64 | |
| Dense/5 | | | 256 | |
| Dropout/2 | Rate = 0.3 | | 256 | |
| Dense/6 | | | 100 | |
| Reshape | | | 10 × 10 × 1 | |
| Up/1 | Conv2DTranspose/LeakyReLU | 4 × 4 | 20 × 20 × 128 | 2 |
| Up/2 | Conv2DTranspose/LeakyReLU | 4 × 4 | 40 × 40 × 64 | 2 |
| Up/3 | Conv2DTranspose/LeakyReLU | 4 × 4 | 80 × 80 × 32 | 2 |
| Up/4 | Conv2D/LeakyReLU | 3 × 3 | 80 × 80 × 3 | 1 |
| Up/5 | Conv2D/selu | 3 × 3 | 80 × 80 × 3(1) | 1 |

two separate processes of randomised masking were utilised).

$$mask = \begin{cases} 1, & rand(0,1) < 0.5 \\ 0, & otherwise \end{cases} \quad (1)$$

### 3.2. AE+AE architecture

The first-stage AEs for AE+MLP and AE+AE (as well as AE*+MLP and AE*+AE*) are identical, a unique aspect of our proposed approaches. Therefore, only single first-stage AE/AE* models are built and trained. AE architectures are identical too, as shown in Table 2, Table 3. Only the output dimensions are different and the physical meaning of inputs and outputs is completely different. Both the second-stage AE/AE* as well as the MLP are trained sequentially after training

**Table 3**
The alternative (shallow) hybrid AE* architecture used during experiments. The AE* is less complex than the AE due to the increased stride and less number of channels. LeakyReLU's $alpha = 0.3$. $x$ = number of output channels control parameter, for the first-stage AE equal to 3, for the second-stage equal to 1. The layer "randomised mask" generates a mask of 80 × 80 either 0s or 1s for each of the 3 dimensions, if $rand(0,1) < 0.5$ and then multiplies this mask element-wise with the inputs. Padding set to "same" for all Conv2D layers, except the selu layer which is set to "valid".

| Block type | Description | Kernel size | Size of feature maps | Stride |
|---|---|---|---|---|
| Input | | | 80 × 80 × 3 | |
| Randomise mask | | | 80 × 80 × 3 | |
| Down/1 | Conv2D/LeakyReLU | 3 × 3 | 27 × 27 × 64 | 3 |
| Down/2 | Conv2D/LeakyReLU | 3 × 3 | 9 × 9 × 128 | 3 |
| Flatten | | | 10.368 | |
| Dense/1 | | | 256 | |
| Dropout/1 | Rate = 0.3 | | 256 | |
| Dense/2 | | | 64 | |
| Dense/3 | | | 64 | |
| Dense/4 | | | 64 | |
| Lambda | Dense/3 and dense/4 inputs | | 64 | |
| z_sampling | | | 64 | |
| Dense/5 | | | 256 | |
| Dropout/2 | Rate = 0.3 | | 256 | |
| Dense/6 | | | 81 | |
| Reshape | | | 9 × 9 × 1 | |
| Up/1 | Conv2DTranspose/LeakyReLU | 4 × 4 | 27 × 27 × 128 | 3 |
| Up/2 | Conv2DTranspose/LeakyReLU | 4 × 4 | 81 × 81 × 64 | 3 |
| Up/3 | Conv2D/selu | 2 × 2 | 80 × 80 × 3(1) | 1 |

the first-stage AE/AE*. The first stage AE/AE* is trained once only for all desired PTs, while the second stage AEA/AE* is trained individually for each PT. Therefore, one first-stage AE/AE* and four different second-stage AEs/AEs* are needed to forecast four different PTs.

Fig. 6 shows the AE+AE/AE*+AE* architecture. Instead of connecting the latent space into the MLP, the predicted anomalies are input into the second-stage AE/AE*. The output of the second-stage AE/AE* is the predicted T2M, given for the whole region. The benefit of AE+AE compared to AE+MLP is that it is able to forecast a big-picture outlook instead of a single value, a potential game-changer in our field. The main drawback of this approach is that the AE+AE is inherently more complex, and therefore, more cautious training is required. Fig. 5 shows the specific (detailed) version of the AE used in this paper.

## 4. Experiments and results

To verify the performance of the proposed AE+MLP/AE*+MLP and AE+AE/AE*+AE* methodologies, the temperature prediction analysis was carried out at the seven locations considered for different prediction time horizons (PT, from 1 to 4 weeks in advance). As mentioned above, the experiments were carried out for the cities shown in Table 1, training with data from 1950 to 2022, and testing in the year when the major heatwave occurred in the zone (see Table 1). For example, if the Paris heatwave occurred in August 2003, methodologies were

