Feature selection for data-driven seasonal forecasts of European heatwaves

Check for updates

Ronan McAdam [1] ✉, Jorge Pérez-Aracil [2,3], Antonello Squintu [1], César Peláez-Rodríguez[2], Felicitas Hansen[4], Verónica Torralba[5], Harilaos Loukos [6], Eduardo Zorita[4], Matteo Giuliani [1,7], Leone Cavicchia [1], Sancho Salcedo-Sanz[2] & Enrico Scoccimarro [1]

The early warning of heatwaves using seasonal forecasting systems has the potential to mitigate economic losses and risk to life. Because of the limited reliability and computational expense of dynamical forecasting systems, efforts in recent years have turned to exploiting Machine Learning. Here, an inexpensive approach to forecasting summer heatwaves over Europe is developed, using an optimisation-based feature selection framework to detect a combination of variables, domains and time-lags used to skilfully predict heatwaves. The purely data-driven forecasts are shown to match, and in places outperform, the skill of the state-of-the-art dynamical multi-model products. Moreover, low skill over Scandinavia, a long-term issue common to most dynamical systems, is improved in our data-driven approach. This work also highlights that the greatest contribution to skill comes from predictors at 4-7 weeks time-lag (e.g. mid-March), and identifies predictors which can form the basis for future studies on mechanisms.

Heatwaves (HWs) are prolonged periods of extreme temperature, and lead to a wide range of impacts, including the collapse of agricultural yields[1], drastic increases in energy usage[2], impacts on human health, and increased mortality[3,4]. Europe has experienced devastating heatwaves in the past decades, including, but not limited to, deadly events in 2003 and 2010[5] and more recently in 2022[6,7]. Climate projections suggest further intensification of HWs in the coming decades[8], which will likely lead to an increase in deaths attributed to extreme heat unless mitigation measures are implemented[9]. Consequently, the ability to predict extreme summer heat several months in advance provides an opportunity for the agricultural industry and national health services to implement mitigation measures[3,10].

Seasonal forecasting, the prediction of seasonal climate conditions several months in advance, has the potential to provide society with time and necessary information to take meaningful action prior to potentially damaging climate events. Such information already serves as the foundation for climate services in various sectors, such as the early warning of droughts for agriculture[11,12] and snow cover for tourism[13]. The generation and maintenance of seasonal forecasts, however, is an enormous computational undertaking, with many centres around the world producing dozens of ensemble members of climate models, each of which couples several components of the climate system, at horizontal resolutions typically less than 1 geographical degree. Moreover, predicting heatwaves beyond the deterministic limit (roughly 10–15 days) is challenging[14–17]. The state-of-the-art operational seasonal forecasting systems from the Copernicus Climate Change Service (C3S) have demonstrated reliable forecast skill over large parts of Europe when predicting, up to 3 months in advance, seasonal heatwave indices[18,19]. However, skill gaps remain; for example, over northern Europe, which is less influenced by more predictable mid-latitude variability[20].

The increasing use of Machine Learning (ML) for weather and climate science applications provides a means of reducing the resources required to make accurate weather forecasts without compromising on skill, as demonstrated in the field of weather forecasting[21]. Dynamical forecasting systems, named after their ability to solve dynamical and thermodynamic equations numerically, are now matched in skill or even outperformed by purely data-driven approaches using a range of machine or deep learning architectures[22–26]. These data-driven approaches leverage techniques designed to identify relationships between multiple variables from large datasets of observations or model simulations. The current frameworks used are being adjusted for subseasonal[27] and seasonal timescales[28], and any data-driven approach to seasonal forecasting requires considerably more data than is available in the observational records[29].

Prior to the widespread use of ML, statistical seasonal forecasting showed that the input of known predictors, such as soil moisture for

European heatwaves, into simple statistical models could provide skill for the prediction of certain climate variables on seasonal timescales[30,31]. Previously, such methods relied on the selection of known drivers and thus were limited by the current scientific understanding of heatwave dynamics. Nowadays, more sophisticated ML models are used and are able to select from a set of potential predictors, as widely demonstrated on the sub-seasonal timescale[32–36]. ML-based seasonal forecasting techniques have shown that spatially and temporally distant predictors of seasonal climate can be identified and used to make accurate predictions[37–41]. However, studies either do not include comparisons to operational systems or use a restricted number of pre-selected and known predictors in their feature selection. Moreover, there is currently no purely data-driven seasonal forecast approach for Europe, nor one that focuses explicitly on temperature extremes.

This study describes a data-driven seasonal forecast system that is computationally inexpensive, provides scientifically relevant information on HW predictors, and is shown to match, and in some instances outperform, the state-of-the-art of operational dynamical seasonal forecasting. This work merges efforts in previous statistical and ML-based approaches with training based on a multi-millennial paleo-simulation dataset. Crucially, it employs a feature selection method that boasts the freedom to identify optimal predictor variables and the time-lags over which they contribute to skill. This framework provides an index-specific forecast and driver detection for summer heatwave propensity at any location.

Results

Feature selection of heatwave predictors

This study begins with a feature selection framework, designed to identify the combination of predictors that provides the optimal seasonal forecast skill of European summer heatwave indicators (Supplementary Fig. 1). The chosen potential predictors describe atmospheric, land and ocean conditions which are known to influence the European climate or extremes. First, a range of dimension-reduced predictors is defined using an enhanced version of k-means clustering (which employs a weight of 5% for distances on the geoid) applied to variables known to impact European summer climate (e.g. soil moisture, sea ice content; Supplementary Fig. 2; Supplementary Table 1; "Methods" section). The target in this study is the number of days in which the temperature exceeds the climatological 90th percentile between May and July (MJJ NDQ90). To identify the most influential predictors, a multi-method ensemble optimisation algorithm[42] is employed to select the variables and the corresponding range of time-lags that provide optimal forecast skill of the target. The multi-method ensemble[43] tests various combinations of predictors, and aims to reduce the forecast error; the optimisation algorithm combines various subsets of variables and time-lags into a Logistic Regression model to predict NDQ90 (Fig. 1). The framework benefits from a paleoclimate simulation of the years 0–1850 with a coupled atmosphere-ocean model (hereafter "past2k"), which provides long-term simulated data of predictors and HWs in a stationary climate. The optimisation-based feature selection is performed using a training period of years 0–1600, and a test period of 1601–1850. Finally, when applied to the modern 1993–2016 period with ERA5 predictors, the optimal predictors are used to train ML-based prediction models to provide fully data-driven seasonal forecasts of heatwave occurrence. The optimisation is performed individually for each grid point; see Supplementary Figs. 3 and 4 for an example of the optimal predictors selected and the corresponding seasonal forecasts of the test period.

