Using Explainable Machine Learning Forecasts to Discover Subseasonal Drivers of High Summer Temperatures in Western and Central Europe

CHIEM VAN STRAATEN,[a,b] KIRIEN WHAN,[a] DIM COUMOU,[b,c,a] BART VAN DEN HURK,[d] AND MAURICE SCHMEITS[a,b]

[a] *Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands*
[b] *Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, Amsterdam, Netherlands*
[c] *Potsdam Institute for Climate Impact Research, Postdam, Germany*
[d] *Deltares, Delft, Netherlands*

ABSTRACT: Reliable subseasonal forecasts of high summer temperatures would be very valuable for society. Although state-of-the-art numerical weather prediction (NWP) models have become much better in representing the relevant sources of predictability like land and sea surface states, the subseasonal potential is not fully realized. Complexities arise because drivers depend on the state of other drivers and on interactions over multiple time scales. This study applies statistical modeling to ERA5 data, and explores how nine potential drivers, interacting on eight time scales, contribute to the subseasonal predictability of high summer temperatures in western and central Europe. Features and target temperatures are extracted with two variations of hierarchical clustering, and are fitted with a machine learning (ML) model based on random forests. Explainable AI methods show that the ML model agrees with physical understanding. Verification of the forecasts reveals that a large part of predictability comes from climate change, but that reliable and valuable subseasonal forecasts are possible in certain windows, like forecasting monthly warm anomalies with a lead time of 15 days. Contributions of each driver confirm that there is a transfer of predictability from the land and sea surface state to the atmosphere. The involved time scales depend on lead time and the forecast target. The explainable AI methods also reveal surprising driving features in sea surface temperature and 850 hPa temperature, and rank the contribution of snow cover above that of sea ice. Overall, this study demonstrates that complex statistical models, when made explainable, can complement research with NWP models, by diagnosing drivers that need further understanding and a correct numerical representation, for better future forecasts.

KEYWORDS: Statistical forecasting; Subseasonal variability; Machine learning; Model interpretation and visualization

1. Introduction

In the recent two decades Europe faced a multitude of impactful heat extremes (e.g., Barriopedro et al. 2011; Russo et al. 2014), that exceeded the modeled expectation (van Oldenborgh et al. 2013). These were disasters that ended in loss of life and a severe disruption of activities. If forecasts would exist that reliably predict such events more than 2 weeks in advance, then new forms of anticipatory risk management can be realized (White et al. 2017). Hints of such predictability have already been found at lead times ranging from subseasonal (2–6 weeks) (Wulff and Domeisen 2019) to seasonal (Weisheimer et al. 2011; Prodhomme et al. 2016, 2022), but the current time window to act upon operational forecasting systems remains limited to shorter lead times (Coughlan de Perez et al. 2018; Casanueva et al. 2019).

The complication with subseasonal lead times is that climate variables have time-varying contributions to predictands. Involvement of a variable is conditional on the state of other variables (Mariotti et al. 2020). Only occasionally do windows of predictability appear in the larger background of chaotic variation. These are conditions in which subseasonal forecasts have a greater opportunity to succeed (Albers and Newman 2019; Mariotti et al. 2020; Mayer and Barnes 2021). Sahelian heatwaves are for instance more successfully forecast during active modes of tropical variability (Guigma et al. 2021). But windows of predictability are hardly regular. Contributions of key drivers of heat extremes do vary from event to event (Wehrli et al. 2019). Also, interactions that lead up to an event, take place over a range of time scales and locations (Sillmann et al. 2017). Such complexities have hampered our current understanding of the set of potential heatwave drivers, and of interactions between them (Perkins 2015). To therefore discover and leverage new opportunities for European subseasonal forecasts, we need to learn what methods of sufficient complexity learn. When we supply the climate variables that we know are important, machine learning (ML) tools can help us to extract better driving features from them (Cohen et al. 2018).

Clearly involved is atmospheric variability, more specifically the synoptic high pressure "blocking" systems that are associated to high surface temperatures (Brunner et al. 2017; Schaller et al. 2018). In the midlatitudes these systems are

---

ⓞ Denotes content that is immediately available upon publication as open access.

Corresponding author: Chiem van Straaten, chiem.van.straaten@knmi.nl

part of quasi-stationary Rossby wave packets (RWPs) (Schubert et al. 2011; Röthlisberger et al. 2019). In transient state such groups of waves travel eastward along the waveguide of the jet stream, instigating low and high pressure systems downstream (Wirth et al. 2018). But in quasi-stationary state the pattern persists over a geographical location, leading to potentially large temperature anomalies (Wolf et al. 2018). Air parcels enter the stagnant high pressure system, descend and adiabatically heat the lower atmosphere (Zschenderlein et al. 2019).

Variability in the land surface is also involved, and conditionally enhances or negates high temperatures. When the land surface is dry, the excess energy from atmospheric temperature and clear-sky conditions will primarily go into diabatic sensible heating, but when the land surface is wet the excess will be used for evapotranspiration too (Seneviratne et al. 2010; Miralles et al. 2019). The potential feedback from a dry land surface in summer can follow after winter and spring precipitation deficits (Quesada et al. 2012). Antecedent drought can also be driven by temporally enhanced transpiration when vegetation greens anomalously in spring (Fischer et al. 2007; Ma et al. 2016). Apart from inter-seasonal interactions, the land surface is also involved in shorter-term interactions that can amplify developing events (Schumacher et al. 2019). Fischer et al. (2007), Haarsma et al. (2009) and Zampieri et al. (2009) find that drier soils can increase upper-level high pressure, and could partially be responsible for stagnating the RWP pattern.

The third type of variability involved at multiple time scales comes from the global ocean. Two-way interaction with anomalous sea surface temperatures (SSTs) can drive meridional jet stream position over the North Atlantic in June–August (JJA) (Duchez et al. 2016; Ossó et al. 2020). More specifically within the season, SST patterns can predictably precede hot events by sourcing involved RWPs (McKinnon et al. 2016), which might be the way the tropical Atlantic influenced the 2003 heatwave (Cassou et al. 2005). SST patterns can also reinforce and thereby stagnate existing RWPs (Black and Sutton 2007; Della-Marta et al. 2007; Feudale and Shukla 2011). At Arctic latitudes the ocean state is equally important, with sea ice cover influencing summertime jet stream position and quasi-stationary RWP amplitude in the Euro-Atlantic region (Hall et al. 2017; Wolf et al. 2020). The same effect, originating from Arctic landmass, is found for snow cover.

Clearly, the features leading up to hot events are not limited to a single time scale or region. Part of this is due to interaction between climate variables, but the involvement of multiple time scales is also just a fact of atmospheric dynamics itself (Schneidereit et al. 2012). Numerical model experiments can be used to disentangle different contributions (e.g., Koster et al. 2010; Stéfanon et al. 2012a; van den Hurk et al. 2012; Wehrli et al. 2019; Osborne et al. 2020). Through one-by-one manipulation, a variable's role in feedbacks or as source of predictability can be diagnosed. In observations such causal "what-if" manipulation is not possible (Runge et al. 2019). Empirical analysis in observations is more likely to measure "association." The empirical approach, however, will include mechanisms that are imperfectly represented in numerical

models. Successful statistical diagnosis has happened through composite driver statistics conditioned on high temperature events (e.g., Stéfanon et al. 2012b; Brunner et al. 2017), composite temperature statistics conditioned on Euro-Atlantic circulation (e.g., Cassou et al. 2005; Jézéquel et al. 2018), plain predictive association like regression in a small set of variables (e.g., Quesada et al. 2012; Hall et al. 2017; Suarez-Gutierrez et al. 2020; Kueh and Lin 2020), analysis of dominant modes in the full multivariate space (e.g., Della-Marta et al. 2007; O'Reilly et al. 2018), or using a potential driver as covariate in a modeled distribution of temperature (e.g., Whan et al. 2015).

What all empirical approaches have in common is a limited scope in terms of variables and time scales. Subsets of variables are selected a priori and interaction between them is often ruled out. This leads to a setup that is easy for humans to understand, but that could falsely attribute underlying, undiscovered features to ones that are only partially involved. Such partial information, not representing the source of predictability itself, might vary from forecast occasion to forecast occasion. When operationalized, we would not know whether the forecasts can be trusted, and whether its learned patterns are actionable, should they arise in real time.

This study presents a data-driven method to extract information from nine important climate variables related to high European temperatures in JJA, limiting a priori choices. Subsequently, an ML method based on random forests reconstructs the interaction between variables on a range of time scales, and forecasts the exceedance of regional temperature above a given threshold. The forecasts are verified in terms of skill and potential value to users. Foremost we are interested in the climate variable features that the method extracts and leverages as sources of predictability. Usually such complex "black box" ML methods are hard to understand, making them untrustworthy in their own way. Here we demonstrate how explainability tools, which have become more and more mature in recent times (McGovern et al. 2019; Molnar et al. 2020), can be used to query the ML method for prior knowledge. For instance for the theory that relevant information is carried across variables and time scales, from oceanic RWP sources and antecedent land surface conditions long before the event, to the atmospheric state close to the event. Last, we show how the method could behave in real time. To that end, we study its predictions of the European heatwave in 2015 (Duchez et al. 2016; Ardilouze et al. 2017). Section 2 introduces the data processing steps and ML methodology. Section 3 presents the verification results, found sources of predictability and 2015 case study. Section 4 provides a discussion and the conclusions.