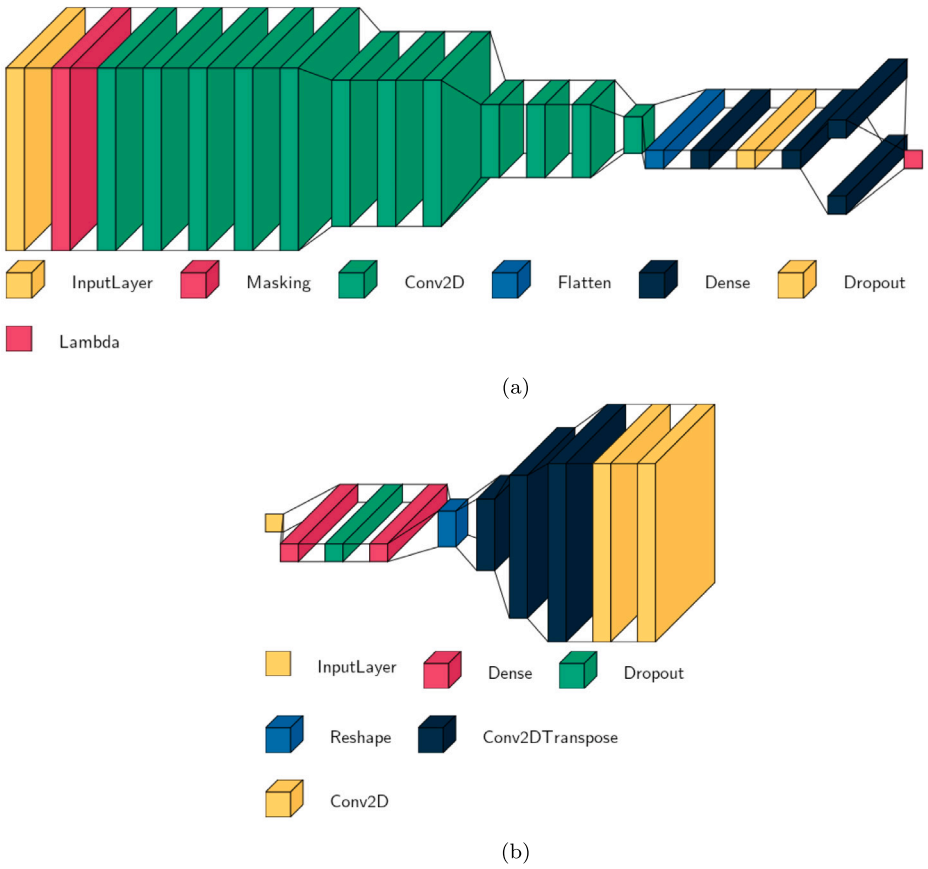



**Fig. 5.** Hybrid AE architecture used during experiments, where (a) is the encoder part and (b) is the decoder part.

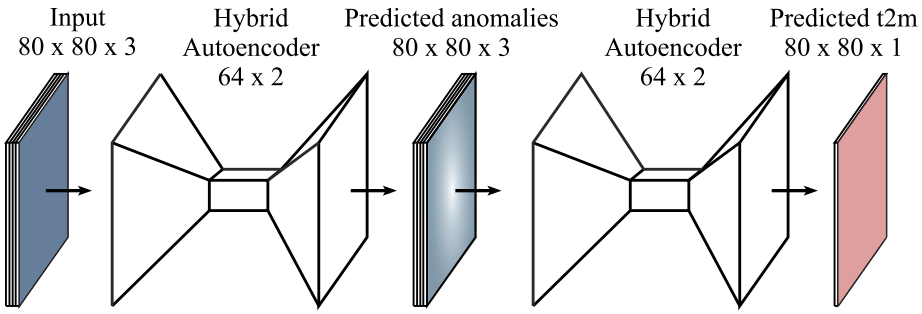



**Fig. 6.** Scheme of the general AE+AE (AE*+AE*) architecture.

tasked to forecast the Paris T2M for the whole year of 2003; the proposed approaches were training with Reanalysis data excluding 2003 and testing in that year. Each experiment forecasted 4 different PTs, varying from the shortest PT1 (1 week) to the longest PT4 (4 weeks in advance). By increasing the PT, the forecast's sample size was reduced by 1 (we have considered 52 weeks per year; 51 forecasts were implemented for PT1 and 48 forecasts for PT4). The proposed hybrid AE-based approaches were compared against two benchmarks (Persistence and Climate Models). The mean squared (MSE) and mean absolute (MAE) errors were recorded annually for each method tested. The goal of the experiments was to show that AE+MLP/AE*+MLP and AE+AE/AE*+AE* can overcome the benchmark models and thus provide better long-term temperature forecasts. Particularly interesting comparison should be between the deep and shallow AE/AE* configurations. During the experiments, the parameter settings described in Table 4, were utilised.

### 4.1. Description of the benchmark models compared

As mentioned, two benchmark models were implemented and utilised, i.e., Climatology and Persistence. The Climatology operator is implemented as follows:

$$C_{(y,w)} = \sum_{t=1950}^{y-1} T2M_{(lat,\ long)}^{(t,w)}, \tag{2}$$

where $y$ represents the year of forecast (given by heatwave occurrence), $w = 1, 2, \ldots, 52$ a weekly index and $(lat, long)$ stands for the geographic

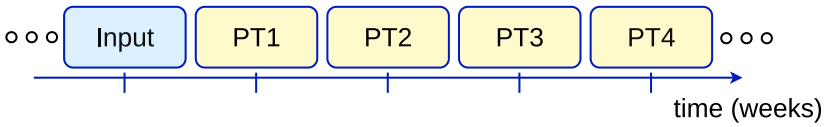



**Fig. 7.** PTs range from 1 to 4.

**Table 4**
Parameter settings for the proposed frameworks. Note: Normalised values of SST for land (originally set to not-a-number) are set to 0.5.

| Parameter | Set value |
|---|---|
| No. of max. epochs | 5000[a] |
| Batch size | 128 |
| Validation split | 15% |
| Input dimension | $80 \times 80 \times 3$ |
| Variables included | z, SST, T2M (normalised) |
| Learning algorithm AE/AE*, MLP | Adam, Adam (Kingma and Ba, 2014) |
| Learning rate AE/AE*, MLP | 0.001, 0.0005 |
| Loss function | 'MSE' |

[a] max. epochs, if no loss function improvement for 100 epochs, the learning is stopped and best weights are restored.

latitude and longitude of the forecasting site. Effectively, Climatology is the arithmetic mean of the T2M's in all previous years (after 1950). The benefit of the Climatology model is that it is stable since it is calculated based on many years. The main drawback of the Climatology operator is that it is only partly able to compensate for the global warming effect and partly able to predict any deviant behaviour (such as heatwaves, rest assured).

On the other hand, persistence is focused on a short-term history (Salcedo-Sanz et al., 2022). Its motivation lies in the presumption that the weather will continue as it was in the near past. Effectively, persistence is implemented as taking the past time series and delaying it for the desired amount of time (defined by PT). In the case of PT1, a T2M at a given location 1 week ago is taken as a result; in the case of PT4, a T2M 4 weeks ago. Eq. (3) shows the persistence operator, mathematically:

$$\mathcal{P}_{(y,w)} = T2M^{(y,w-p)}_{(lat,long)}, \tag{3}$$

where $p = 1, 2, 3, 4$ is the prediction time-horizon (up to 4 weeks, Fig. 7), $y$ stands for the year and $w$ stands for the week of the prediction. The benefit of the persistence operator is that it can account for global warming as it does not incorporate so much history. In addition, the Persistence model is known to be extremely difficult to beat in extreme events, mainly those that last for a long time. Thus, Persistence should perform extremely well in forecasting heatwave temperatures, mainly those with a long last of the event. Of course, the main drawback of the Persistence operator as a forecasting machine is that it cannot predict any deviant behaviour, so it is expected to show a quick underperformance of this benchmark in the previous and posterior weeks of a heatwave.

Note that Persistence and Climatology are basic operators which, in some special cases and problems, are extremely difficult to be beaten by any other approach (including ML). One example is long-term prediction problems or seasonal prediction. In these problems, climatology is usually a very good approach, very difficult to be beaten by alternative approaches, unless they are able to process information extremely well. As previously mentioned, in extreme events such as heatwaves or fog events, the Persistence operator is very difficult to beat. In this paper, we have carried out a comparison with Persistence and Climatology since we are dealing with a problem involving very long-term predictions involving extreme events (heatwaves), in which it is not easy at all to obtain better results than those by Climatology and Persistence.

### 4.2. Results comparison and analysis

In this work, the predictions are generated sequentially, one at a time. The first prediction is done taking the first week of January and generates predictions 1-, 2-, 3-, 4- weeks ahead. For each of the prediction horizons, another model is used (hence 4 different models). Considering 52 weeks within a year, the testing sample size was larger for PT1 (equal to 51) than for the PT4, 48. In practice, this implies that for each week, a new set of predictions is required. No recursive predictions are employed, only actual data.

A statistical comparison among the forecasting methodologies is provided. Two different tables, i.e., one for the MSE (Table 5) and the other for MAE (Table 6), show the results obtained in the experiments carried out in the different locations considered (results for the whole year analysis).

For Athens, the Climatology model performs the best for PTs2-4, while for PT1 the AE*+MLP. A notable increase of the standard deviation of AE+MLP from PT 2 to 4 is observed, while the standard deviation of the AE*+MLP remains steady while increasing the PT. Physically, forecasting the long-term temperature in Athens might be difficult due to its geographic location (close to the sea, where weather patterns are less predictable). However, AE+AE exhibits a very steady performance with a steady standard deviation (except for PT1). The configuration AE*+AE* does not distinguish much from the deep AE+AE.