The European-scale view of optimised forecast skill (root-mean-squared-error normalised by the interannual variability of the target, N-RMSE; "Methods" section; Fig. 2) demonstrates a zone of low skill (N-RMSE >1) stretching from northern central Europe and Scandinavia, while the highest skill is found over central Europe, and the Mediterranean and Black Sea basins. Two grid points representative of either relatively "poor" or "good" regions of forecast skill (Fig. 1) show that the degree of possible improvement, relative to the initial first guess, depends on location. In the "poor" example, optimisation leads to an improvement of 0.18 in N-RMSE,
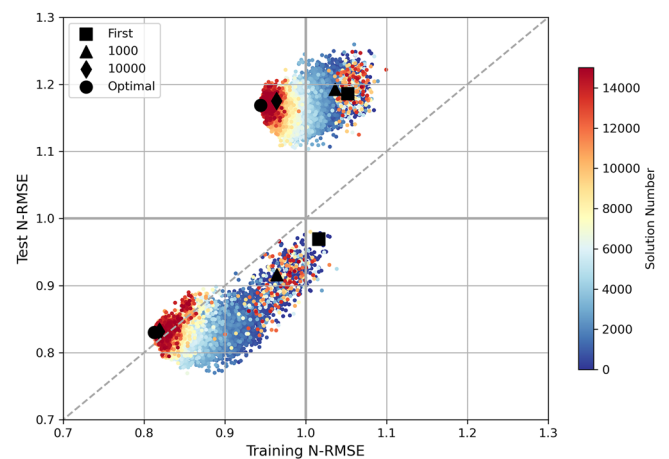

Fig. 1 | Optimisation of seasonal forecast skill (N-RMSE) of seasonal European heatwave indicators (NDQ90) using data from a paleoclimate simulation. The training period and test period are 0–1650 and 1601–1850 respectively. Two examples of optimisation are shown: 43.13° E, 58.76° N (poor, upper cluster of solutions) and 24.38° E, 41.97° N (good, lower cluster). Symbols and colours represent stages of optimisation (latest stages in red), with the "optimal" solution (black circle) corresponding to the solution with the lowest training N-RMSE. The diagonal (dashed grey) represents a perfect fit between training and test scores, while the vertical and horizontal lines at N-RMSE = 1 indicate where the error is equivalent to interannual variability (Supplementary Fig. 5).

while in the "good" example, the improvement is 0.29. Although in both examples the optimal training N-RMSE obtained is below 1 (0.94 and 0.78), indicating that error is within the range of interannual variability, the same is not true for the test period in the "poor" example. The European pattern of data-driven skill in the model world (Fig. 2) resembles the skill of dynamical seasonal forecast systems in predicting temperature[44,45] and its extremes[18]. By applying the data-driven approach first to the model world, we isolate where the use of reduced-dimensionality predictors provides insufficient predictability. If the framework cannot recreate the paleoclimate model training data, then it is highly unlikely to perform well on the test data or in real-world forecasting.

Collecting the optimal predictors from all individual points across Europe (see Supplementary Fig. 3 for an example) provides an overview of the model-world HW drivers at a regional level (Fig. 3). The most commonly selected variables across the domain are the European soil moisture, temperature and geopotential height (z500) clusters. The identified key role of these local predictors agrees with studies of many HWs that have occurred in Europe[15,30]. Commonly selected predictors that represent more distant precursors include sea surface temperature (SST) over the equatorial Pacific and outgoing longwave radiation (OLR) over the tropical Atlantic. While the former represents the phase of the El Niño Southern Oscillation, which is known to play a role in European climate extremes[46], the contribution of the OLR over the tropical Atlantic is not obvious. The feature selection also allows us to study the time-lags in which the variables play a role. The most frequently selected time-lags occur on average around six weeks prior to initialisation (i.e. mid-March; Fig. 3, Supplementary Fig. 6). However, the key temporal lag depends on the variables. Temperature and z500 clusters are selected more frequently in the few weeks prior to initialisation and decay gradually with longer lag, while soil moisture and sea ice selection peak between 7 and 8 weeks prior to initialisation. The number of predictors selected before February is negligible. The most commonly selected short-term European predictors are unsurprisingly selected within or near to the area they represent (Supplementary Fig. 7); TMXEur-1 at 1 1-week time-lag is selected across central Europe, while SSTMed-3 at a 5-week time-lag is largely selected around the Black Sea. For the geographically distant predictors, the tropical Atlantic OLR (OLRTro-2 at 4-week lag) is chosen as a predictor for
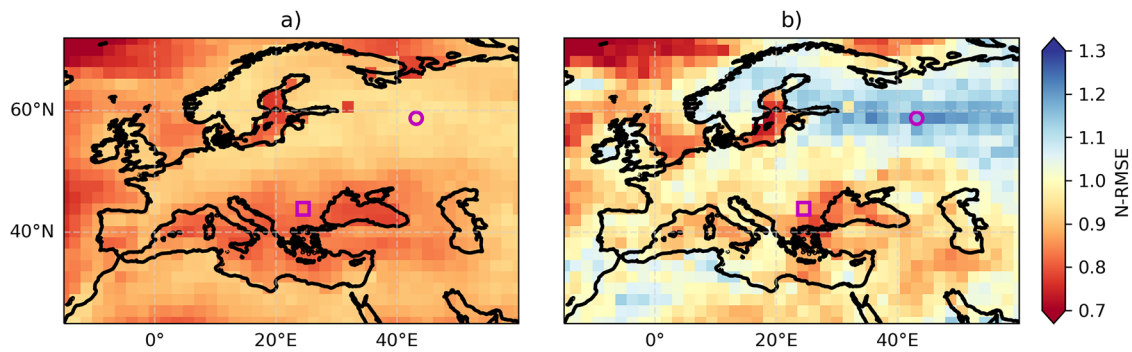
Fig. 2 | Seasonal forecasts of paleo-climate heatwaves across Europe. Optimised seasonal forecast skill (N-RMSE) of seasonal European heatwave indicators (MJJ NDQ90) using data from a paleoclimate simulation for the training period (**a**, 0–1600) and optimisation test period (**b**, 1601–1850). Each point represents the optimal forecast skill based on the point-specific optimisation of predictors (e.g. Fig. 1). The locations corresponding to the good and poor skill examples from Fig. 1 are represented by a square and a circle, respectively.
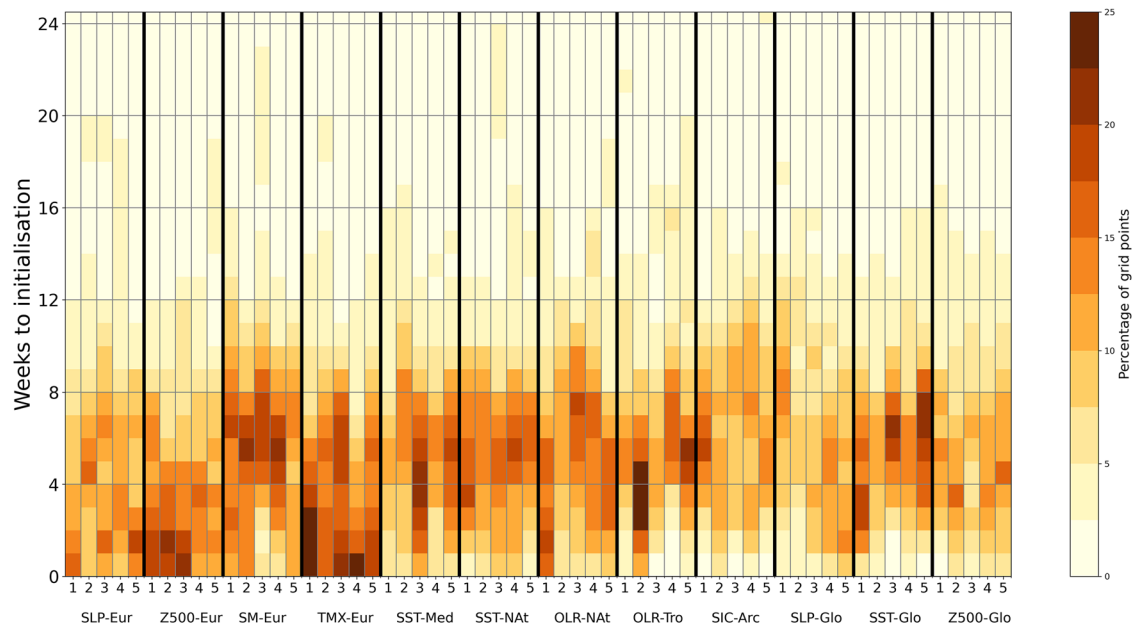