2. Data and methods

a. Reanalysis data

Nine climate variables and the target 2-m air temperature (t2m) are obtained from the ERA5 dataset. This modeled reality, based on assimilated observations (Hersbach et al. 2020), has the benefit of being multivariate, spatially dense and temporally homogeneous. Atmospheric variability is
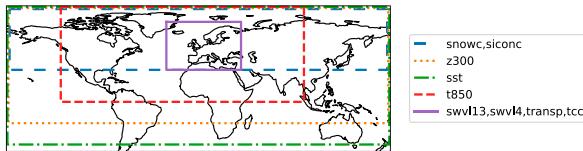
FIG. 1. Spatial domains used to extract climate variables from ERA5. See Table 1 for the meaning of variable abbreviations.

represented by 300-hPa geopotential (z300), 850-hPa temperature (t850), and total cloud-cover fraction (tcc). Respectively, these capture upper-level wave patterns, lower-level heating, and the clear-sky conditions of high pressure systems. Oceanic states are represented by sea ice concentration (siconc) and sea surface temperature (sst), also from ERA5. Land surface variability is represented by snow-cover fraction (snowc), transpiration (transp) and deep and shallow volumetric soil moisture (swvl4 and swvl13), all from ERA5-Land (Muñoz-Sabater et al. 2021).

Since ERA5-Land starts in 1981, 2 years later than ERA5, we extract the data from 1981 to 2019, at high spatial resolution ($0.25° \times 0.25°$ and $0.1° \times 0.1°$ for ERA5 and ERA5-Land, respectively). Domains are variable-specific (Fig. 1). When teleconnections, like influence from the tropical Atlantic, were expected, the domain was kept large enough for the ML-method to use potential features from these regions. For variables without expected global teleconnections we used smaller domains, to ease the total computational load. Hourly values at all grid cells were resampled to a daily resolution (Table 1). For each variable we removed the seasonal cycle by subtracting for each calendar date the average value of similar calendar dates (± 5 days), found in the period 1981–2019.

At this point in the data processing we grouped the resulting gridded fields of daily anomalies into 5 distinct datasets or "folds." The 39 summer seasons were repeatedly split into a 31- or 32-season subset for training and a 7- or 8-season subset for verification. The verification sets consist of consecutive and complete summers. Temperature evolution in one part of summer can depend on the evolution in other parts, and information leakage between training and verification sets needs to be avoided.

b. Target variables

After resampling from ERA5, the gridded t2m anomalies are at daily temporal resolution. A target at such resolution would appear unpredictable at subseasonal lead times. Subseasonal signals can only be extracted from the total variability by aggregating multiple days or even weeks (Hoskins 2013). However, the optimal level of temporal aggregation as well as the level of spatial aggregation is hard to establish a priori (van Straaten et al. 2020). For summer temperatures, spatial domains smaller than the continent are preferred (Jézéquel et al. 2018), because hot spell types and their relation to subseasonal drivers vary across Europe (Sousa et al. 2018; Stéfanon et al. 2012b). Besides, subcontinental domains are also preferable from a forecast user perspective.

We obtain our sub-European target region by means of agglomerative hierarchical clustering. This algorithm groups grid cells that are similar, starting strict, with single-cell groups only, but gradually allowing more and more dissimilarity, merging those groups that comply. First we let binary time series indicate whether local t2m anomalies exceed the grid cell's 95th climatological percentile. A high synchronicity between two such series, i.e., two grid cells sharing many daily exceedances, indicates that they are governed by the same regional hot spells. We therefore measure the number of non-shared daily exceedances relative to the number of shared ones, with the Jaccard dissimilarity (as in McKinnon et al. 2016). We compute it between all gridcell pairs, and cluster them hierarchically. At the level where on average, in each cluster, 10% of the exceedances are shared, a suitable central-west European cluster emerged. This region is selected as target for this study (Fig. 2b). A similar geographic region appeared when climatological percentiles like the 66th were used to define dissimilarity.

From the spatially averaged t2m anomalies over this region, binary prediction targets can be defined. These targets have a

TABLE 1. Nine climate variables related to high summer 2-m air temperatures in Europe. Resampled to daily values from the hourly ERA5 dataset.

| Variable	Abbreviation	ERA-5	Daily resampling	Unit
2-m temperature	t2m	Single level	24-h mean	K
300-hPa geopotential	z300	Pressure level	1200 UTC	$m^2 s^{-2}$
850-hPa temperature	t850	Pressure level	1200 UTC	K
Total cloud cover	tcc	Single level	24-h mean	—
Sea surface temperature	sst	Single level	24-h mean	K
Sea ice concentration	siconc	Single level	24-h mean	—
Snow cover	snowc	Land	2400 UTC	—
Transpiration	transp	Land	24-h accumulation	m
Shallow volumetric soil water	swvl13	Land	24-h mean, depth-weighted average of upper three layers (0–100 cm)	$m^3 s^{-3}$
Deep volumetric soil water	swvl4	Land	24-h mean, bottom layer (100–289 cm)	$m^3 s^{-3}$
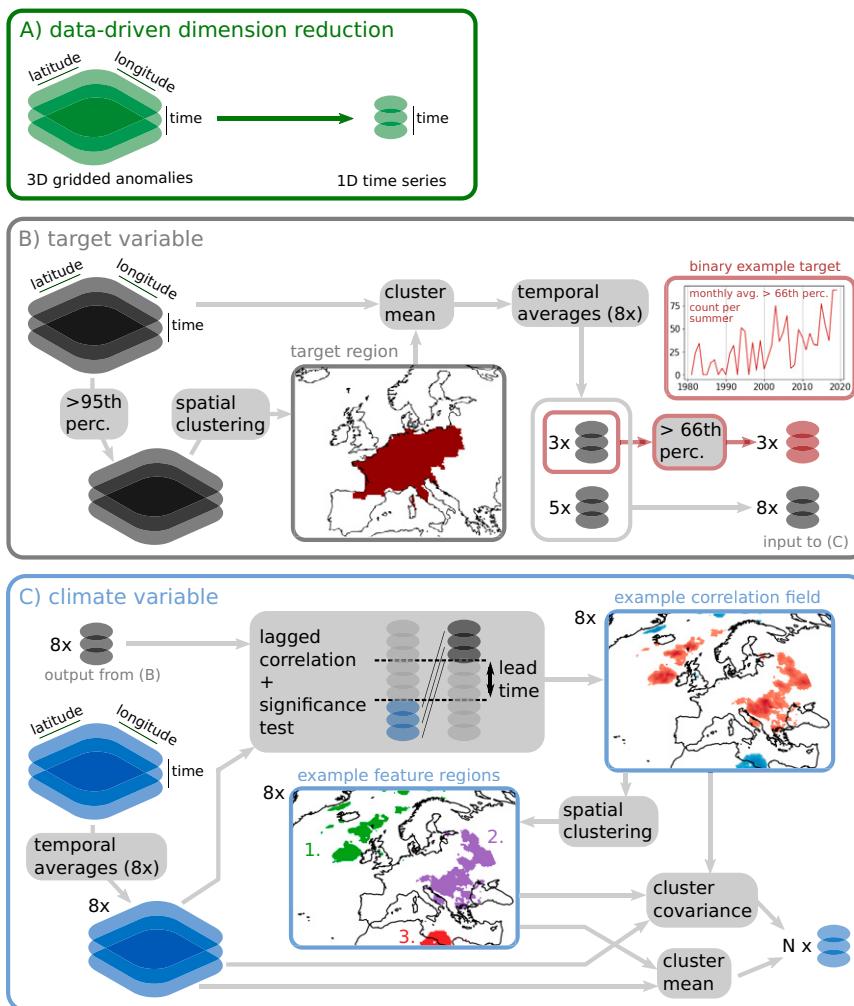
FIG. 2. Method for data-driven feature extraction from gridded climate variables. (a) Example of reducing the dimensionality of 3D gridded anomalies to 1D feature time series. (b) Extraction of three target time series: average regional temperature, also averaged in time, exceeding a given threshold, resulting in the red binary series (see also section 2b). Eight intermediate continuous time series, the result of averaging with windows of varying length, are used in the next step. (c) The extraction of potentially predictive features from a climate variable (blue; total cloud cover in this example), through correlation with temperature across a lead time gap, at eight different time scales. Multiple discovered feature regions, with two-dimension reduction steps each (covariance and mean) lead to N time series (see also section 2c).

value of 1 when the weekly (7-day), biweekly (15-day) or monthly (31-day) average temperature exceeds the 50th, 66th, or 90th climatological percentile (Fig. 2b), otherwise they are 0. The thresholds are computed for each temporal averaging window separately. In addition, we want the ML method to detect at which time scales the climate variables relate to the weekly, biweekly and monthly temperature target. The eight possible relation time scales are set to 1, 3, 5, 7, 11, 15, 21 and 31 days. Each time scale is defined as a rolling time window applied to the spatially averaged t2m and climate variables simultaneously, when climate variable features belonging to the time scale get extracted (Fig. 2c) (section 2c).

c. Data-driven features

The climate variable features related to high temperatures will likely depend on the lead time at which they are sought. We allow for this possibility without assuming it a priori, by always supplying all nine variables on all eight time scales to the feature extraction in each lead time. Eight time scales are available because the rolling windows provide an average value each day, regardless of the applied window size. Crucially, this allows the complex ML method (described later and also fitted per lead time) to let features from multiple time scales interact. Chosen lead times mirror the eight

aggregation time scales (1, 3, 5, 7, 11, 15, 21, and 31 days) and are defined as the number of days between the end of input information and the start of (predicted) output information. The rolling windows are thus applied in backward mode to the climate variables (timestamp at the end of each period) and in forward mode to the predicted 2-m temperatures (timestamp at the start of each period). Also short lead times (up to 7 days) are examined to confirm whether our method shows the expected high skill under fairly deterministic conditions.