For Córdoba, the Persistence model exhibits large differences in MSEs and MAEs among different PTs, while the Climatology model performs steadily. AE+MLP exhibits a decreasing but still augmenting performance compared to all other techniques. AE+AE improves the Climatology model for PT1 and PT4, but at a cost of high standard deviation. The AE*+MLP seems as the best option here, both regarding the best mean values as well as lowest standard deviations. Again, no large differences are observed between the AE+AE and AE*+AE*. Here, we see the potential for improving the model through additional coarse and fine training. Córdoba is naturally less affected by the sea than Athens but more affected by mountains on the northwest and southeast sides. Due to large differences compared to previous years, and autocorrelated behaviour, the Persistence model overcomes the Climatology in PT1. However, AE*+MLP is the best model overall in the rest of prediction time-horizons.

Frankfurt and Paris are both mainland cities. Statistical analysis shows that AE*+MLP undoubtedly overcomes all other techniques for all PTs, both in the MSE and MAE metrics, except for Paris PT4. In both cases, the Persistence model exhibits a constant rise of MSEs and MAEs, and Climatology exhibits steady or slightly better performance by increasing the PT.

The city of Szczecin is located near the Baltic Sea, and its temperatures are usually colder than those of the rest of the cities. The performance of the Climatology model is steady among the PTs, whereas the MSEs and MAEs of the Persistence model increase as the PT increases. AE+MLP performs with trendily increasing MSEs and MAEs, while AE+AE performs with trendily decreasing MSEs and MAEs. Still, AE*+MLP performs the best among the proposed techniques, exhibiting extraordinary performance not only for mainland cities but also for seaside cities.

Sofia presents a variate result. The Persistence model is the best for PT1, AE*+MLP is the best for PT2, and Climatology for PTs 3-4. Climatology exhibits a constant downward trend of MSE. The longer

**Table 5**
Results obtained (MSE values, $N = 10$) for AE+MLP/AE*+MLP and AE+AE/AE*+AE* algorithms. Parentheses show standard deviations of methodologies (standard deviation of persistence and climatology methodologies equal to zero).

| | | $\mathcal{P}$ | $\mathcal{C}$ | AE+MLP | AE+AE | AE*+MLP | AE*+AE* |
|---|---|---|---|---|---|---|---|
| Athens | PT1 | 11.721 | 7.227 | 6.455 (0.442) | 7.549 (0.442) | **6.409** (0.348) | 7.648 (0.018) |
| | PT2 | 17.443 | **7.333** | 7.747 (0.153) | 7.722 (0.093) | 7.862 (0.256) | 7.753 (0.037) |
| | PT3 | 24.210 | **6.673** | 7.080 (0.251) | 7.126 (0.065) | 6.928 (0.315) | 7.130 (0.028) |
| | PT4 | 30.370 | **6.811** | 7.452 (0.345) | 7.196 (0.039) | 7.520 (0.318) | 7.206 (0.016) |
| Córdoba | PT1 | 6.449 | 9.116 | 6.427 (0.980) | 8.502 (1.700) | **6.190** (0.751) | 8.663 (0.914) |
| | PT2 | 11.114 | 9.290 | 7.747 (0.707) | 9.836 (1.181) | **7.539** (0.614) | 9.489 (0.916) |
| | PT3 | 15.972 | 9.463 | 8.352 (0.821) | 9.580 (1.208) | **7.956** (0.565) | 8.818 (0.604) |
| | PT4 | 21.666 | 9.660 | 8.362 (0.985) | 9.104 (1.094) | **7.886** (0.703) | 9.689 (1.153) |
| Frankfurt | PT1 | 10.071 | 12.737 | 7.556 (1.601) | 10.721 (1.527) | **7.009** (0.569) | 11.704 (0.780) |
| | PT2 | 18.490 | 12.979 | 10.264 (0.826) | 12.204 (0.690) | **9.949** (0.337) | 11.994 (0.767) |
| | PT3 | 23.318 | 12.986 | 10.864 (0.487) | 11.786 (0.717) | **10.218** (0.317) | 11.710 (0.746) |
| | PT4 | 29.506 | 13.252 | 11.709 (0.382) | 12.256 (0.800) | **11.124** (0.448) | 11.753 (0.668) |
| Paris | PT1 | 12.722 | 10.635 | 7.317 (1.459) | 9.509 (1.357) | **6.891** (0.513) | 9.591 (1.451) |
| | PT2 | 18.962 | 9.699 | 8.333 (0.935) | 9.358 (1.082) | **7.764** (0.443) | 8.795 (0.894) |
| | PT3 | 20.549 | 9.895 | 9.078 (0.602) | 9.458 (0.956) | **8.644** (0.165) | 8.798 (0.702) |
| | PT4 | 25.616 | 9.951 | 9.547 (0.879) | 9.747 (0.981) | 9.509 (0.328) | **9.300** (0.993) |
| Szczecin | PT1 | 8.898 | 8.608 | 5.362 (1.287) | 9.380 (0.454) | **5.156** (0.368) | 9.417 (0.483) |
| | PT2 | 15.044 | 8.513 | 7.373 (0.885) | 9.300 (0.589) | **7.290** (0.418) | 9.085 (0.506) |
| | PT3 | 21.039 | 8.148 | 7.632 (1.067) | 8.671 (0.644) | **7.362** (0.263) | 8.724 (0.459) |
| | PT4 | 28.660 | 8.230 | 8.154 (1.001) | 8.732 (0.512) | **7.955** (0.416) | 8.491 (0.371) |
| Sofia | PT1 | **7.504** | 11.048 | 8.791 (1.847) | 12.599 (1.177) | 8.468 (0.284) | 12.472 (0.893) |
| | PT2 | 15.849 | 10.596 | 10.520 (0.763) | 12.039 (0.579) | **10.428** (0.583) | 11.957 (1.013) |
| | PT3 | 19.476 | **9.785** | 10.830 (0.921) | 11.809 (1.051) | 10.547 (0.424) | 10.893 (0.732) |
| | PT4 | 21.582 | **9.218** | 9.880 (0.737) | 10.879 (0.702) | 9.544 (0.540) | 10.240 (0.633) |
| Smolensk | PT1 | 14.822 | 20.021 | 14.193 (2.688) | 18.284 (0.810) | **13.838** (1.957) | 18.213 (0.954) |
| | PT2 | 31.682 | 20.267 | 17.694 (1.209) | 19.044 (1.043) | **17.445** (0.858) | 18.150 (3.571) |
| | PT3 | 49.326 | 20.049 | 18.554 (0.621) | 18.088 (0.674) | **18.018** (1.232) | 17.859 (0.407) |
| | PT4 | 63.67 | 18.26 | 17.199 (1.004) | **16.471** (1.497) | 17.412 (2.350) | 16.898 (1.354) |

**Table 6**
Results obtained (MAE values, $N = 10$) for AE+MLP/AE*+MLP and AE+AE/AE*+AE* algorithms. Parentheses show standard deviations of methodologies (standard deviation of persistence and climatology methodologies equal to zero).