Fig. 3 | Feature selection of predictors for seasonal heatwave indicators across Europe. The matrix displays the percentage of grid points with respect to the entire European domain in which the cluster and time-lag appear in the optimal solutions. Initialisation is on May 1st. Variable labels: mean sea level pressure (SLP), geopotential height at 500 hPa (z500), soil moisture (SM), daily maximum 2 m temperature (TMX), sea surface temperature (SST), outgoing longwave radiation (OLR), and sea ice concentration (SIC). Cluster maps are shown in Supplementary Fig. 2.

areas across the Barents Sea and Scandinavia, as well as parts of the southern Mediterranean Sea, while the selection of the tropical Pacific SST is more sporadic. Studying these points in a SHAP (SHapley Additive exPlanations[47]; "Methods" section) analysis, used to quantify their relative contribution to forecasts, confirms that the local predictors carry more predictive value, while the more distant predictors contribute more weakly (Supplementary Fig. 7). While distant teleconnection-based predictors would be expected to have less direct impact on HW occurrence and therefore a relatively low predictive power according to the SHAP analysis, we cannot rule out that such drivers are present only in the model world. However, their widespread selection, in particular, that of OLR, suggests they are not spurious results and merit further analysis in future studies.

Data-driven seasonal forecasts of European heatwaves
Using the selected optimised predictors for each grid point, the data-driven forecast system is adjusted to be trained on the entire past2k simulation period (0–1850) and then tested for the period 1993–2016 using predictors from ERA5. Whereas the optimisation-based feature selection training and testing is performed with linear regression to avoid large computational cost, the real-world forecasts use ML models in an attempt to boost the skill achieved by the same predictors (e.g. Random Forest; Supplementary Fig. 8). In the "Methods" section, skill for each ML model used is reported; here, the skill graphs reflect the best performing model for each grid point. Data-driven re-forecasts display significant correlation skill scores over 56% of the European domain, including central Europe and the Mediterranean Basin. The skill patterns in the data-driven forecasts (Fig. 4) match those of the optimisation test period (Fig. 2), indicating the successful transfer of learning from the paleoclimate.

Existing operational systems from the C3S represent the state-of-the-art of dynamical seasonal forecasting and can also provide forecasts of summer heatwaves, but the skill has previously only been tested for ECMWF-51[18]. Individual systems (CMCC-35, MF-8, DWD-21 and ECMWF-51, Supplementary Table 2) display similar patterns of skill, such as the zone of low skill extending across Scandinavia and northern central Europe (Supplementary Fig. 9). A multi-model mean is often
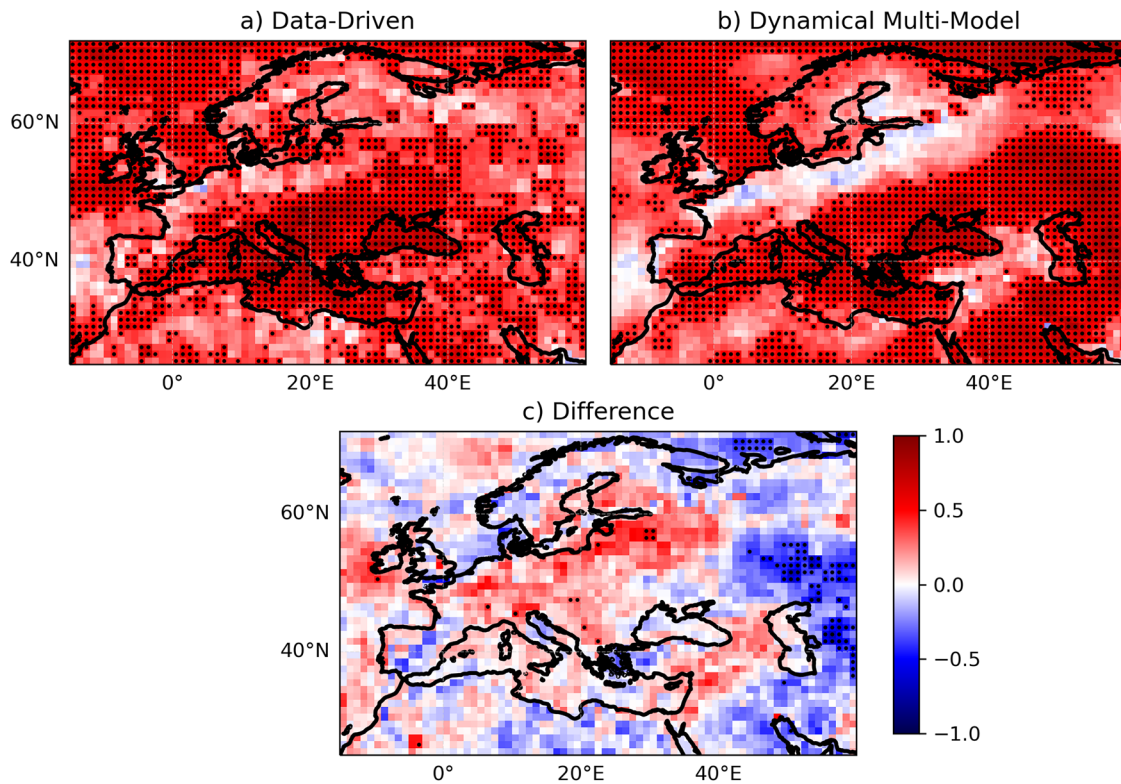
Fig. 4 | Data-driven seasonal forecast skill of European summer heatwaves. Correlation skill score of seasonal European heatwave indicators (MJJ NDQ90) in the data-driven (**a**) and C3S multi-model ensemble (**b**) forecasts over the forecast test period 1993–2016, validated against ERA5. Black stippling represents statistically significant correlation (**a** & **b**) or correlation difference (**c**) at the 95% confidence interval.

used in dynamical forecasting to smooth out errors in individual systems and boost forecast skill; this holds true for forecasts of NDQ90 for which the multi-model product provides statistically significant skill over 58% of Europe (Fig. 4). As a result, over the majority of the domain, there is no statistically significant difference between the data-driven and dynamical skills (with the exception of western Russia; Fig. 4c). Therefore, regions that are skilfully predicted by the dynamical system are also skilfully predicted by the data-driven system. Moreover, the skill in the zone extending over northern central Europe and Scandinavia, a well-known issue in dynamical systems, is also higher in the data-driven approach. When compared to the individual systems (Fig. 5), the data-driven approach is more skilful over certain areas, such as over Eastern Europe when compared to CMCC-35 and ECMWF-51, the previously mentioned Scandinavian zone in MF-8, and over France in DWD-21. However, the skill increase is rarely statistically significant. To predict summertime HWs over Europe, the data-driven approach is as capable as the state-of-the-art multi-model dynamical product and, in some places, more skilful than individual operational forecasting systems.