The extraction applied per lead time is a dimensionality reduction of the gridded variable data (Fig. 2a), that differs from conventional tools for dimension reduction. Principal component analysis for instance, reduces variability to a few dominant components (e.g., Kämäräinen et al. 2019), but without guarantee that the variables projecting highly into those components contain information relevant to a particular forecasting problem (Lakshmanan et al. 2015). Theoretically, the same issue occurs when upper-level (e.g., 300-hPa) geopotential is reduced to a set of regimes first, from which temperature is predicted at second instance. Although such methods have established conceptual modes of variability, like the summer North Atlantic Oscillation (Folland et al. 2009; Bladé et al. 2012), that proved insightful for heatwave research (e.g., Cassou et al. 2005; Kueh and Lin 2020), this study deliberately uses a dimension-reduction that ensures direct relevance to the target variable (e.g., Kretschmer et al. 2017). This choice penalizes simplicity and interpretability in terms of well-established modes, but serves the goal of discovering (potentially new) sources of predictability in nine climate variables.

First, we compute for each lead time and each climate variable grid cell c a partial rank correlation $r_{\mathrm{partial},c}$ between the time series of the climate variable $x_{t,c}$ and the spatial mean temperature $\overline{y}_t$ in the target region, where the time dimension for each entity is aggregated to one of the eight time scales (Fig. 2c, largest gray box). For example, at a lead time of 15 days we correlate 31-day cloud cover with 31-day temperature, and we correlate 1-day cloud cover with 1-day temperature. The rank correlation is called partial because we remove common drivers that can inflate the correlation (Runge et al. 2014). Seasonality was already removed in the creation of anomalies, so at each time t we use linear regression to further remove the long-term climatic trend and the influence of autocorrelation from values at $t - \tau$:

$$\hat{\overline{y}}_t = \beta_1 \overline{y}_{t-\tau} + \beta_2 t, \qquad (1)$$

$$\hat{x}_{t,c} = \beta_{3,c} x_{t-\tau,c} + \beta_{4,c} t, \qquad (2)$$

$$r_{\mathrm{partial},c} = r\left(\overline{y}_t - \hat{\overline{y}}_t, x_{t,c} - \hat{x}_{t,c}\right), \qquad (3)$$

where τ is the rolling window size and where β_1, β_2, β_3, and β_4 are regression coefficients. $\beta_{3,c}$ and $\beta_{4,c}$ are estimated for each grid cell separately, in order to apply the detrending procedure for every grid cell separately.

Each correlation value is tested for significant two-sided difference from zero, with a confidence level α that becomes increasingly strict with window size: $\alpha = 5 \times 10^{-4 - 0.3(\tau-1)}$.

This relation was found experimentally, and provides good correction for the increase in dependence by rolling window aggregation. The joint result of all correlation values is a correlation field (example in Fig. 2c) to which we apply an extra false discovery rate correction (Benjamini and Hochberg 1995). This corrects for the inclusion of grid cells that are significant by chance and results in a single field per lead time, climate variable and time scale. And a single field per cross-validation fold, because we repeat the computation for each training set, as it is known that correlation patterns depend on the years in which they are computed (Garcia-Serrano and Frankignoul 2014).

The number of features extracted from each field is data-driven and therefore variable. The cloud cover example (Fig. 2c) illustrates that multiple contiguous groups of significantly correlated cells can be present in a single field. The patterns of negatively and positively related anomalies need not arise at the same time. More likely, especially at larger distances between cell-groups, is that each group arises as an independent regional feature. We use the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) algorithm (McInnes et al. 2017) to identify these features. HDBSCAN is preferred over standard clustering methods like k-means to better handle the large grid cell density differences pertaining to the regular latitude-longitude grid (Saunders et al. 2021). Variants of HDBSCAN have also been applied by e.g., Zhang et al. (2019) and Tilloy et al. (2021). Here we use haversine to measure the geographical distance between all grid cell pairs.

In the example HDBSCAN leads to three distinct clusters (Fig. 2c). We call these clusters "feature regions" to distinguish them from the "target region" produced by the hierarchical clustering procedure for the target. A close look at region 1 reveals how it can comprise both negatively and positively related anomalies (blue and red in the correlation field, respectively). Though not necessarily the case in this cloud cover example, such dipoles are common driving features, for instance in Atlantic SST (Ossó et al. 2020), but also in the sequence of low and high pressure systems that constitute an RWP (Schubert et al. 2011; Wolf et al. 2018). In these cases we allow a larger geographical distance to exist within the cluster, such that all negative and positive constituents become a single regional feature. While the complex model could in principle learn the dependency between such constituents, we prefer to treat well-known dipole or wave features as one, which gives room for the model to learn unknown links. To this end, we experimentally established a specific set of HDBSCAN parameters for each climate variable (Table 2). These parameters are minimum size of the feature region, minimum number of samples for a subregion to not be considered noise, and the distance ε below which a region is not split up any further. The different parameter values reflect differences in a variable's characteristic length scale, e.g., z300 being a much smoother atmospheric field than tcc, and the higher gridcell density of ERA5-Land as compared to ERA5.

After clustering, we extract two time series per feature region. First is the mean anomaly at each time step. Second is the covariance between the time-varying anomalies $x_{t,c}$ and the static correlation values r_c from the correlation field:

TABLE 2. Parameters in feature extraction: clustering gridded correlation patterns on a sphere. The Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) considers three parameters. Minimum final cluster size, minimum number of samples for a sub-clusters to not be considered noise, and the spherical distance ε below which clusters are not split up.

Variable	Min cluster size (cells)	Min samples (cells)	ε (radians)
t850	3000	1000	0.17
z300	6000	2000	0.17
tcc	400	200	0.09
sst	1000	300	0.20
siconc	400	200	0.10
swvl13	450	200	0.08
swvl4	450	200	0.08
transp	600	200	0.08
snowc	3000	1000	0.11

$$\text{covariance}_t = \frac{1}{n}\sum_{c=1}^{n}\left(x_{t,c} - \frac{1}{n}\sum_{c=1}^{n}x_{t,c}\right)\left(r_c - \frac{1}{n}\sum_{c=1}^{n}r_c\right), \qquad (4)$$

where n is the number of grid cells in the cluster. This covariance expresses the spatial coherence of anomalies with the corresponding correlation field. If anomalies are an exact copy (inverse) of the correlation pattern, that time step would be assigned a high positive (negative) value. Covariance thus ensures that dipole features, whose positive and negative anomalies would cancel in a cluster mean, are extracted, too. As the number of regions is data-driven, the feature extraction results in a total of N feature time series per lead time and training fold (Fig. 2c).

d. Machine learning model

Per lead time and cross-validation fold one ML model is fitted to predict either the weekly, biweekly or monthly binary target. Because of the way we defined the target, this prediction is simplified by climate change. The frequency with which temperature exceeds a fixed threshold increases with time due to thermodynamic effects (Vogel et al. 2020). Forecasts can make reliable use of this (Suarez-Gutierrez et al. 2020). Climate change is thus the simplest prediction a model can make, as it does not need to understand or leverage any other driver of high temperatures. Therefore, we define our "base" model as the Logistic Regression of climate change–driven probability p_{base} against time t (Julian day):

$$p_{\text{base}} = \frac{1}{1 + e^{-(\gamma_0 + \gamma_1 t)}}, \qquad (5)$$

where γ_0 and γ_1 are regression coefficients (see also Fig. 3a).

Subseasonal windows of predictability, being sourced from the features and interactions that we wish to research, would exist on top of this climate change signal (Hamill and Juras 2006). Our "full" machine learning model is thus defined as the complex function $g(X)$ with which a random forest

regressor (Breiman 2001) lowers or heightens the base probability, by leveraging the set of N data-driven features X:

$$p_{\text{full}} = p_{\text{base}} + g(X) \text{ with } p \in [0, 1]. \qquad (6)$$

A combination of a random forest on top of a base model has been applied earlier (Kirkwood et al. 2021). However, as seen in Fig. 3b, the target variable of this random forest regressor is not fully continuous. Instead it is the residual between (trended) binary observations o and the base probability that increases with time: $o - p_{\text{base}}$, where o is either 0 or 1, so the regression target is bounded by $[-1, 1]$. This approach is not elegant, because after addition any final negative probabilities need to be transformed to 0, but it works in practice. Other reasons for this approach, and possible alternatives, are discussed in section 4.