| | | $\mathcal{P}$ | $\mathcal{C}$ | AE+MLP | AE+AE | AE*+MLP | AE*+AE* |
|---|---|---|---|---|---|---|---|
| Athens | PT1 | 2.91 | 1.989 | 1.938 (0.072) | 2.127 (0.013) | **1.913** (0.052) | 2.122 (0.004) |
| | PT2 | 3.333 | **2.001** | 2.117 (0.028) | 2.144 (0.021) | 2.133 (0.051) | 2.137 (0.008) |
| | PT3 | 4.099 | **1.913** | 2.003 (0.037) | 2.054 (0.009) | 2.000 (0.048) | 2.052 (0.004) |
| | PT4 | 4.723 | **1.948** | 2.093 (0.044) | 2.085 (0.013) | 2.095 (0.033) | 2.077 (0.007) |
| Córdoba | PT1 | **1.953** | 2.514 | 2.107 (0.166) | 2.413 (0.241) | 2.055 (0.146) | 2.451 (0.106) |
| | PT2 | 2.698 | 2.552 | 2.347 (0.079) | 2.617 (0.147) | **2.322** (0.092) | 2.575 (0.113) |
| | PT3 | 3.19 | 2.585 | 2.430 (0.114) | 2.598 (0.151) | **2.382** (0.091) | 2.506 (0.077) |
| | PT4 | 3.83 | 2.638 | 2.437 (0.144) | 2.555 (0.141) | **2.393** (0.108) | 2.637 (0.152) |
| Frankfurt | PT1 | 2.571 | 2.966 | 2.220 (0.233) | 2.688 (0.249) | **2.146** (0.112) | 2.829 (0.107) |
| | PT2 | 3.422 | 3.009 | 2.636 (0.115) | 2.911 (0.086) | **2.599** (0.058) | 2.885 (0.095) |
| | PT3 | 4.019 | 2.998 | 2.718 (0.076) | 2.846 (0.094) | **2.626** (0.057) | 2.832 (0.096) |
| | PT4 | 4.639 | 3.051 | 2.862 (0.054) | 2.917 (0.113) | **2.785** (0.052) | 2.862 (0.089) |
| Paris | PT1 | 2.643 | 2.503 | 2.168 (0.178) | 2.391 (0.170) | **2.134** (0.090) | 2.399 (0.197) |
| | PT2 | 3.39 | 2.402 | 2.263 (0.122) | 2.389 (0.157) | **2.189** (0.069) | 2.311 (0.129) |
| | PT3 | 3.635 | 2.444 | 2.388 (0.083) | 2.419 (0.129) | **2.341** (0.030) | 2.328 (0.095) |
| | PT4 | 4.172 | 2.439 | 2.421 (0.113) | 2.446 (0.139) | 2.427 (0.052) | **2.380** (0.134) |
| Szczecin | PT1 | 2.587 | 2.293 | 1.788 (0.179) | 2.363 (0.070) | **1.756** (0.054) | 2.369 (0.075) |
| | PT2 | 2.984 | 2.266 | **2.040** (0.157) | 2.346 (0.088) | 2.051 (0.052) | 2.306 (0.075) |
| | PT3 | 3.541 | 2.207 | 2.070 (0.136) | 2.239 (0.106) | **2.055** (0.044) | 2.249 (0.074) |
| | PT4 | 4.729 | 2.211 | 2.161 (0.204) | 2.242 (0.085) | **2.125** (0.048) | 2.198 (0.058) |
| Sofia | PT1 | **2.146** | 2.791 | 2.460 (0.222) | 2.898 (0.182) | 2.442 (0.057) | 2.882 (0.142) |
| | PT2 | 3.04 | 2.73 | 2.671 (0.090) | 2.815 (0.096) | **2.663** (0.093) | 2.776 (0.134) |
| | PT3 | 3.595 | **2.641** | 2.737 (0.122) | 2.833 (0.181) | 2.690 (0.074) | 2.668 (0.136) |
| | PT4 | 3.81 | **2.57** | 2.629 (0.117) | 2.729 (0.144) | 2.582 (0.077) | 2.586 (0.134) |
| Smolensk | PT1 | 3.159 | 3.574 | 3.011 (0.300) | 3.369 (0.122) | **2.950** (0.225) | 3.361 (0.139) |
| | PT2 | 4.473 | 3.59 | 3.333 (0.131) | 3.464 (0.142) | **3.287** (0.088) | 3.444 (0.218) |
| | PT3 | 5.6 | 3.55 | 3.395 (0.054) | 3.321 (0.107) | **3.323** (0.115) | 3.282 (0.075) |
| | PT4 | 6.653 | 3.409 | 3.250 (0.092) | **3.186** (0.180) | 3.236 (0.074) | 3.252 (0.176) |

the PT, the better the Climatology performs. Persistence, on the other hand, performs inverted: it performs well for the shortest PT but for longer PTs its performance quickly degrades. Standard deviations of the AE+MLP and AE+AE methodologies for PT1 are unusually high compared to other PTs. Also, performances for longer PTs cannot overcome the Climatology performances. Hence, the reliability of the AE+MLP/AE*+MLP and AE+AE/AE*+AE* performances for Sofia remains questionable.

Smolensk's long-term temperature prediction shows that AE*+MLP provides the best results for PTs, 1-3, whilst AE+AE seems to perform better for PT4. AE+MLP exhibits a steadily decreasing behaviour, while AE+AE shows a trendily decreasing behaviour. Persistence performs strictly rising, while Climatology maintains steady behaviour but is worse than the AE+ techniques in all cases. Interestingly, Smolensk was associated with one of the so-called mega-heatwaves in year 2010.

As a summary of the results obtained, taking into account the geographic locations of the cities considered, the following points can be highlighted: The specific features (geographic factors) of each city where the temperature prediction has been obtained, such as its distance to the sea, mountains, lakes or even deserts, significantly impacts the accuracy of temperature forecasts, mainly when dealing with long-term predictions, as in this case. Cities near large bodies of water, like oceans or lakes, often experience moderated temperature variations due to the thermal inertia of water, resulting in more stable climatic conditions. This moderation can facilitate more accurate predictions, although simpler models may perform well in some cases. For example, in Athens, the maritime influence led to Climatology consistently outperforming other methods, showing the advantage of historical trends in regions with more predictable weather patterns. However, in Szczecin, our AE*+MLP model showed the lowest MAE and MSE across all time horizons, benefiting from the climate moderation by the Baltic Sea. Mountainous regions create microclimates and significant temperature variability due to elevation changes and orographic effects, leading to complex weather patterns that are harder to predict. In Sofia, the increased complexity posed challenges for our models, and persistence and climatology methods often provided better predictions. This fact highlights the advantage of historical trends in regions with significant geographic variability. Inland cities, away from the moderating influence of seas or lakes, typically exhibit more significant temperature extremes and larger seasonal variability. In Frankfurt, for instance, our models, particularly AE*+MLP, performed well across all time horizons, demonstrating their ability to handle continental climate influences effectively. In Smolensk, despite its continental climate and significant seasonal variability, our models showed improved performance over persistence and climatology methods. Mediterranean climates, characterised by hot, dry summers and mild, wet winters, exhibit distinct seasonal patterns influencing model performance. In Córdoba, our models, especially AE*+MLP, outperformed Persistence and Climatology operators in all prediction time horizons. The pronounced seasonality was well captured by our models, leading to lower MAE and MSE values, demonstrating their robustness in handling extreme seasonal contrasts.

### 4.3. Detailed analysis of long-term forecasting during heatwaves

The graphic representations of the errors obtained by each method are shown in Figs. 8 and 9. These figures show the air temperature prediction for each location and each PT (from 1 week to 4 weeks in advance) for the whole year when a major heatwave occurred in the city. PTs are divided into 4 subplots, where each subplot represents a PT of a given horizon, e.g., from PT1 to PT4. Only a single run of the 10 (not necessarily the best one) is shown instead of averaging. Additionally, the periods of each heatwave are marked with straight red lines in the figures to make it easier to discuss the results in that particular period.