It is crucial for newly proposed systems to demonstrate skill in forecasting the most exceptional events and, crucially, the data-driven forecasts display this capability in some cases (Fig. 6). In northern Italy, where both data-driven and dynamical systems generally display high skill, the top-performing models in data-driven system forecasts are remarkably close to the observed values for the two years with the greatest number of HW days (2003 and 2015). In this region, a simple linear regression model is as effective as ML-based models and outperforms the dynamical systems (Supplementary Fig. 8), while some display stronger biases in NDQ90 (e.g. Light Gradient Boost - DD-LGB) than others. The extent of the HW in 2003 across western Europe and the Mediterranean basin is also well forecast by the data-driven approach (Supplementary Fig. 10), although the exceptionally deadly event of 2010 over western Russia was not predicted by either type of forecast[20].

Discussion
Although dynamical forecasts of heat extremes display skill over much of Europe[18,19], the zone of low skill over northern central Europe and Scandinavia is a problem that has persisted despite continued updates to dynamical systems[44,45,48]. Recent efforts have demonstrated that hybrid dynamical-ML approaches, in which only ensemble members that best represent the North Atlantic Oscillation are selected, can boost dynamical forecast skill of summer conditions in this region[49]. The purely data-driven approach described here also achieves the goal of improving upon dynamical systems (Figs. 4 & 5), with the benefit of doing so at a considerably lower cost by identifying region-specific predictors. Although Linear Regression displays very high skill in Central Europe and the Mediterranean basin, ML models such as Random Forest and Light Gradient Boosting have shown greater skill across Europe as a whole.

The computational expense of the data-driven approach is very low. For each grid cell, the optimisation of predictors requires roughly 1 CPU-hour (on the DKRZ Levante BullSequana XH2000 supercomputer with 3rd generation AMD EPYC CPUs), and the forecasts require only minutes; scaling to cover the 1° rectangular grid of C3S over Europe (1066 grid points) requires roughly 1000 CPU-hours in total. The optimisation-based feature selection is required only once per start date. Here, the data-driven system was initialised in May by choosing predictors prior to May 1st, and was applied to forecast an HW index of May–June–July. Unlike a dynamical system designed to output many variables at many start dates, our approach focuses on a specific task. In the future, the system can be easily re-optimised for other start dates, target dates and even other extreme events, and can include other potential predictors; for example, humidity for nighttime heatwaves[50].

The predictors identified in the climate simulation (Fig. 3) are not necessarily equivalent to those in the real world, especially given that the training dataset past2k may present biases in both predictor and target variables. Moreover, drivers may change over time, and past2k may provide
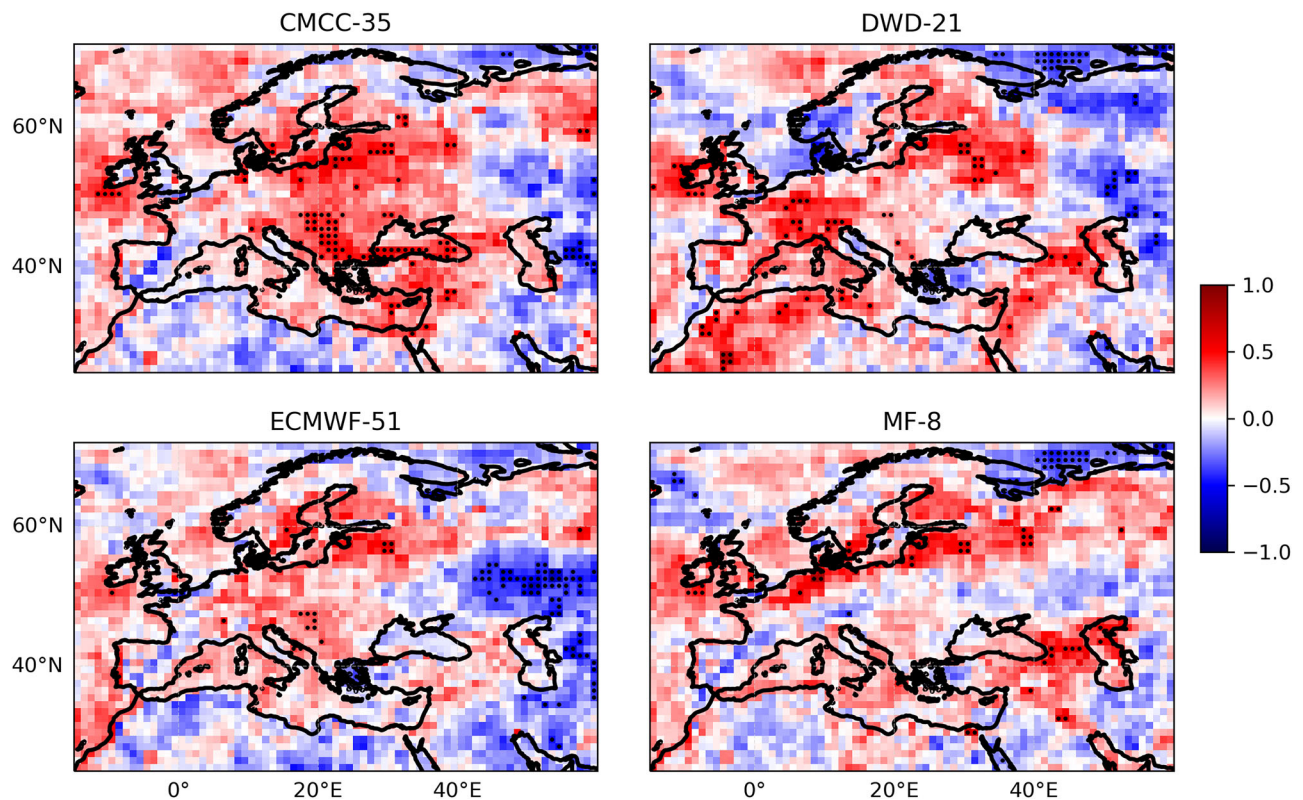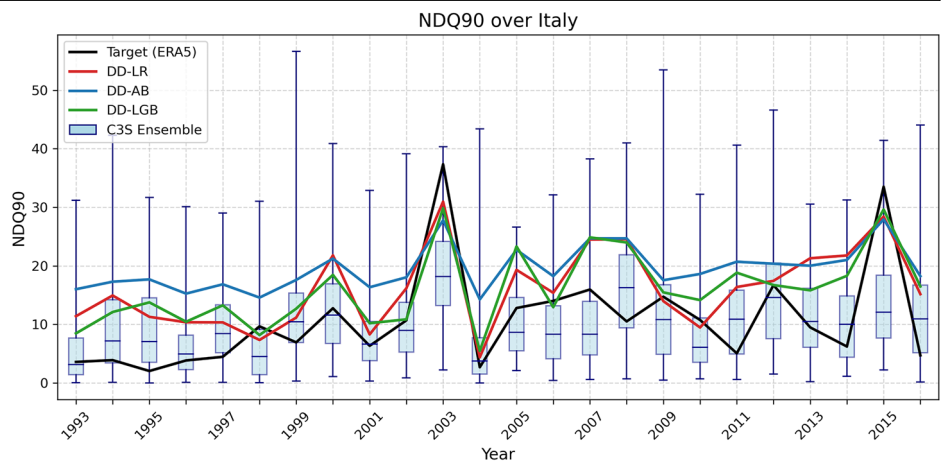
Fig. 5 | Differences in correlation skill score between the data-driven and individual C3S forecast systems. Positive values indicate higher skill in the data-driven system for forecasts of seasonal heatwave indicators (MJJ NDQ90). Black stippling represents a statistically significant correlation difference (c) at the 90% confidence level.