The set of about $N \approx 300$ input features at each lead time is large compared to the number of independent observations. Although daily resolution was retained by the rolling window averaging, the actual amount of non-overlapping observations ranges from about 3000 for forecasts of the 1-day average target, to about 100 for forecasts of the 31-day average target. Many statistical models are prone to overfitting with such low observation-to-feature ratios (100/300 at worst). Random forests, however, consist of an ensemble of decision trees (Fig. 3c) and are suited for this task (Wei et al. 2015). Each tree uses a random subset of data and features to split its target values into collections (also "nodes") with maximum homogeneity. Splitting rules are combinations of a feature and a threshold (see example rules in Fig. 3c). The trees then keep partitioning the data into smaller and smaller nodes until a stopping criterion is reached. As an ensemble, the trees converge to stable average estimates of the target, and have proven useful for meteorological applications (e.g., Taillardat et al. 2016; Whan and Schmeits 2018; van Straaten et al. 2018; Bakker et al. 2019; Hill et al. 2020; Mecikalski et al. 2021).

Individual trees are tuned by setting their maximum tree depth, the minimum number of samples required to split a node, and the number of features considered per split. The ensemble size is determined by setting the number of trees. These hyperparamaters are usually tuned during the training process and then kept fixed when the final model is trained. For each and every lead time, we constrained the tree depth to a maximum of 5, required a minimum number of 30 samples, set a random pick of 35 out of 300 features, and used 2500 trees (Fig. 3c). The first two settings are known to limit model capacity, resulting in shallower trees, which avoids overfitting in datasets where overfitting is possible (Segal 2004).

The first three hyperparameters were found by iteratively testing different combinations for performance in terms of the Brier score (section 2e), measured on the verification folds. This introduces the risk of selecting hyperparameters that may overfit the same verification folds, also used for final verification (section 2e). However, this risk was reduced by accepting only combinations that display a low generalization error (similar performance in both the training and the verification set). A test of our approach with an unseen dataset (a backward 1950–79 ERA5 extension) demonstrated the robustness of found hyperparameters (not shown).
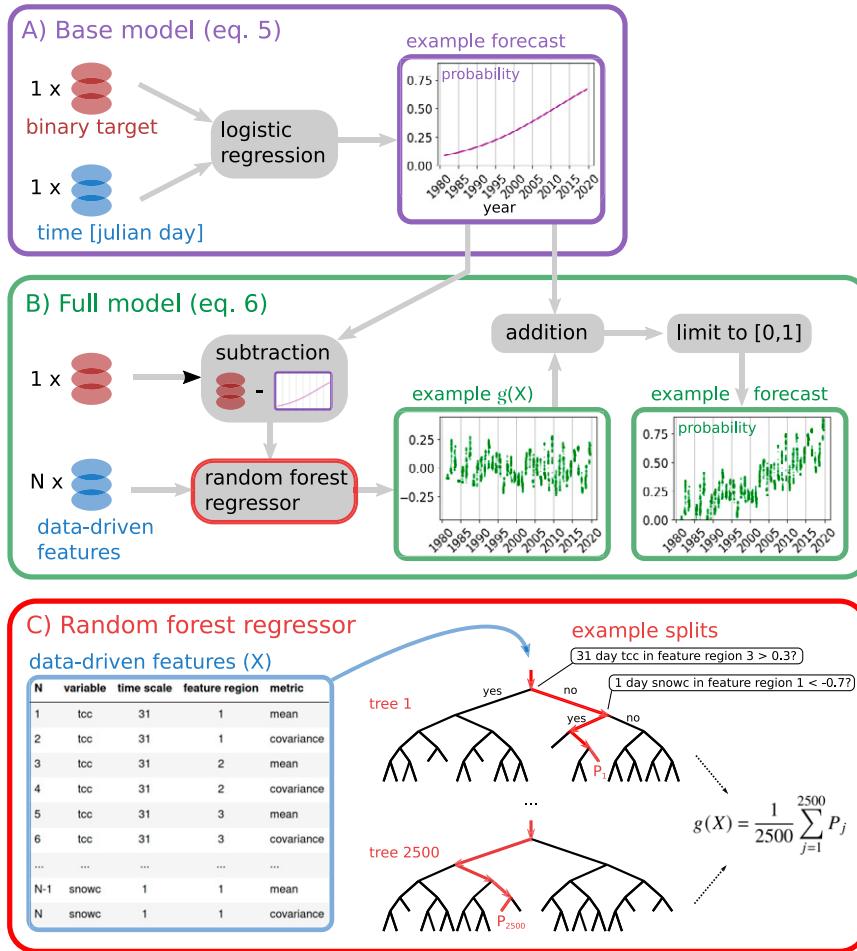
FIG. 3. Method for fitting the sources of predictability with an ML model that forecasts a temperature exceedance target (red-brown) from climate variable features (blue; see also Fig. 2). (a) Base model predicting the increased probability of events due to climate change [Eq. (5)]. (b) Full model predicting deviations from the base, by leveraging the large set of N data-driven features available at a single lead time [Eq. (6)]. (c) Random forest component of the full model, fitted per lead time and cross-validation fold, consisting of 2500 decision trees with a maximum depth of 5, each using a random selection of 35 features in its splits. The predicted value for $g(X)$ is the average over the trees. The table on the left side of (c) illustrates some of the N available features, for instance the three cloud cover features illustrated in Fig. 2c. Random forest diagram after Mecikalski et al. (2021).

Our fourth hyperparameter, i.e., 2500 for the number of trees, was chosen for practical reasons. The random pick of 35 (out of 300) features implies a 35/300 chance to be available for one of the top-most splits in a decision tree. The first splits dominate the predictions, which means that any feature has little chance to be important in a single tree (Wei et al. 2015). To distribute feature selection probability equally, we choose a large but computationally feasible number of 2500 trees.

e. Verification

The existence of subseasonal predictability can be evaluated by comparing full- to base-model performance. Both models produce probability forecasts for which we compute the Brier score (BS) over all forecast–observation pairs:

$$\text{BS}_{[\text{full,base}]} = \sum_{i=1}^{K} \Big(o_i - p_{i,[\text{full,base}]} \Big)^2, \qquad (7)$$

where K is the total amount of pairs present in the five verification folds. The BS is then converted to a Brier skill score: $\text{BSS} = 1 - (\text{BS}_{\text{full}}/\text{BS}_{\text{base}})$. Besides BS_{base} we also compare against the BS of the climatological frequency $\bar{o}$ observed over all data, which is a reference probability forecast that is commonly used. Uncertainty in the BSS can be large due to dependence between samples, and due to the relatively small number
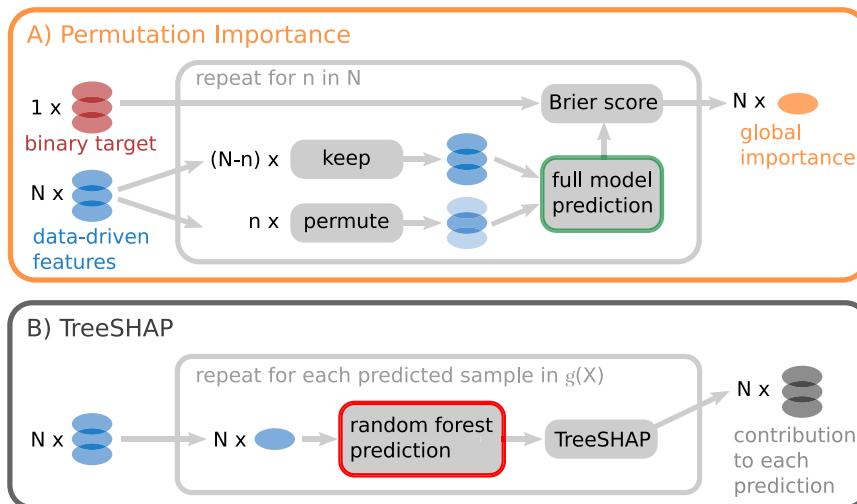
FIG. 4. Explainability tools to interpret fitted source of predictability. (a) Explanation of the full model with permutation importance. Sources of predictability are identified iteratively. They are the first *n* features that after random reordering of their time series result in the worst forecast scores. (b) Explanation of the full model's random forest regressor with TreeSHAP (see also section 2f).

of positive cases when the most extreme temperature threshold (90th percentile) is used to define the target variable. We therefore repeatedly recompute BSS on *K* forecast–observation pairs that are drawn at random with replacement, i.e., bootstrapping. The dependence between samples due to rolling window aggregation would cause uncertainty to be underestimated when we draw day-by-day. Therefore we draw in consecutive blocks, with sizes ranging from 1 to 60 days.