It is easy to see that the longer the PT, the more uniform patterns exist between Climatology and both AE+ methodologies. For PT1, some differences between these are noted at initial sight for some cities; while prolonging the PT, fewer differences are observed. The nature of Persistence is completely different; hence, notable and significant differences arise compared to the other 3 methodologies. During the heatwave period, Persistence acts in the following pattern. During the first week of the heatwave, due to the significant change compared to the previous week, the forecasting error is large (usually the highest during 4 weeks). Next, the Persistence error drops to lower values, sometimes even to the negative values. Finally, when the heatwave commences its last phase, the Persistence error is drastically increased and negative, contributing heavily to the annual averaged MSE. The more uniform, non-changing, static, and prolonged heatwaves there are, the better the persistence performs (as in the case of Smolensk, where persistence performs significantly better than any other methodology).

Climatology is a powerful indicator of anomalous temperature behaviour. Where the Climatology is positive, the more-than-average temperatures are present, and vice-versa. Interesting phenomena occur with the AE+ methodologies, as they are trained on the MSE loss function. Their forecasts closely align with the climate forecasts, even during different periods of the year. Some automated mechanism that enables the AE+ to perform very similarly or at moments better than the Climatology must be present in the learning process. We learn that AE+ methodologies are complex and, as such, due to the selection of MSE loss function and given the large-scale dataset, at best, able to capture the mean behaviour between the input–output transformation. Generally, this should give behaviour that is very close or equal to Climatology. Therefore, some overfitting mechanism is needed to ensure additional correctness of outlying forecasts.

Table 7 shows the statistical results limited to the heatwave periods in each location. Here, only a single run (the same run as for the graphics) is accounted for. Average MSE values on the basis of 4 weeks ($N = 4$) are calculated; bold denotes the result with the least error (best result). Note that due to the mathematical approach of the Climatology model, the results are identical.

The empirical work demonstrates that the proposed methods can be reliable alternatives for long-term forecasting of heatwaves or normal periods. They have performed exceptionally well for locations such as those corresponding to Frankfurt, Paris, Szczecin, and Smolensk, all of which experienced severe heatwave intensities. In contrast, Climatology proved to be the superior choice only for Athens, while for other cities, its performance was less impressive. The inherent nature of the Climatology model makes it incapable of accurately forecasting extreme temperatures during heatwaves. The Persistence model beats our proposed model in some cities, but for some PTs, except for Smolensk, in which the Persistence model is clearly the best one. Between the two proposed architectures, the AE+MLP/AE*+MLP emerges as a more favourable alternative. It demonstrates the ability to forecast Córdoba temperatures during the heatwave most accurately for all PTs. It also significantly enhances other methodologies for the case of Frankfurt PT3 and PT4, while AE+MLP appears competitive for Sofia PT1 and PT2.

### 4.4. Discussion

Long-term air temperature forecasting, specially during heatwave periods, is all about (1) forecasting outliers and (2) dealing with imbalanced data samples (datasets). Traditional AI-based algorithms are efficient at capturing the mean behaviour of data samples between the inputs and outputs, but predicting out-of-mean samples can sometimes be problematic. The sequence of two AEs (for AE+AE/AE*+AE*) should produce enough anomalies in the first step that would then enter the second step, enabling forecasts for the magnitude of these anomalies. The two-step architecture was not the uniformly best solution (meaning that other methodologies may perform superior for some scenarios). To combat this, a sensitivity analysis or something similar should be performed to realise the bottleneck. If the susceptibility of the AE handling the input data could be improved, even a single-step AE architecture would be sufficient. Therefore, only a single type of error would be used, which would not be further propagated through the architecture, as is now the case with the AE+AE/AE*+AE*. The clues for potential weather in future do not lie uniformly distributed across the map (as are treated at the moment) but rather at some significant parts of the map, which contribute more to the understanding of future weather

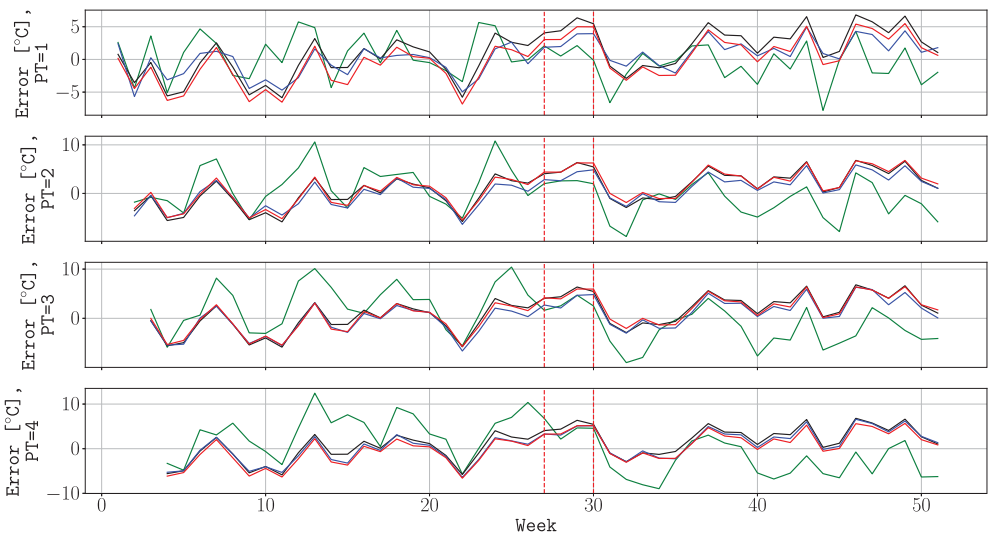



**Fig. 8.** Temperature errors, Frankfurt 2010. Green = persistence, black = climatology, blue = AE+MLP, red = AE+AE, red dotted = heatwave duration. For AE+MLP and AE+AE are displayed one of the runs among the 10 runs (not necessarily the best one). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

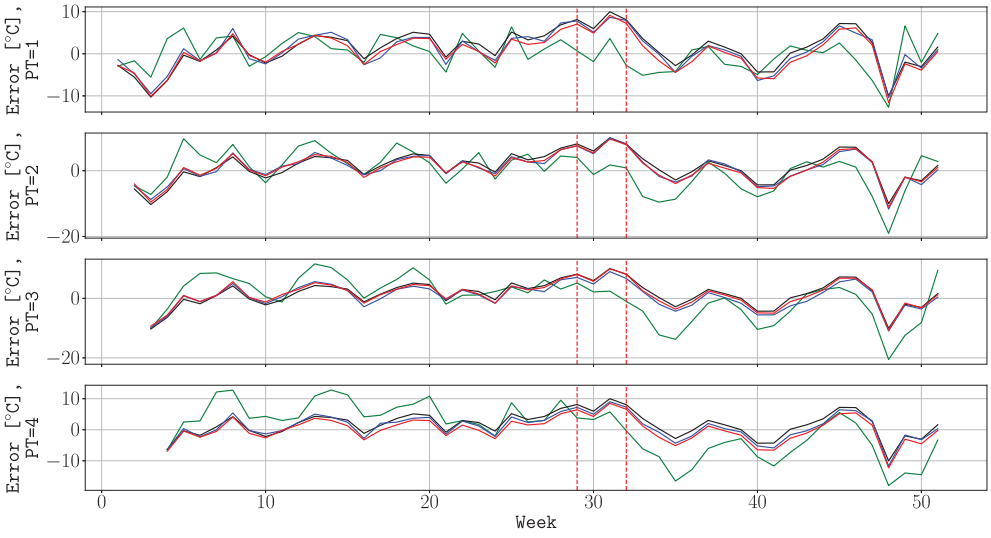



**Fig. 9.** Temperature errors, Smolensk 2010. Green = persistence, black = climatology, blue = AE+MLP, red = AE+AE, red dotted = heatwave duration. For AE+MLP and AE+AE are displayed one of the runs among the 10 runs (not necessarily the best one). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

than others. This phenomenon, which could be named susceptibility (or maybe attention), should be mechanically addressed to a better degree. Generally, it could be enhanced using (1) the single- or multichannel attention mechanisms, (2) alternative loss function, and (3) multi-latent space architecture, among others. Also, hierarchical forecasting could be implemented, where during the first step the input is classified into one of the several classes, e.g., below average temperature, non-significant average or above average temperature. Then, these would be further processed using a specifically allocated AE model.