Fig. 6 | Forecasts of summer heatwave indicators over northern Italy. NDQ90 in ERA5 (black) is compared to the full ensemble (120 members) of the dynamical systems (C3S Ensemble) and the data-driven approach with three high-performing ML models (Linear Regression, DD-LR; AdaBoost, DD-AB; Light Gradient Boost, DD-LGB). Boxplots represent the medians, interquartile ranges and maxima and minima for each forecast year in the C3S Ensemble. The correlation skill scores over the 1993–2016 period are as follows: C3S ensemble median (0.63), DD-LR (0.74), DD-AB (0.78) and DD-LGB (0.76). The box used to define northern Italy is shown in Supplementary Fig. 10.

"outdated" predictors. For example, there has been a shift in the role of Arctic sea ice on European atmospheric circulation during recent decades[51,52]. However, it is clear that sufficient knowledge has been gained from the paleoclimate simulation to make accurate predictions. The feature selection identifies the principal role of predictors (e.g. soil moisture) at 4-8 weeks prior to initialisation, i.e. around March (Supplementary Fig. 6). This analysis provides indications for future studies on heatwave drivers, and can assist in describing the physical mechanisms behind their influence. Moreover, it serves as a means to study the recently highlighted differences in predictability between day and night extremes[19,50].

Here, the pool of potential predictors used is wider than typically used in feature selection studies on S2S or seasonal forecasting; the number of cluster variables is 70, with each time-lag (up to 28 weeks prior to the target)

counting as an individual predictor, thus leading to a total of roughly 2000 potential predictors. The framework does not rely only on expected or known predictors, but instead clusters predictor variables as a means of reducing the dimensionality of the problem without including human bias or relying on prior knowledge. A benefit of the optimisation algorithm used is its capacity to filter out unnecessary predictor information and identify the key predictors, thereby allowing the input of many potential predictors. Across the domain studied here, between 3.2 and 10.6% (6.5% on average) of the cluster-lag combinations are selected as predictors. Generally, including more predictors slows down ML forecasts, ruling out the possibility of including all 2000 in this case. To illustrate the benefits of the optimisation-based method, we compare it to an alternative and simpler method of feature selection based on linear correlation analysis, in which the selected

predictors are those significantly (positively or negatively) correlated to the target data during the period 1993–2016. Although the correlation approach selects a similar number of predictors (7.8% on average across the domain), the resulting forecasts are considerably poorer in skill (Supplementary Fig. 11). This highlights the greater ability of our optimisation-based approach to capture physically plausible predictors compared to simpler statistical-based approaches.

By focusing on one climate simulation (past2k) for training, it is demonstrated that there is a successful transfer of learning between the model and the real world. Increasing the training period from 50 to 1850 years of paleo-simulation data has a noticeable impact on the forecast skill (Supplementary Fig. 12), although there is limited growth in skill between 1000 and 1850 years of training data. This plateau implies that increasing the data beyond what is available from a single source, for example, by extending the paleo-simulation further back in time, would contribute little to improving the skill. Thus, the next avenue of research should be to attempt a multi-model training approach, for example, using the dynamical forecasting systems themselves as training data. Recent advances in short-term forecasting have also demonstrated that an ML-based ensemble outperforms deterministic data-driven forecast models, as in dynamical systems[53]. Emulating the dynamical multi-model approach with the data-driven system has the potential to further increase skill.

Although already successful, future improvements can be made to this prediction system. Parameter tuning for the ML models used could identify potential improvements, but it is an enormous undertaking for several models covering a wide geographical domain. K-means can be replaced by clustering algorithms which provide more interpretable and physically meaningful output[54]. We find that local variables are important for accurate predictions (Fig. 3), meaning that each target region should have a diverse range of potential predictors located close to it. The setup used in this study is ideal for forecasting central European conditions. The edge of the domain (e.g. Western Russia) displays lower skill (Fig. 4), likely due to the use of fewer local variables centred around this part of the domain. A crucial difference to the dynamical systems is that the current data-driven approach is inherently deterministic; future efforts must either explore probabilistic alternatives[53] or how to leverage the use of a single skilful member, such as for dynamical ensemble member selection[49].

Methods

ERA5 reanalysis

The ERA5 reanalysis[55,56] provides the target and predictor data for the modern period 1993–2016. Daily maximum 2m temperature (TMX) is used to calculate the heatwaves. To allow for comparison with the dynamical seasonal forecasts from the Copernicus Climate Change Service (C3S), TMX is regridded from the 0.25° regular grid to a regular 1° grid. The following variables are used as predictors: mean sea level pressure (SLP), volumetric soil moisture content in the upper 7 cm (SM), sea ice concentration (SIC), sea surface temperature (SST), geopotential height at 500 hPa (z500), outgoing longwave radiation (OLR), and TMX.

Previous studies have shown that ERA5 accurately reproduces both mean and extreme temperatures[57], confirming that it is a reliable source of climate information over Europe, in particular for heatwave indicators[58].

MPI-ESM paleo-simulation "past2k"

The "past2k" simulation is a simulation of the climate system with a state-of-the-art Earth System Model over the past two millennia. It was performed with the MPI-ESM1.2-LR model, which couples the ECHAM6.3 as its atmospheric component (1.875° horizontal resolution with 47 vertical levels) and the MPIOM1.63 as its ocean component (1.5° horizontal resolution reaching 30–40 km in the subpolar North Atlantic, with 40 vertical levels). The spin-up time of the simulation was 1200 model years prior to the year 0. The model is forced by reconstructions of past atmospheric greenhouse gases, volcanic forcing, solar forcings (with an artificial 11-year cycle), derived from the analysis of polar ice-core data; land-use changes derived from historical and palynological data; and ozone concentrations resulting

from an atmospheric photochemistry model forced by past solar irradiance[59,60]. No de-biasing or correction was made to the predictor or target data in past2k.

The soil moisture content in past2k is provided as the mass of water per m^2 in the upper 10 cm, as opposed to the volumetric water content in the upper 7 cm provided by ERA5. To convert from $m^3\,m^{-3}$ (ERA5) to $kg\,m^{-2}$ (past2k), we scale by density and extrapolate the value to 10 cm by assuming even distribution of water in the top 10 cm and a water density of $1000\,kg\,m^{-3}$.

The representation of summer temperature and heatwaves in past2k is found to agree with ERA5 (Supplementary Fig. 5). First, the patterns of interannual variability of NDQ90 are similar across Europe, although the magnitude is consistently higher in past2k. For example, peaks appear across the Mediterranean basin, the Caucasus and north-western Russia. In the two leading Empirical Orthogonal Functions (EOFs) of average summer TMX, past2k and ERA5 resemble each other in terms of patterns and magnitude; the leading EOF is dominated by variability over Russia, while the second clearly separates land and ocean variability.

Dynamical seasonal forecast systems

The C3S provides several operational dynamical seasonal forecast systems, each with a different number of ensemble members (i.e. individual realisations with perturbed initial conditions used to sample the uncertainty) and set-ups. Four systems from the C3S (Supplementary Table 2) were selected for this study: SPS3.5 from the Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC-35, 40 ensemble members), System2.1 from the Deutscher Wetterdienst (DWD-21, 30 ensemble members), SEAS5.1 from the European Centre for Medium-Range Weather Forecasts (ECMWF-51, 25 ensemble members), and System8 from Météo-France (MF-8, 25 ensemble members). These systems are initialised in "burst-mode", meaning the full set of ensemble members is run on the first day of the month. Other forecasting systems release ensemble members periodically throughout the month, in "lagged" mode. To remain consistent with the data-driven approach, which is also initialised on the first of each month, we use only burst-mode forecasts. The horizontal spatial resolution is 1°, and 6-hourly data are used to extract the daily maximum (TMX).