A single verification metric is often not enough to understand performance (Gneiting et al. 2007). We complement the BSS with reliability diagrams and an evaluation of forecast value. Reliability diagrams are graphic tools to assess a forecast's reliability and resolution (Wilks 2011). The potential economic value (PEV) is the value that a forecast has to a hypothetical decision-maker, compared to having no forecast available (Richardson 2000). The user's decision problem is characterized by a cost–loss ratio *c*, being the cost of taking action over the potential loss if no action is taken. PEV becomes

$$\text{PEV} = \frac{\min(c,\overline{o}) - Fc(1 - \overline{o}) + H\overline{o}(1 - c) - \overline{o}}{\min(c,\overline{o}) - c\overline{o}}, \quad (8)$$

where hit rate *H* and false alarm rate *F* are obtained from a contingency table after binarizing the forecast with probability thresholds 0.1, 0.3, 0.5, 0.7, and 0.9, and where $\overline{o}$ is the observed frequency of the event. We evaluate PEV for a range of cost–loss ratios between 0 and 1.

f. Explainability

Enhanced full model performance as compared to the base, can only occur when the full model has learned to leverage some of its input features as sources of predictability, either direct, or as an interaction on multiple time scales. We investigate what the model has learned in two ways, one being the permutation importance of each feature for the overall correctness of full-model predictions (Fig. 4a), the other being the contribution of each feature to a low or high forecast probability, as quantified by TreeSHAP, an application of SHapley Additive exPlanations specifically designed for tree-based methods (Lundberg et al. 2020) (Fig. 4b).

Permutation importance quantifies the decrease in performance over all predictions when a feature is wrongly assigned (i.e., permuted) (Breiman 2001; Lakshmanan et al. 2015). This means it results in one "global importance" per feature, "global" meaning "over all samples." We express the decrease in performance in terms of BS. These BS values depend on lead time, so to equalize situations far-before and close to the event, we rank importance within each model from 0 to 1 (from least important, lowest increase in BS, to most important, highest increase in BS). We permute in a repeated "multipass" manner, which, as opposed to "single-pass," can better discriminate between correlated features (Lakshmanan et al. 2015; McGovern et al. 2019). It involved iteratively searching and permuting the next-most important feature, given a set of already permuted features, until a total number of *n* = 30 features were found (Fig. 4a). We also found little difference in the results of "multipass" and "single-pass," when evaluated on training data.

The application of TreeSHAP to our random forest produces a different, "local" measure of importance (Lundberg et al. 2020). In each sample, the random forest receives *N* feature values, and produces a single prediction *g*(*X*) [see Eq. (6) and Fig. 3c]. TreeSHAP is a method originating from game theory that can attribute a game's single outcome to the contributions from each player. In this case it computes the set of positive
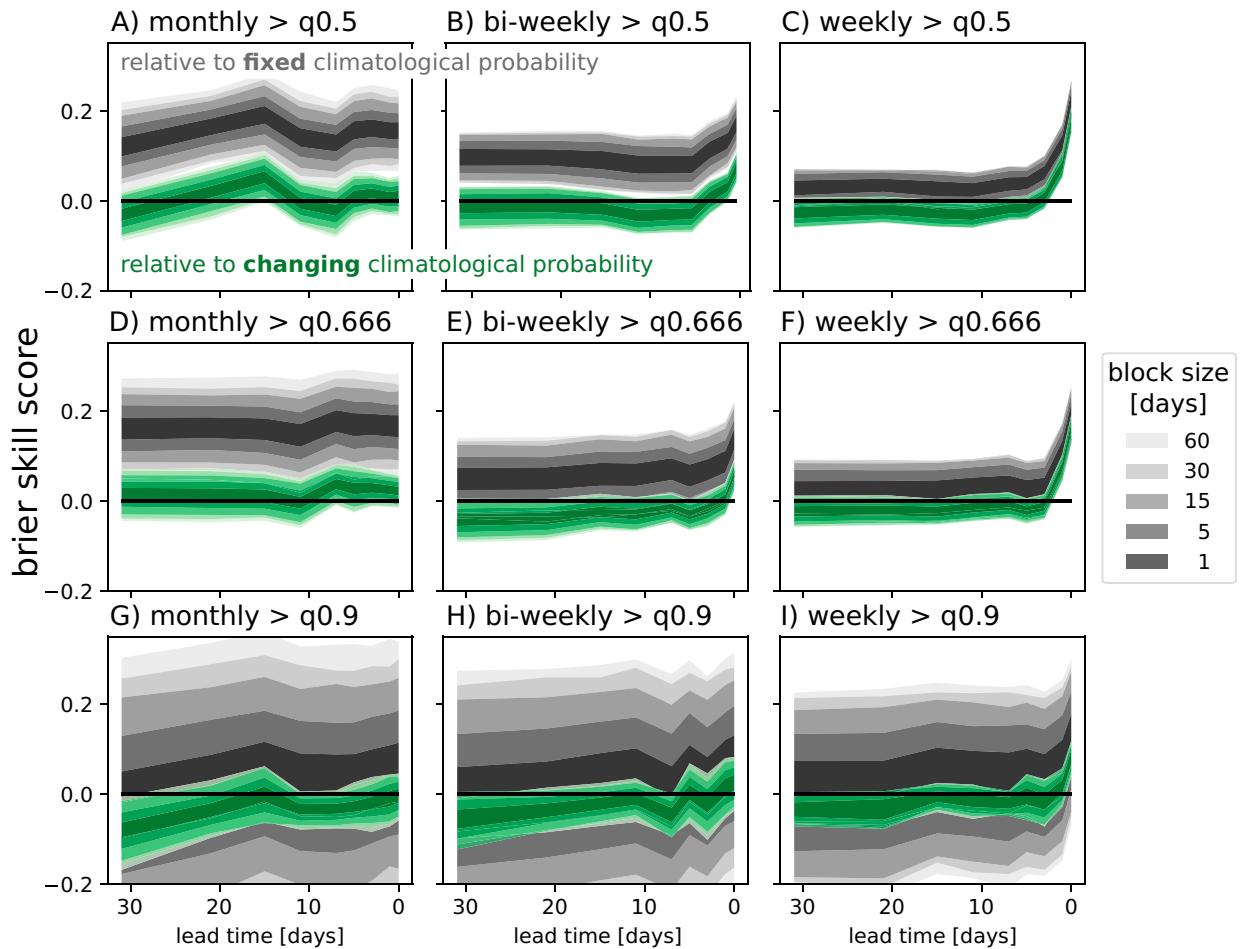
FIG. 5. BSS-based identification of skillful lead-time windows. Full model machine learning forecasts are made for different target events, being mean temperature in a given averaging window (columns), exceeding a given climatological quantile threshold (rows). Performance is measured as a function of lead time and relative to two different reference forecasts. Gray: relative to the event's fixed probability of occurrence over all samples. Green: relative to the gradual change in probability due to climate change, as forecast with the base model. Values above the zero line indicate positive skill. Uncertainty in the metric is illustrated with the 5th–95th-percentile uncertainty bounds, obtained by bootstrapping available samples with different block sizes (see legend, 5000 repeats).

and negative contributions from all features, that together add up to the predicted probability (more details in Lundberg et al. 2020). When repeated for all samples, we obtain a time series of contributions for each feature. We use these series in our case-study of summer 2015, but also extract a measure of global importance by averaging the absolute values of all contributions in each series.

3. Results

a. Verification and predictability

We consider the performance of the full machine-learning model, fitted to forecast different types of target events at multiple lead times. Events are mean temperatures in a given averaging window (columns in Fig. 5), exceeding a given temperature threshold (rows in Fig. 5). Since the threshold is fixed in time, we expect a big part of the total predictability of

events to come from climate change. We compare the full model BSS computed with two different reference forecasts (Fig. 5).

The first reference forecast is the commonly used climatological event frequency over all samples in the verification sets. It assumes that the probability of occurrence is fixed, and not low in the beginning and high toward the end, as in reality. Relative to this reference, the full model shows positive BSS at many lead times (gray shading in Fig. 5). At extended lead times (15–31 days) the monthly target shows larger BSS values than other targets, presumably because noise has been suppressed by a larger averaging window, increasing the usability of the climate change signal in forecasts (e.g., Fischer et al. 2013). We confirm the contribution of climate change to predictability with BSS relative to the base reference. This reference does model the gradual change in probability. Skill at extended lead times is hardly different from zero (green

shading in Fig. 5), meaning that when we define subseasonal predictability as "the ability to forecast deviations from the climate change signal at extended lead times," it is low.

The low amount of apparent subseasonal predictability can be expected. First, it is characteristic of the target region (Prodhomme et al. 2022). Second, the occasional part of predictability that exists in forecasts of opportunity might be masked by the fact that we use all samples to compute BSS (Mariotti et al. 2020). Still, BSS does indicate a skillful lead-time window in forecasting median-exceedance in the monthly target. The full model BSS is higher at a lead time of 15 days than at other, also shorter lead times (Fig. 5a). Usually we expect forecasts to be more skillful for shorter lead times, as events extend less far into the future. The situation gets for instance increasingly certain from a 5- to 1-day lead time (Fig. 5c). But this is not the case for the monthly target (Fig. 5a).

Differences in skill relate to the discriminatory power of the forecasts. The skillful window at a lead time of 15 days suggests that at 15 days before the event a physical link from input features to the target can be leveraged and distinguishes high probabilities from low probabilities of monthly exceedance. In verification terms, the model would temporarily show a better "resolution" at this lead time (for a mathematical definition of resolution see Wilks 2011).