The proposed AE+ algorithms can obtain good long-term air temperature predictions for all PTs considered (up to four weeks in advance). In most cases, the proposed approaches, specially AE+MLP/AE*+MLP can beat the Climatology and Persistence models, which evaluate long-term prediction in meteorology and climatology applications. The performance of the AE+ models is more similar to that of Climatology than Persistence but improves Climatology in most cities, with some exceptions. The performance of the models considered for long-term temperature forecasts during the weather season is highly dependent on the geographic location of the forecasting site. For example, geographic latitude significantly tailors common weather patterns. In addition, the site's proximity to the seas, mountains, lakes, and deserts profoundly affects its weather, thus the temperature. In Athens, the Climatology approach performs better for prediction time horizons from 2 weeks in advance. The differences are not large, but they denote a behaviour different from other considered cities. Also, in Sofia, where Persistence seems to be important in the prediction, at least for 1-week in advance prediction. Moreover, in Sofia, Climatology seems to be better for 3 and 4 weeks in advance prediction. In the other cities under consideration, the proposed AE+ approaches stand out as the best prediction models. Specifically, AE*+MLP is the best approach in most prediction time-horizons and cities.

An intriguing discussion details the prediction results obtained in the different heatwaves considered, where the Persistence model should obtain the best results. Definitely, this happens in just a single city that was associated with a mega-heatwave, i.e., Smolensk. Otherwise, no universal best heatwave predictor is observed. Persistence is especially not competitive in Athens, Córdoba and Sofia. In the rest of the cities,

**Table 7**
Statistical results (MSE values, $N = 4$) obtained in the heatwave periods for all the methodologies tested.

| | | $\mathcal{P}$ | $\mathcal{C}$ | AE+MLP | AE+AE | AE*+MLP | AE*+AE* |
|---|---|---|---|---|---|---|---|
| Athens | PT1 | 5.043 | **1.815** | 2.820 | 1.994 | 1.965 | 2.150 |
| | PT2 | 7.326 | **1.815** | 2.623 | 1.825 | 1.487 | 2.417 |
| | PT3 | 22.065 | 1.815 | 2.755 | **1.200** | 2.232 | 2.570 |
| | PT4 | 32.353 | 1.815 | 2.149 | 1.893 | **1.538** | 2.516 |
| Córdoba | PT1 | 13.441 | 8.957 | 8.366 | **3.810** | 3.995 | 8.092 |
| | PT2 | 17.147 | 8.957 | 8.486 | 6.921 | **6.647** | 8.066 |
| | PT3 | 25.536 | 8.957 | 8.157 | 6.662 | **6.427** | 8.089 |
| | PT4 | 17.502 | 8.957 | 8.534 | 7.660 | **6.285** | 8.060 |
| Frankfurt | PT1 | **2.256** | 20.169 | 10.926 | 5.770 | 6.271 | 11.144 |
| | PT2 | **4.472** | 20.169 | 20.364 | 8.767 | 11.419 | 12.500 |
| | PT3 | 13.358 | 20.169 | 17.716 | **8.422** | 8.773 | 10.902 |
| | PT4 | 44.920 | 20.169 | 12.091 | 11.816 | **10.074** | 11.636 |
| Paris | PT1 | 20.538 | 32.064 | 31.506 | 35.035 | **20.050** | 31.119 |
| | PT2 | 36.068 | 32.064 | 40.013 | 39.579 | 29.412 | **29.398** |
| | PT3 | **11.242** | 32.064 | 38.576 | 26.678 | 28.967 | 29.268 |
| | PT4 | **14.735** | 32.064 | 34.395 | 34.315 | 26.149 | 32.220 |
| Szczecin | PT1 | **13.694** | 37.962 | 36.247 | 18.877 | 18.003 | 39.006 |
| | PT2 | **9.471** | 37.962 | 40.961 | 28.930 | 23.898 | 33.430 |
| | PT3 | 26.453 | 37.962 | 34.446 | 31.930 | **20.840** | 40.963 |
| | PT4 | 46.764 | 37.962 | 40.653 | 36.213 | **20.333** | 38.410 |
| Sofia | PT1 | 7.974 | 8.231 | **4.844** | 7.887 | 8.587 | 4.854 |
| | PT2 | 18.046 | 8.231 | **6.262** | 6.858 | 7.457 | 9.098 |
| | PT3 | 22.697 | 8.231 | 8.030 | 7.326 | 8.839 | **5.041** |
| | PT4 | 5.268 | 8.231 | 8.083 | 8.358 | 7.803 | **5.077** |
| Smolensk | PT1 | **7.019** | 62.19 | 47.691 | 53.835 | 35.221 | 39.201 |
| | PT2 | **10.121** | 62.19 | 55.037 | 55.894 | 57.107 | 59.821 |
| | PT3 | **11.716** | 62.19 | 59.635 | 48.158 | 55.271 | 60.874 |
| | PT4 | **37.251** | 62.19 | 39.645 | 48.121 | 53.406 | 63.628 |

the Persistence model is better when predicting the heatwave temperature, mainly in the longer-lasting heatwaves (such as Smolensk). In other words, considering temperature predictions throughout the whole year, the AE+AE/AE*+AE* have shown comparable performance than the Climatology model (which was neither the best, neither the worst). The AE*+MLP has shown more superior performance than the AE+MLP. In fact, both the AE+MLP and the AE+MLP disappointed for such kind of predictions, with the AE+MLP showing as the worst of all methods here. The AE+AE seemed to have worked better for longer PTs, such as PT3 and PT4. The heatwave in Smolensk 2010 is considered the largest mega-heat wave of the 21st century (Russo et al., 2015). Its period lasted almost six weeks in a row with maximal temperatures of more than 35 °C. Therefore, the error magnitudes are much higher than for other cities.