Definition of heatwave index

The target data used in this study is the number of days, from 1st May to 31st July, in which TMX exceeds the climatological 90th percentile (NDQ90). Each product uses its own respective climatology, thereby adjusting the mean biases inherent to the dynamical seasonal prediction systems. For ERA5, the 90th percentile is calculated by first averaging each calendar day over the 1993–2016 period and then applying an 11-day running mean to smooth the climatology[61]. For past2k, the period used is the last 30 years of the simulation (1821–1850), a choice which is justified by the stability of the climate throughout the simulation[60]. The NDQ90 index provides a measure of the propensity of a season to experience extreme temperatures, and displays similar variability and predictability to other indices based on intensity[8,18].

Optimisation-based feature selection and seasonal forecasts

Here, a data-driven seasonal forecast system is designed based on an optimisation-based feature selection framework[62] (Supplementary Fig. 1). The framework is composed of the following steps: a dimensionality reduction of potential predictor variables; a feature selection that identifies the optimal combination of variables, spatial domains and time-lags; the use of selected features to train statistical or ML prediction models. Previous work has successfully tested the ability of the framework to detect HWs based on short-term drivers in a detection-mode context[62]; here, the framework is adapted to only use predictor information on seasonal timescales, thereby providing seasonal forecasts.

First, a pool of potential predictors is defined. Here, a predictor refers to a variable within a domain at a particular time-lag. The chosen variables describe atmospheric, land and oceanic conditions, some of which are

known to influence the European climate or extremes, such as atmospheric circulation (e.g. z500[63–65]), ocean-atmospheric interactions (e.g. SST[66,67], soil moisture[68,69], and sea ice[51,52]). Among these, variables such as SST and OLR are used to capture modes of climate variability[30,70]. Using variables in a global domain or from outside Europe allows for the potential identification of teleconnections. An upgraded k-means clustering is applied to each variable to extract five clusters per domain (Supplementary Fig. 2; Supplementary Table 1). The innovation, compared to the traditional k-means, is performed using weighted multi-dimensional distances, where the Euclidean distances between the time series are given 95% weight and the remaining 5% is assigned to spatial distances on the geoid between the grid point and the centroid of the cluster. The clustering is performed on variable anomalies in ERA5 with respect to the 1993–2016 climatology. The ERA5-derived cluster shapes are then used to calculate weekly area-averages in both ERA5 and past2k, covering the period from November 1st to April 30th. Given that our objective is to demonstrate the capability of such a framework, we use $k = 5$ as a compromise between choosing a large pool of predictors and performing a suitable reduction of the dimensionality of the problem. The numbered week of the year is also included as a dummy predictor variable.

In the second step, the identification of predictors of extreme summers is treated as an optimisation problem. The optimisation algorithm employed is the Probabilistic Coral Reef Optimization algorithm with Substrate Layers (PCRO-SL)[43], which uses a multi-model method to combine different search procedures. In particular, it has recently been adapted to create a Spatio-Temporal Cluster-Optimized Feature Selection (STCO-FS) for heatwaves[62]. While previously the STCO-FS has been applied to the detection of HWs, here it is adapted to seasonal forecasting. We provide a description of the seasonal forecast setup of STCO-FS and refer readers to the aforementioned references for more complete technical descriptions of the optimisation algorithm PCRO-SL and the feature selection setup STCO-FS.

The aim of the second step is to select the combination of predictors which together provide the optimal skill for the target time series. Here, the problem to be optimised is the forecasting of NDQ90 using a multiple linear regression model. The optimisation is performed on past2k data, with training and test periods of 0–1600 and 1601–1850, respectively, and on each grid point individually. The skill score used is the root-mean-squared-error normalised by the standard deviation (interannual variability) of the target data (N-RMSE). The training score is calculated with a 5-fold cross-validation of the training period to reduce overfitting. Three parameters are simultaneously adjusted during the evolution of the optimisation: the variable cluster, the time-lag and the sequence length. The variable cluster is treated as a binary selection process (either selected or not). The time-lag, together with the sequence length, determines the times prior to May 1st in which the cluster is selected. Specifically, the sequence length represents the period during which the cluster is important. The value ranges for each parameter are as follows: variable cluster (0–1), time-lag (0–24 weeks prior to May 1st) and sequence length (0–8 weeks).

The optimisation aglorithm used begins with a first guess, after which the evolution of solutions during the optimisation improves both the training and test scores until the algorithm converges on an optimal solution, which typically occurs between 10,000–15,000 evolutions (Fig. 1). The solution with the lowest N-RMSE of all evolutions is selected as the optimal solution (see an example of selected variables and lags for an optimal solution in Supplementary Fig. 3). In summary, the optimal solution is obtained by repeating seasonal forecasts in the model world and adjusting the input predictors. The forecast is considered seasonal because the target data correspond to May–June–July, whereas the predictor data are obtained from the months prior to May.

The final step is to apply the method to real-world data. This requires simply changing the test period to 1993–2016 and the test predictors to those of ERA5, and extending the training period to cover the full 1850-year period of past2k. Given that past2k and ERA5 have different grids, a nearest-neighbour mapping function is used to associate past2k grid cells with those of ERA5.

While the optimisation is based on a multi-linear regression forecast for the sake of computational time, the second step of producing real-world forecasts can be performed with ML models in order to boost skill. Several candidates are tested (Supplementary 8), and in all cases, the default values provided in the Python modules (see Code Availability statement) are used. For instance, the Random Forest regressor used n_estimators=100, criterion="squared_error", max_depth=None, and max_features=1.0. Random Forests provide the greatest area of significant correlation over Europe, corresponding to a 10% increase over Linear Regression. However, the most skilful model depends on the grid point (Supplementary 8). In the most skilfully predicted regions (e.g. the central Mediterranean), all models provide significant skill, except Decision Trees. In the low skill zone extending over northern central Europe and Scandinavia, significant skill is rare among models, but the best performing models are Random Forest, Light Gradient Boost and AdaBoost. Multi-Layer Perceptron is a deep learning (DL) neural network model, ideal for larger datasets in which there are more non-linear relationships between predictors and the target. In this study, it is outperformed by ML-based models such as Random Forest, which has been found to be more suited to similar tasks[29,33]. The data-driven approach displayed in this study (e.g. Fig. 4) is derived from the most skilful model, depending on the grid point. While all models provide similar patterns of skill, there is no coherent pattern in which a model provides the best skill in certain regions (Supplementary Fig. 8).

The framework allows us to quantify the relative importance of each variable and cluster, and crucially, to identify the time-lag from short-term to seasonal timescales. By ensuring that potential predictors are restricted to certain time-lags, the system resembles a dynamical forecast system that receives climate information only before the initialisation date. The cut-off time for potential predictors determines the effective "initialisation" time; for example, using predictor data prior to May 1st to target summer HWs is equivalent to a May initialisation of the dynamical system.

SHapley Additive exPlanations - SHAP analysis
SHAP is a method used to interpret machine learning models by quantifying the contribution of each feature to individual predictions[47]. Here, we apply SHAP to Random Forest forecasts to explore the contribution of predictors selected in past2k to the forecasts using ERA5 predictors from 1993–2016. For each example predictor studied (Supplementary Fig. 7), and in each grid cell, we calculate the average of the SHAP value magnitudes for the target predictors.