In Fig. 6 we plot reliability diagrams for the skillful monthly exceedance of the median, 66th and 90th percentiles, forecast with a 15-day lead time. The higher quantiles are of interest to explore the predictability of upper tail-events with metrics that provide a more complete picture than BSS (as in Dorrington et al. 2020). The reliability diagrams show that the upper-tail forecasts of the full model outperform the base model (Figs. 6d,e,g,h). A reliability diagram compares the forecast probability with the observed frequency: for a forecast probability p ($0 \leq p \geq 100$ percent), the event should be observed in p percent of the cases. Forecasts that are reliable in that sense lie on the diagonal 1:1 line. Panel D shows that the base model probabilities range from about 0.1 to 0.6 and are close to the 1:1 line. This implies that the forecasts are reliable, but not that the forecasts are perfect (which would only be realized when binary probabilities of either 0 or 1 were issued). The full model's range is wider than the base model, with for example probabilities of 0 or 0.8 being more frequently issued (Figs. 6d,e). Since the full model remains close to the 1:1 line and has widened the probability range, we can conclude that it has reliably increased the resolution of forecasts, on top of the climate change signal.

That the full model adds value to the base model is visible in the vertical difference between their PEV curves (Fig. 6f). Base model upper tercile forecasts are valuable for decision-makers with cost–loss ratios ranging from 0.1 to 0.6. The full model widens this range, and especially adds value for users with cost–loss ratios < 0.2 (typical for many real-world users). Also for predictions of extremer events, namely, exceedance of the 90th percentile, value is added (Fig. 6i). We see that the full model has learned to issue probability forecasts up to 0.6, compared to the maximum of 0.3 in the base model (Fig. 6h). This extension is not perfectly reliable (Fig. 6g shows

deviations from the perfect reliability curve), but is still adding value to users with cost–loss ratios of 0.2–0.5 (Fig. 6i). The useful increase of the resolution shown in Figs. 6d,e also extends to monthly forecasts at different lead times (not shown). This performance improvement cannot be derived from BSS values alone.

The ability to leverage features for forecasting is expected to not only depend on lead time but also on the properties of the target. The full model BSS at a 15-day lead time is higher for the monthly target (Fig. 5a) than for the biweekly target (Fig. 5b). The former event extends 46 days into the future (15-day lead time and 31 days of event), while the latter event extends 30 days into the future. Again we could expect the shortest extension into the future to be the most certain. However, in Fig. 7 we see that forecasts of the monthly (solid green) get closer to perfect reliability, and with a wider range of probabilities, than forecasts of the biweekly (dashed green), using the exact same set of features extracted on eight time scales. It needs saying that the monthly target, due to the larger averaging window, enables the full model to use a more apparent climate change signal in its base (not shown). But still the results suggest that driving features exist and are leverageable in especially this skillful lead-time window. Predictability in a monthly target thus does not need to stem from successful prediction of its first 2 weeks.

b. Sources of predictability

The convincing resolution enhancement in the reliability diagrams at a 15-day lead time, and the moderate enhancement at other lead times, lead to a logical question: which of the features has the ML model learned to leverage as sources of predictability? We consider this question over a range of lead times.

We expect the learned relations to depend on lead time in two ways. Physically we expect a transfer of predictability across variables, from oceanic RWP sources and antecedent land surface conditions long before the event, to the atmospheric state close to the event. Second, we expect a transfer of predictability across time scales: Multiple time scales are involved in lead up to the event, and close to the event we expect the short time scales, representing the state closest in time to the event, to become dominant in the interaction. If per the second expectation, all forecasts would be dominated by a time scale equal to the lead time, then all panels in Fig. 8 would look like the top-left example. Unused features are in blue and leveraged features are in orange and black (permutation importance and TreeSHAP, respectively).

We first discuss the forecasts of monthly temperature, exceeding the 66th-percentile threshold (left columns, Fig. 8). Judging by the amount of black and orange dots, and the patterns that both metrics agree on, SST is the largest source of predictability, followed by snow cover, 850-hPa temperature, and sea ice concentration. Of the antecedent land surface conditions, deep soil moisture is a more important source than shallow soil moisture and transpiration. Interestingly, we also see that the black dots are distributed horizontally, along the 21- and 31-day feature time scale, instead of diagonally like in
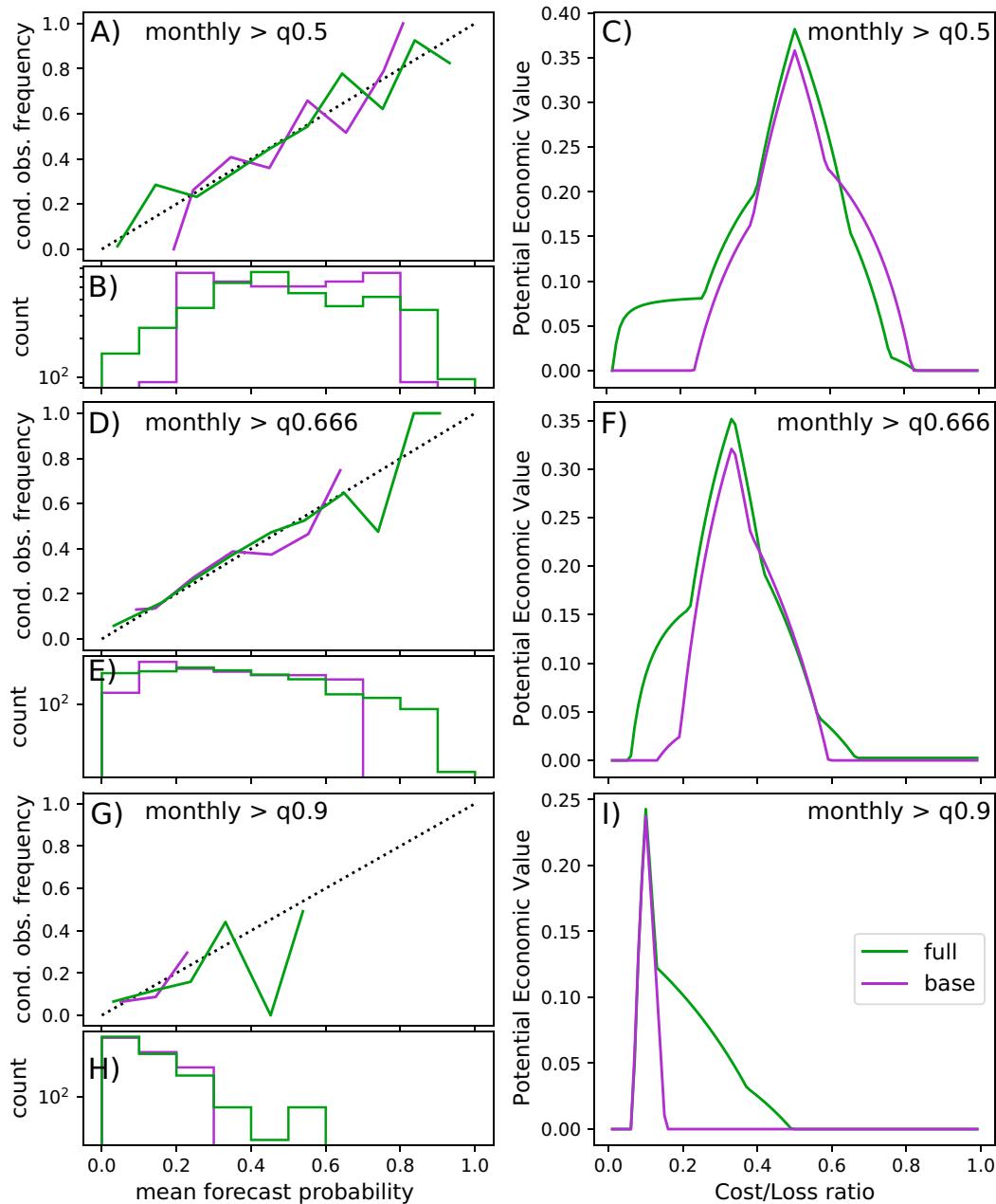
FIG. 6. Verification of probabilistic machine learning forecasts for an identified skillful lead-time window. Forecasts are for monthly temperature exceeding the (top) 50th, (middle) 66th, or (bottom) 90th percentiles made with a 15-day lead time. The full model (green) is compared to the base model that forecasts only the gradual change in probability due to climate change (purple). (a),(d),(g) Reliability diagrams of the conditional observed frequency per bin vs binned forecast probabilities. Perfect reliability is depicted by the dotted 1:1 line. (b),(e),(h) Histograms of forecast probabilities. Given good reliability in the panel above, a wider histogram shows better resolution. (c),(f),(i) The maximum potential economic value over probability thresholds 0.1, 0.3, 0.5, 0.7, and 0.9, for users acting on these forecasts in the context of different cost–loss ratios (x axis). Vertical difference between green and purple shows value added by the full model.