With respect to the proposed architecture, AE+MLP seems to be a better option than AE+AE in general. By nature, AE+MLP is much less complex than AE+AE, making it easier to adapt. Tuning a single output specifically for a given forecasting site is inherently a much simpler task than tuning a whole set (quadrant) of outputs. However, it seems that the AE+MLP sometimes misses the big-picture look. The standard deviation of the forecast results (expressed in parentheses) for AE+AE is frequently lower than the standard deviation for AE+MLP, which negatively affects the reliability of the results. The AE+AE, on the other hand, offers better reliability of the results, although at the cost of diminished performance (mean value). It seems especially suitable for forecasting extreme cases. AE+AE calls for an alternative statistical learning approach to improve it for more general cases. The benefit of using the AE+AE is that it offers a whole set of information since it can deliver multiple outputs. However, its performance should be further improved to deliver a performance similar to that of AE+MLP. Here, (1) coarser initial tuning of both stages of AEs, (2) overfitting the second stage AE using the heatwave extracted data seems reasonable options, and (3) implementing a multi-latent space AE structure instead of a single latent space. For future work and research, the AE+AE seems a more interesting alternative to continue with from the point of view of the possibilities to improve its performance.

Finally, note that the long-term prediction of air temperatures from predictive variables (spatio-temporal prediction in this case), including periods of extreme events (heatwaves), is an extremely hard problem, in which AI-based systems hardly obtain results similar to persistence or climatology. In this work, we have shown two deep autoencoder-based systems (data driven) which obtain improvements over persistence and climatology, which means that they are able to exploit the information in data better than previous AI-based systems dealing with data-driven long-term air temperature prediction. Note that the proposed systems can be used to further improve alternative prediction systems based on meteorological models, which consider the physics of the problem, to obtain a much better and operational result. In this paper, we present the system and show its capacity for processing information for long-term air temperature prediction. We do it by comparing persistence and climatology as baseline models in long-term prediction.

## 5. Conclusions

In this paper, we have implemented and tested two different deep learning techniques based on Autoencoders (AEs), to carry out long-term air temperature forecasts with prediction time-horizon up to 4 weeks in advance.

Our study involved the implementation and testing of two different deep learning techniques based on Autoencoders (AEs) for long-term air temperature forecasts. Specifically, we described two hybrid approaches: First, an AE+AE/AE*+AE* approach that combines two AEs, the first one to learn patterns associated with the temperature in the zone, and the second one to detect possible anomalies and obtain the temperature prediction. The second approach consists of an AE, where the obtained latent space is used as the input for a neural network to obtain the temperature prediction. Two different configurations of these architectures, i.e., the deep and the shallow, were implemented. We conducted exhaustive experiments on 7 significant European cities where major heatwaves occurred. The results obtained indicate that the hybrid AE+MLP/AE*+MLP approach is more suitable for long-term temperature forecasting, as it scores better statistical indicators and improves the performance of Persistence and Climatology models in the majority of cities considered. In general, the AE+AE obtains worse results than AE*+MLP, but it seems more appropriate for providing forecasts for a whole region, not only for a single city instead (given by geographic coordinates). Thus, we see future potential for improving the AE+AE performance to achieve at least the same performance as the AE*+MLP in specific locations.

Regarding future lines of work, one promising direction is to incorporate additional climatic variables that can offer a more nuanced understanding of atmospheric conditions. Variables such as humidity, precipitation, solar radiation, and soil moisture could provide critical insights into the interactions influencing temperature anomalies, especially during extreme weather events like heatwaves. The application of transfer learning offers another avenue for improvement. By pre-training models on large, diverse datasets encompassing broad climatic patterns and then fine-tuning them on specific regional data, the models can leverage generalised knowledge and adapt it to local conditions. This approach can significantly enhance model performance and adaptability. Exploring more hybrid modelling approaches that combine AE-based models with other machine learning techniques, such as ensemble methods or physical climate models, could further enhance predictive accuracy. Hybrid models can leverage the strengths of different algorithms, providing a more robust and comprehensive approach to temperature prediction.

## CRediT authorship contribution statement

**J. Pérez-Aracil:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Investigation, Data curation, Conceptualization. **D. Fister:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis. **C.M. Marina:** Writing – original draft, Visualization, Methodology, Investigation, Data curation. **C. Peláez-Rodríguez:** Writing – original draft, Visualization, Methodology, Investigation. **L. Cornejo-Bueno:** Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **P.A. Gutiérrez:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Funding acquisition, Formal analysis. **M. Giuliani:** Writing – original draft, Validation, Supervision. **A. Castelletti:** Writing – original draft, Methodology, Investigation, Funding acquisition. **S. Salcedo-Sanz:** Writing – original draft, Methodology, Investigation, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

This research has been partially supported by the European Union, through H2020 Project "CLIMATE INTELLIGENCE Extreme events detection, attribution and adaptation design using machine learning (CLINT)", Ref: 101003876-CLINT. This research has also been partially supported by the project PID2020-115454GB-C21 of the Spanish Ministry of Science and Innovation (MICINN). The present study has been supported by the European Commission, project Test and Experiment Facilities for the Agri-Food Domain, AgriFoodTEF (grant ref.: DIGITAL-2022-CLOUD-AI-02, 101100622). This research is part of the ENIA International Chair in Agriculture, University of Córdoba (TSI-100921-2023-3), funded by the Secretary of State for Digitalisation and Artificial Intelligence and by the European Union - Next Generation EU. Recovery, Transformation and Resilience Plan.

## References

Abdel-Aal, R., Elhadidy, M., 1995. Modeling and forecasting the daily maximum temperature using abductive machine learning. Weather Forecast. 10 (2), 310–325.

Ahmed, K., Sachindra, D., Shahid, S., Iqbal, Z., Nawaz, N., Khan, N., 2020. Multi-model ensemble predictions of precipitation and temperature using machine learning algorithms. Atmos. Res. 236, 104806.

Barriopedro, D., García-Herrera, R., Ordóñez, C., Miralles, D., Salcedo-Sanz, S., 2023. Heat waves: Physical understanding and scientific challenges. Rev. Geophys. e2022RG000780.

Bergmann, A., Stechemesser, K., Guenther, E., 2016. Natural resource dependence theory: Impacts of extreme weather events on organizations. J. Bus. Res. 69 (4), 1361–1366.

Bhend, J., Franke, J., Folini, D., Wild, M., Brönnimann, S., 2012. An ensemble-based approach to climate reconstructions. Clim. Past 8 (3), 963–976.

Bishop, C.M., 1995a. Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York.

Bishop, C.M., 1995b. Training with noise is equivalent to Tikhonov regularization. Neural Comput. 7 (1), 108–116.

Change, P.C., 2018. Global warming of 1.5° C. World Meteorological Organization, Geneva, Switzerland.

Chevalier, R.F., Hoogenboom, G., McClendon, R.W., Paz, J.A., 2011. Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks. Neural Comput. Appl. 20 (1), 151–159.

Dai, G., Mu, M., Li, C., Han, Z., Wang, L., 2021. Evaluation of the forecast performance for extreme cold events in east Asia with subseasonal-to-seasonal data sets from ECMWF. J. Geophys. Res.: Atmos. 126 (1), 2020JD033860.

DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Díaz, J., Garcia, R., De Castro, F.V., Hernández, E., López, C., Otero, A., 2002a. Effects of extremely hot days on people older than 65 years in Seville (Spain) from 1986 to 1997. Int. J. Biometeorol. 46 (3), 145–149.

Díaz, J., Jordán, A., García, R., López, C., Alberdi, J., Hernández, E., Otero, A., 2002b. Heat waves in madrid 1986–1997: effects on the health of the elderly. Int. Arch. Occup. Environ. Health 75 (3), 163–170.

Du, Y., Shao, W., Chai, Z., Zhao, H., Diao, Q., Gao, Y., Yuan, X., Wang, Q., Li, T., Zhang, W., et al., 2022. Synaptic 1/f noise injection for overfitting suppression in hardware neural networks. Neuromorphic Comput. Eng. 2 (3), 034006.