Statistics
Statistical significance of correlations (e.g. Fig. 4a) is calculated using the two-sided test included in the stats.pearsonr function from the Python module scipy. Statistical significance of the difference between correlations (e.g. Fig. 4c) is calculated using a Fisher's Z-test, suitable for correlations with overlapping data (i.e. the ERA5 data used for validation). In both cases, a confidence interval of 95% is used, and the sample size is 24 (the number of available re-forecast years).

Data availability
The ERA5 data and dynamical seasonal forecast output used for this study are openly available at the Climate Data Store of the Copernicus Climate Change service: https://cds.climate.copernicus.eu/. Data from the paleoclimate simulation is available through the Earth System Grid Federation (ESGF), using identifiers "PMIP" (Paleoclimate Modeling Intercomparison Project) and "past2k".

Code availability
The machine learning models used in this study are from the scikit-learn (v1.4.1) Python programming language package and Microsoft LightGBM (v4.3.0). The default parameter values corresponding to the specific versions are used. The optimisation-based feature selection framework and seasonal forecasting framework are freely available at https://github.com/climateintelligence/DDHWSF.git. The repository contains the scripts used

to perform the optimisation and forecasting in this paper, and also guided step-by-step Jupyter notebooks used for training.

References

1. Lesk, C., Rowhani, P. & Ramankutty, N. Influence of extreme weather disasters on global crop production. *Nature* **529**, 84–87 (2016).
2. Zuo, J. et al. Impacts of heat waves and corresponding measures: a review. *J. Clean. Prod.* **92**, 1–12 (2015).
3. Hoegh Guldberg, O. et al. Impacts of 1.5°C global warming on natural and human systems (2018).
4. Mora, C. et al. Global risk of deadly heat. *Nat. Clim. Chang.* **7**, 501–506 (2017).
5. Russo, S., Sillmann, J. & Fischer, E. M. Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environ. Res. Lett.* **10**, 124003 (2015).
6. Ballester, J. et al. Heat-related mortality in Europe during the summer of 2022. *Nat. Med.* **29**, 1857–1866 (2023).
7. Feser, F., van Garderen, L. & Hansen, F. The summer heatwave 2022 over Western Europe: An attribution to anthropogenic climate change. *Bull. Am. Meteorol. Soc.* **105**, E2175–E2179 (2024).
8. Russo, S. et al. Magnitude of extreme heat waves in present climate and their projection in a warming world. *J. Geophys. Res. Atmos.* **119**, 12–500 (2014).
9. Kendrovski, V. et al. Quantifying projected heat mortality impacts under 21st-century warming conditions for selected European countries. *Int. J. Environ. Res. Public Health* **14**, 729 (2017).
10. Lowe, R. et al. Evaluation of an early-warning system for heat wave-related mortality in Europe: Implications for sub-seasonal to seasonal forecasting and climate services. *Int. J. Environ. Res. Public Health* **13**, 206 (2016).
11. Klemm, T. & McPherson, R. A. The development of seasonal climate forecasting for agricultural producers. *Agric. For. Meteorol.* **232**, 384–399 (2017).
12. Darbyshire, R. et al. Insights into the value of seasonal climate forecasts to agriculture. *Aust. J. Agric. Resour. Econ.* **64**, 1034–1058 (2020).
13. Terzago, S., Bongiovanni, G. & Von Hardenberg, J. Seasonal forecasting of snow resources at alpine sites. *Hydrol. Earth Syst. Sci.* **27**, 519–542 (2023).
14. Luo, L. & Zhang, Y. Did we see the 2011 summer heat wave coming? *Geophys. Res. Lett.* **39**, L09708 (2012).
15. Ardilouze, C., Batté, L. & Déqué, M. Subseasonal-to-seasonal (s2s) forecasts with CNRM-CM: a case study on the July 2015 West-European heat wave. *Adv. Sci. Res.* **14**, 115–121 (2017).
16. Batté, L., Ardilouze, C. & Déqué, M. Forecasting West African heat waves at subseasonal and seasonal time scales. *Mon. Weather Rev.* **146**, 889–907 (2018).
17. Domeisen, D. I. et al. Prediction and projection of heatwaves. *Nat. Rev. Earth Environ.* **4**, 36–50 (2023).
18. Prodhomme, C. et al. Seasonal prediction of European summer heatwaves. *Clim. Dyn.* **58**, 2149–2166 (2022).
19. Torralba, V. et al. Nighttime heat waves in the Euro-Mediterranean region: definition, characterisation, and seasonal prediction. *Environ. Res. Lett.* **19**, 034001 (2024).
20. Katsafados, P., Papadopoulos, A., Varlas, G., Papadopoulou, E. & Mavromatidis, E. Seasonal predictability of the 2010 Eussian heat wave. *Nat. Hazards Earth Syst. Sci.* **14**, 1531–1542 (2014).
21. Lam, R. et al. Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
22. Keisler, R. Forecasting global weather with graph neural networks. Preprint at https://arxiv.org/abs/2202.07575 (2022).
23. Chen, L. et al. FuXi_ A cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.* **6**, 190 (2023).
24. Kurth, T. et al. Fourcastnet: accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proc. Platform for Advanced Scientific Computing Conference* 1–11 (Association for Computing Machinery, 2023).
25. Bi, K. et al. Accurate medium-range global weather forecasting with 3d neural networks. *Nature* **619**, 533–538 (2023).
26. Lang, S. et al. AIFS-ECMWF's data-driven forecasting system. Preprint at https://arxiv.org/abs/2406.01465 (2024).
27. Chen, L. et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nat. Commun.* **15**, 6425 (2024).
28. Watt-Meyer, O. et al. ACE2: accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj clim. atmos. sci.* **8**, 205 (2025).
29. Materia, S. et al. Artificial intelligence for climate prediction of extremes: state of the art, challenges, and future perspectives. *Wiley Interdiscip. Rev. Clim. Chang.* **15**, e914 (2024).
30. Della-Marta, P. M. et al. Summer heat waves over Western Europe 1880–2003, their relationship to large-scale forcings and predictability. *Clim. Dyn.* **29**, 251–275 (2007).
31. Kämäräinen, M. et al. Statistical learning methods as a basis for skillful seasonal temperature forecasts in Europe. *J. Clim.* **32**, 5363–5379 (2019).
32. Ford, T. W., Dirmeyer, P. A. & Benson, D. O. Evaluation of heat wave forecasts seamlessly across subseasonal timescales. *NPJ Clim. Atmos. Sci.* **1**, 20 (2018).
33. Van Straaten, C., Whan, K., Coumou, D., Van den Hurk, B. & Schmeits, M. Using explainable machine learning forecasts to discover subseasonal drivers of high summer temperatures in western and central Europe. *Mon. Weather Rev.* **150**, 1115–1134 (2022).
34. Pyrina, M. & Domeisen, D. I. Subseasonal predictability of onset, duration, and intensity of European heat extremes. *Q. J. R. Meteorol. Soc.* **149**, 84–101 (2023).
35. Weirich-Benet, E. et al. Subseasonal prediction of central European summer heatwaves with linear and random forest machine learning models. *Artif. Intell. Earth Syst.* **2**, e220038 (2023).
36. Rouges, E., Ferranti, L., Kantz, H. & Pappenberger, F. Pattern-based forecasting enhances the prediction skill of European heatwaves into the sub-seasonal range. *Clim. Dyn.* **62**, 9269–9285 (2024).
37. Xu, L., Chen, N., Zhang, X. & Chen, Z. A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Clim. Dyn.* **54**, 3355–3374 (2020).
38. Sattari, M. T., Feizi, H., Samadianfard, S., Falsafian, K. & Salwana, E. Estimation of monthly and seasonal precipitation: a comparative study using data-driven methods versus hybrid approach. *Measurement* **173**, 108512 (2021).
39. Zhang, R. Z., Jia, X. J. & Qian, Q. F. Seasonal forecasts of Eurasian summer heat wave frequency. *Environ. Res. Commun.* **4**, 025007 (2022).
40. Zhu, Y. et al. Deep learning-based seasonal forecast of sea ice considering atmospheric conditions. *J. Geophys. Res.: Atmos.* **128**, e2023JD039521 (2023).
41. Lee, Y. et al. Unveiling teleconnection drivers for heatwave prediction in South Korea using explainable artificial intelligence. *npj Clim. Atmos. Sci.* **7**, 176 (2024).
42. Wu, G., Mallipeddi, R. & Suganthan, P. N. Ensemble strategies for population-based optimization algorithms–a survey. *Swarm Evolut. Comput.* **44**, 695–711 (2019).
43. Pérez-Aracil, J. et al. New probabilistic, dynamic multi-method ensembles for optimization based on the cro-sl. *Mathematics* **11**, 1666 (2023).
44. Bhend, J., Mahlstein, I. & Liniger, M. A. Predictive skill of climate indices compared to mean quantities in seasonal forecasts. *Q. J. R. Meteorol. Soc.* **143**, 184–194 (2017).