the example. It means that the model prefers the longer time scales despite getting closer to the event with decreasing lead time. The likely reason is that the monthly target still extends far into the future. The necessary long-term information is not sufficiently present in the short-term states. With t850

being an exception, this statement applies to atmospheric features at all time scales. Especially TreeSHAP shows that z300 and tcc do not contribute to forecasts of the 31-day target (almost no black dots in the left atmospheric column of Fig. 8).
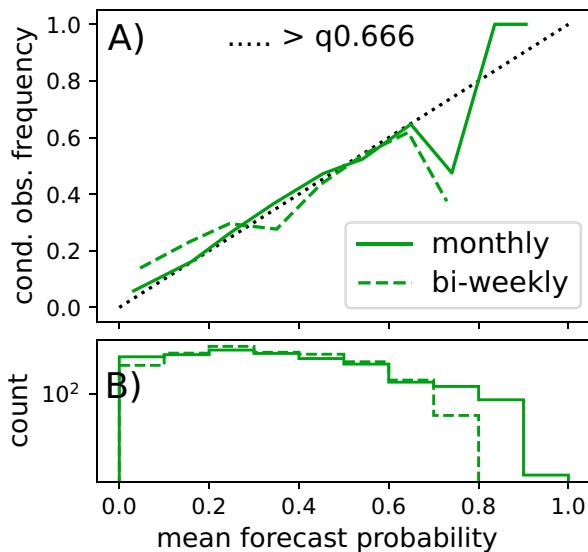
FIG. 7. Time-scale dependence for an identified skillful lead-time window. As in Figs. 6d,e, but with full model forecasts for temperature exceedance in two different averaging windows: monthly (solid) and biweekly (dashed), made with a 15-day lead time.

This changes when we look at predictions of the shorter weekly target (right columns, Fig. 8). SST, snow cover and deep soil moisture still are dominant sources of predictability at lead times longer than 5 days. The model, however, does not use the 21- and 31-day feature time scales exclusively. The black and largest orange dots now lie in the window range of 10 to 31 days for SST, 5–21 days for snow cover and deep soil moisture. At lead times shorter than 5 days, when the event gets closer, features from z300, followed by t850 and transpiration, become the dominant sources of predictability: they increase or decrease the chance of weekly high temperature events. In other words, the ML models confirm, by being able to pick and learn freely, that information transfers from longer-term oceanic and land surface conditions to the atmospheric state.

Before further physical interpretation, we consider the difference between the two importance measures. Usually, the highest ranking variables according to permutation importance, i.e., the largest orange dots, are also important according to TreeSHAP and accordingly accompanied by a black dot (Fig. 8). The two methods often do not agree on lower ranking variables. It appears from Fig. 8 that average absolute TreeSHAP is a stricter measure of global importance, whereas permutation importance admits a more wide-spread set of important features and time scales. Part of the reason lies in the difference between rank and TreeSHAP contribution. A feature that is conditionally the n-most important variable, might not contribute noticeably to forecast probability in each and every sample. Qualitative differences in emphasized features can also follow from permutation importance itself. As permutation breaks dependencies between features, it can move the model to situations it was not trained for, leaving one to interpret extrapolation behavior instead of

normal predictive links (Hooker and Mentch 2019). Consequentially, we have most confidence in the patterns that both measures agree on.

The visualized importance in Fig. 8, like the surprising usability of 31-day average atmospheric t850, are linked to specific regions. Each dot is, namely, the maximum global importance of the multiple possible feature regions with two time series each, i.e., one for the mean and one for the covariance, that were all at the full model's disposal. So in Fig. 9 we map where the 31-day average input features of SST, t850, snowc, and siconc are important for predicting monthly temperature exceedance of the 66th-percentile threshold at a 15-day lead. Features from SST are present in a large portion of its domain, as many grid cells correlate significantly to temperature, even after autocorrelation and linear trends have been accounted for (Fig. 9a). Only a selection of these cells are robust and present in at least 4 of the 5 sets of training data (orange to yellow). Robust groups of cells for instance appear east of the Maritime Continent, suggesting that Pacific SST variability in that region can be important. However, such a potential relation does not imply that a feature will be a useful source of predictability. So for SST and the other variables (in rows), we plot each feature's permutation and TreeSHAP importance, for the robust cells only, averaged over at least 4 of the 5 cross-validation sets (Fig. 9, middle and right column). Consequentially, the plotted shapes do not perfectly resemble the feature regions of a single subset.

Features from SST (Fig. 9b) are stronger sources of predictability than features of other variables (Figs. 9e,h,k). But not all SST cells will be of equal relevance. Especially smaller patches are often part of a feature region like the whole Indian Ocean, and therefore share the importance of a predictive signal coming from large patches that dominate a feature's mean or covariance. Focusing for that reason on large contiguous patches, we notice that besides the mentioned Pacific features, the Indian Ocean provides an extensive source of predictability, and that a lesser source of predictability lies off the east coast of South America (Figs. 9b,c).