Fister, D., Pérez-Aracil, J., Peláez-Rodríguez, C., Del Ser, J., Salcedo-Sanz, S., 2023. Accurate long-term air temperature prediction with machine learning models and data reduction techniques. Appl. Soft Comput. 136, 110118.

Gómez-Orellana, A.M., Guijo-Rubio, D., Pérez-Aracil, J., Gutiérrez, P.A., Salcedo-Sanz, S., Hervás-Martínez, C., 2023. One month in advance prediction of air temperature from reanalysis data with explainable artificial intelligence techniques. Atmos. Res. 284, 106608.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., Hoefler, T., 2021. Deep learning for post-processing ensemble weather forecasts. Phil. Trans. R. Soc. A 379 (2194), 20200092.

Hatfield, J.L., Boote, K.J., Kimball, B.A., Ziska, L., Izaurralde, R.C., Ort, D., Thomson, A.M., Wolfe, D., 2011. Climate impacts on agriculture: implications for crop production. Agron. J. 103 (2), 351–370.

He, S., Li, X., Trenary, L., Cash, B.A., DelSole, T., Banerjee, A., 2021. Machine learning and dynamical models for sub-seasonal climate forecasting. In: NeurIPS Workshop on Machine Learning and the Physical Sciences. pp. 1–7.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146 (730), 1999–2049.

Huang, B., Qin, G., Zhao, R., Wu, Q., Shahriari, A., 2018. Recursive Bayesian echo state network with an adaptive inflation factor for temperature prediction. Neural Comput. Appl. 29, 1535–1543.

Johnstone, C., Sulungu, E.D., 2021. Application of neural network in prediction of temperature: a review. Neural Comput. Appl. 33 (18), 11487–11498.

Karevan, Z., Suykens, J.A., 2020. Transductive LSTM for time-series prediction: An application to weather forecasting. Neural Netw. 125, 1–9.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krishnamurthy, V., 2019. Predictability of weather and climate. Earth Space Sci. 6 (7), 1043–1056.

Li, Y., Liu, F., 2016. Whiteout: Gaussian adaptive noise regularization in deep neural networks. arXiv preprint arXiv:1612.01490.

Lorenz, R., Jaeger, E.B., Seneviratne, S.I., 2010. Persistence of heat waves and its link to soil moisture memory. Geophys. Res. Lett. 37 (9).

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., et al., 2021. Climate change 2021: the physical science basis. In: Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press Cambridge, UK, p. 2.

Mellit, A., Pavan, A.M., Benghanem, M., 2013. Least squares support vector machine for short-term prediction of meteorological time series. Theor. Appl. Climatol. 111 (1), 297–307.

Oettli, P., Nonaka, M., Richter, I., Koshiba, H., Tokiya, Y., Hoshino, I., Behera, S.K., 2022. Combining dynamical and statistical modeling to improve the prediction of surface air temperatures 2 months in advance: A hybrid approach. Front. Clim. 4.

Olabi, A., Abdelkareem, M.A., 2022. Renewable energy and climate change. Renew. Sustain. Energy Rev. 158, 112111.

Paniagua-Tineo, A., Salcedo-Sanz, S., Casanova-Mateo, C., Ortiz-García, E., Cony, M., Hernández-Martín, E., 2011. Prediction of daily maximum temperature using a support vector regression algorithm. Renew. Energy 36 (11), 3054–3060.

Peng, T., Zhi, X., Ji, Y., Ji, L., Tian, Y., 2020. Prediction skill of extended range 2-m maximum air temperature probabilistic forecasts using machine learning post-processing methods. Atmosphere 11 (8), 823.

Ren, F., Ding, C., Zhang, D.-L., Chen, D., Ren, H.-l., Qiu, W., 2020. A dynamical-statistical-analog ensemble forecast model: Theory and an application to heavy rainfall forecasts of landfalling tropical cyclones. Mon. Weather Rev. 148 (4), 1503–1517.

Russo, S., Sillmann, J., Fischer, E.M., 2015. Top ten European heatwaves since 1950 and their occurrence in the coming decades. Environ. Res. Lett. 10 (12), 124003.

Salcedo-Sanz, S., Casillas-Pérez, D., Del Ser, J., Casanova-Mateo, C., Cuadra, L., Piles, M., Camps-Valls, G., 2022. Persistence in complex systems. Phys. Rep. 957, 1–73.

Salcedo-Sanz, S., Pérez-Aracil, J., Ascenso, G., Del Ser, J., Casillas-Pérez, D., Kadow, C., Fister, D., Barriopedro, D., García-Herrera, R., Giuliani, M., et al., 2023. Analysis, characterization, prediction, and attribution of extreme atmospheric events with machine learning and deep learning techniques: a review. Theor. Appl. Climatol. 1–44.

Schultz, M.G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L.H., Mozaffari, A., Stadtler, S., 2021. Can deep learning beat numerical weather prediction? Phil. Trans. R. Soc. A 379 (2194), 20200097.

Seager, R., Cane, M., Henderson, N., Lee, D.-E., Abernathey, R., Zhang, H., 2019. Strengthening tropical Pacific zonal sea surface temperature gradient consistent with rising greenhouse gases. Nature Clim. Change 9 (7), 517–522.

Sen, D., Huseyinoglu, M.F., Günay, M.E., 2023. Prediction of global temperature anomaly by machine learning based techniques. Neural Comput. Appl. 1–14.

da Silva, Y.U., França, G.B., Ruivo, H.M., de Campos Velho, H.F., 2022. Forecast of convective events via hybrid model: WRF and machine learning algorithms. Appl. Comput. Geosci. 16, 100099.

Stendel, M., Francis, J., White, R., Williams, P.D., Woollings, T., 2021. The jet stream and climate change. In: Climate Change. Elsevier, pp. 327–357.

Taylor, J., Feng, M., 2022. A deep learning model for forecasting global monthly mean sea surface temperature anomalies. arXiv preprint arXiv:2202.09967.

Venkadesh, S., Hoogenboom, G., Potter, W., McClendon, R., 2013. A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks. Appl. Soft Comput. 13 (5), 2253–2260.

Vitart, F., Robertson, A.W., 2018. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. Npj Clim. Atmos. Sci. 1 (1), 3.

Vos, E.E., Gritzman, A., Makhanya, S., Mashinini, T., Watson, C.D., 2021. Long-range seasonal forecasting of 2m-temperature with machine learning. arXiv preprint arXiv:2102.00085.

Wehrli, K., Guillod, B.P., Hauser, M., Leclair, M., Seneviratne, S.I., 2019. Identifying key driving processes of major recent heat waves. J. Geophys. Res.: Atmos. 124 (22), 11746–11765.

Weirich Benet, E., Pyrina, M., Jiménez-Esteve, B., Fraenkel, E., Cohen, J., Domeisen, D.I., 2023. Sub-seasonal prediction of central European summer heatwaves with linear and random forest machine learning models. Artif. Intell. Earth Syst. 1–52.

Weyn, J.A., Durran, D.R., Caruana, R., Cresswell-Clay, N., 2021. Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. J. Adv. Modelling Earth Syst. 13 (7), e2021MS002502.

Yu, X., Shi, S., Xu, L., 2021. A spatial–temporal graph attention network approach for air temperature forecasting. Appl. Soft Comput. 113, 107888.