45. Johnson, S. J. et al. Seas5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.* **12**, 1087–1117 (2019).
46. Luo, M. & Lau, N.-C. Summer heat extremes in northern continents linked to developing ENSO events. *Environ. Res. Lett.* **15**, 074042 (2020).
47. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (eds Guyon, I. et al.) 4765–4774 (Curran Associates, Inc., 2017).
48. Fröhlich, K. et al. The German Climate Forecast System: GCFS. *J. Adv. Model. Earth Syst.* **13**, e2020MS002101 (2021).
49. Famooss Paolini, L., Ruggieri, P., Pascale, S., Brattich, E. & Di Sabatino, S. Hybrid statistical–dynamical seasonal prediction of summer extreme temperatures in Europe. *Q. J. R. Meteorol. Soc.* **151**, p.e4900 (2024).
50. Wu, S. et al. Local mechanisms for global daytime, nighttime, and compound heatwaves. *npj Clim. Atmos. Sci.* **6**, 36 (2023).
51. Coumou, D., Di Capua, G., Vavrus, S., Wang, L. & Wang, S. The influence of arctic amplification on mid-latitude summer circulation. *Nat. Commun.* **9**, 2959 (2018).
52. Zhang, R., Sun, C., Zhu, J., Zhang, R. & Li, W. Increased european heat waves in recent decades in response to shrinking arctic sea ice and eurasian snow cover. *NPJ Clim. Atmos. Sci.* **3**, 7 (2020).
53. Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* 1–7 (2024).
54. Bonetti, P., Metelli, A. M. & Restelli, M. Interpretable linear dimensionality reduction based on bias-variance analysis. *Data Mining and Knowledge Discovery* 1–69 (2024).
55. Hersbach, H. et al. The era5 global reanalysis. *Q. J. R. Meteorological Soc.* **146**, 1999–2049 (2020).
56. Soci, C. et al. The ERA5 global reanalysis from 1940 to 2022. *Quarterly Journal of the Royal Meteorological Society* (2024).
57. Velikou, K., Lazoglou, G., Tolika, K. & Anagnostopoulou, C. Reliability of the ERA5 in replicating mean and extreme temperatures across Europe. *Water* **14**, 543 (2022).
58. Skrynyk, O., Aguilar, E. & Cimolai, C. The sensitivity of heatwave climatology to input gridded datasets: a case study of Ukraine. *Atmosphere* **16**, 289 (2025).
59. Jungclaus, J. H., Lohmann, K. & Zanchettin, D. Enhanced 20th-century heat transfer to the Arctic simulated in the context of climate variations over the last millennium. *Climate* **10**, 2201–2213 (2014).
60. Jungclaus, J. H. et al. The PMIP4 contribution to CMIP6–part 3: the last millennium, scientific objective, and experimental design for the PMIP4 past1000 simulations. *Geosci. Model Dev.* **10**, 4005–4033 (2017).
61. Perkins, S. E. & Alexander, L. V. On the measurement of heat waves. *J. Clim.* **26**, 4500–4517 (2013).
62. Perez-Aracil, J. Identifying key drivers of heatwaves: a novel spatio-temporal framework for extreme event detection. *Weather Clim. Extremes* **49**, 100792 (2025).
63. Sousa, P. M., Trigo, R. M., Barriopedro, D., Soares, P. M. & Santos, J. A. European temperature responses to blocking and ridge regional patterns. *Clim. Dyn.* **50**, 457–477 (2018).
64. Kornhuber, K. et al. Amplified Rossby waves enhance risk of concurrent heatwaves in major breadbasket regions. *Nat. Clim. Change* **10**, 48–53 (2020).
65. Kautz, L.-A. et al. Atmospheric blocking and weather extremes over the Euro-Atlantic sector–a review. *Weather Clim. Dyn. Discuss.* **2021**, 1–43 (2021).
66. Duchez, A. et al. Drivers of exceptionally cold North Atlantic Ocean temperatures and their link to the 2015 european heat wave. *Environ. Res. Lett.* **11**, 074004 (2016).
67. Lipfert, L., Hand, R. & Brönnimann, S. A global assessment of heatwaves since 1850 in different observational and model data sets. *Geophys. Res. Lett.* **51**, e2023GL106212 (2024).
68. Stefanon, M., D'Andrea, F. & Drobinski, P. Heatwave classification over Europe and the mediterranean region. *Environ. Res. Lett.* **7**, 014023 (2012).
69. Materia, S. et al. Summer temperature response to extreme soil water conditions in the mediterranean transitional climate regime. *Clim. Dyn.* **58**, 1943–1963 (2022).
70. Kenyon, J. & Hegerl, G. C. Influence of modes of climate variability on global temperature extremes. *J. Clim.* **21**, 3872–3889 (2008).

Author contributions
Ronan McAdam conceived the study, performed the analysis and drafted the manuscript. Jorge Pérez-Aracil contributed to analysis, visualisation and software provision. Antonello Squintu contributed to data curation, designing the study, and software provision. Cesar Peláez-Rodríguez contributed to analysis, visualisation and software provision. Felicitas Hansen contributed to data curation and validation of datasets. Verónica Torralba contributed to data curation, software provision and validation of datasets. Harilaos Loukos contributed to designing the study and was responsible for funding acquisition. Eduardo Zorita contributed to designing the study and validation of datasets and was responsible for funding acquisition. Matteo Giuliani contributed to designing the study and was responsible for funding acquisition. Leone Cavicchia contributed to designing the study. Sancho Salcedo-Sanz contributed to designing the study and was responsible for funding acquisition. Enrico Scoccimarro conceived the study and was responsible for funding acquisition. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43247-025-02863-4.

Correspondence and requests for materials should be addressed to Ronan McAdam.

Peer review information *Communications Earth and Environment* thanks Cheolhee Yoo, Jonas Bhend, Jing-Jia Luo for their contribution to the peer review of this work. Primary Handling Editors: Seung-Ki Min and Martina Grecequet, Aliénor Lavergne. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.