One of T850's predictive features is collocated with an SST feature over the Indian Ocean (Figs. 9e,f). Their coincidence off India's west coast hints at monsoon dynamics, which previously have been found to affect Euro-Atlantic summer circulation (Beverley et al. 2019). The most surprising T850 feature is the region extending from the tropical Atlantic into the western Sahara. In T850, the feature mean is more predictive than feature covariance, which in combination with climate change, gives the impression that the feature is leveraged to explain a thermodynamic trend. However, our full model learns deviations from the climate trend (i.e., deviations from the base model), so this is likely not the case. Without a causal framework it remains speculative how this source of predictability, present at the relatively low level of 850 hPa, affects western and central European temperature. Given its location a link to upper-level disturbances in the tropical Atlantic and Sahel region is expected (Cassou et al. 2005; Nakanishi et al. 2021). Crucially, it would not have been discovered, had we not applied a data-driven dimension
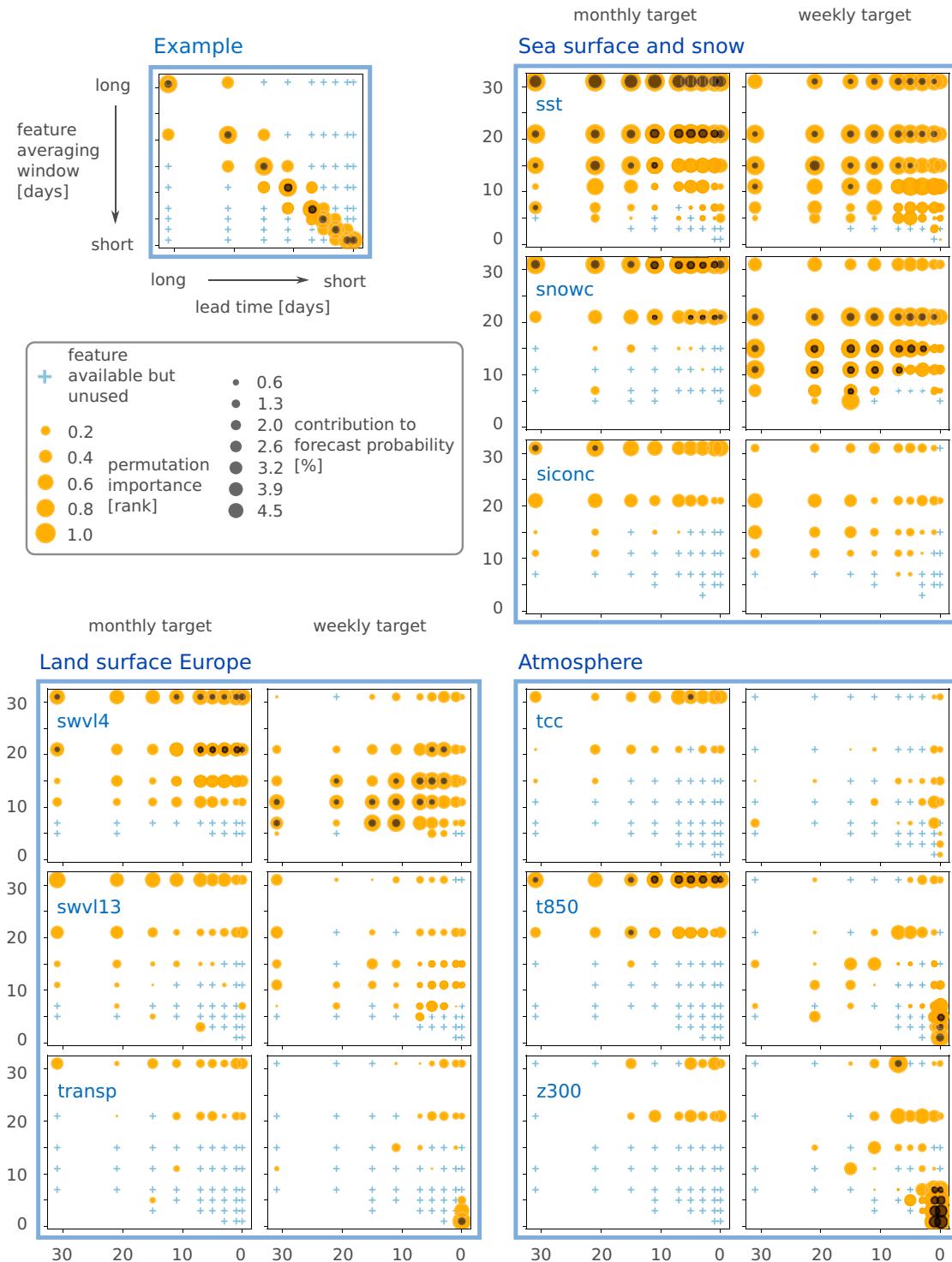
FIG. 8. Sources of predictability learned by the full model. Features from nine climate variables are available on eight time scales, depending on lead time. Their importance is revealed by two metrics: average absolute TreeSHAP (black) and multi-pass permutation importance (orange). Unused features are in blue. The emphasized time scales (*y* axis) and climate variables (rows) vary with the lead time at which the forecast model operates (*x* axis) and the type of forecast target. The two targets are 66th-percentile exceedance in (left) monthly temperatures and (right) weekly temperatures. TreeSHAP and permutation importance values are the maximum over all features per time scale and climate variable (comprising multiple time series of mean and covariance from multiple feature regions). TreeSHAP contributions below 0.5% are not plotted.
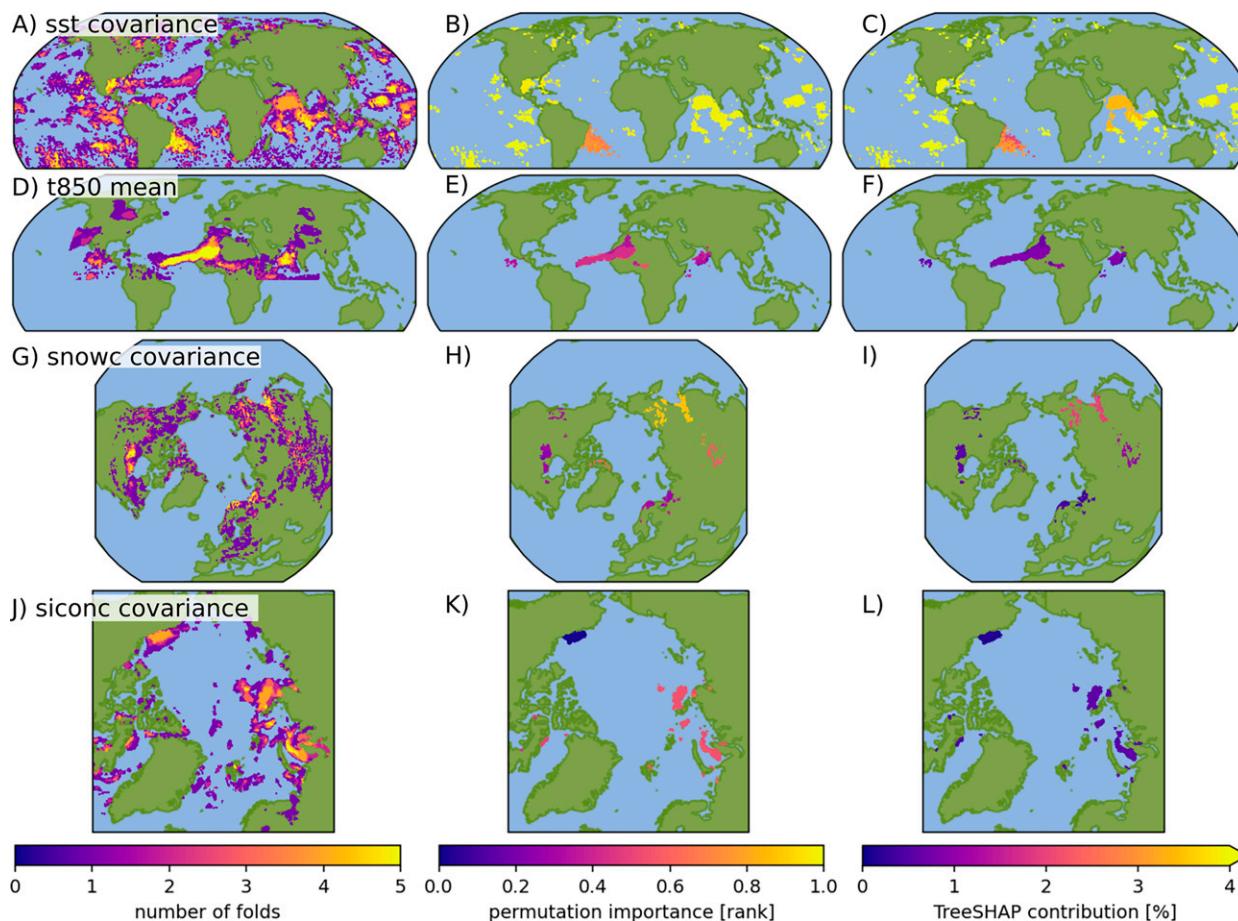
FIG. 9. Geographic distribution of sources of predictability, learned by the full model. (from top to bottom) Visible are the feature regions in SST, T850, snow cover, and sea ice concentration that the model leverages to predict monthly average temperature exceedance with a lead time of 15 days (threshold is q0.66). The climate variables' grid cells, belonging to distinct regional features, are colored (middle) with each feature's average permutation importance and (right) with the average absolute TreeSHAP contribution to the forecast probability. (left) The importance values are presented only for grid cells that are found significant in at least four of the five cross-validation folds. The type of time series used by the model is annotated (mean or covariance). Time scale of the features is also monthly.

reduction and let our ML model process many features with little a priori selection.

The location of snow-cover features differs from fold to fold. This is visible in the minor extent to which they overlap (Fig. 9g). Still, these are important sources of predictability (Figs. 9h,i). Importance of Eurasian snow-cover anomalies trumps that of the North American ones (Hall et al. 2017). And within Eurasia the regions located farther east are more important than those located in the west. The most eastern snow-cover feature is also more important than any of the features in sea ice concentration (Figs. 9h,k), which is the reverse of the order suggested by Zhang et al. (2020). Nonetheless, within sea ice concentration, the ML model has diagnosed the relative importance of the Kara Sea, whose decreased ice concentration is known to relate to a more southern and stronger polar front jet (Hall et al. 2017), inhibiting stagnant high pressure systems over the target region.

c. Summer of 2015

Physical interpretation of sources of predictability remains limited when, like above, emphasized features are seen, but the sign of their predictive relations are not. One needs to know whether a specific emerging anomaly inhibits or increases the likelihood of an event, by how much, and what the state of the other features is, since links can be conditional. This information is usually hard to extract from complex ML models. With so-called local explanations of individual forecasts we can uncover such details (see also Lundberg et al. 2020; Gibson et al. 2021). When a feature's TreeSHAP value shows increased forecast contribution from one point in time to the next, we can trace that in real time to a specific set of shifting anomalies, that for instance has started to resemble the feature's underlying correlation pattern. We demonstrate such a breakdown of contributions with forecasts of the hot summer in 2015.

The summer of 2015 was characterized by exceptional temperatures in western and central Europe that were clustered in two intense periods. Duchez et al. (2016) found that a cold North Atlantic SST anomaly from the months before might have displaced the subtropical jet to a stationary southern position, favoring buildup of heat over the continent. They find that the displacement commenced in the last days of June and persisted into September. High temperatures followed on 1–6 July. Despite the claim that the temperatures were driven by the cold SST anomaly, this first period was hard to predict by an operational ensemble (Ardilouze et al. 2017). The high temperatures were followed by a rainy intermission, before a second temperature peak started on 16 July, for which Wehrli et al. (2019) found that SST was of low importance. Their analysis also showed that although soil moisture (on a monthly scale) was decreased during the event, no significant feedback to atmospheric heat was found.

For both peak periods, we examine the features contributing most to the probability forecasts of the ML model. A horizontal bar (Fig. 10a) shows contributions to the probability that monthly temperature exceeded the 66th percentile, from 6 June to 6 July, a period that encompasses the first heatwave period at its end. Below is the forecast that encompasses the second period at its beginning (Fig. 10f). The forecasts were made with a 15-day lead time, which was shown to have predictive value in section 3a.

For both periods the full model raises forecast probability above the 59% that we expect from the climate change base model (Figs. 10a,f). The increase occurs because the joint state of all driving features produces larger positive contributions (pink), than negative inhibitory contributions (blue, Figs. 10a,f). In both Figs. 10a and 10f the largest positive contributions come from the state of 21-day average SST, in feature region 4, and from 31-day average SST, in feature region 1. This suggests the influence of a long-term SST anomaly (Duchez et al. 2016). However, region 1, where the 31-day covariance signal originates, encompasses more than just the Atlantic. When we look at the respective SST anomaly (Fig. 10c) we see the cold North Atlantic temperatures south of Iceland. A close look at the underlying correlation pattern (Fig. 10e) shows that at this location, no significantly negative, even slightly positive SST anomalies are associated to higher target temperatures, for a driving effect at this lead time. The positive contribution of covariance (resemblance to an underlying correlation pattern), has thus to be sought elsewhere, for instance in the positive anomalies extending from the Gulf of Mexico eastward. Not shown for the 31-day state, but shown for the 21-day state, is the feature east of the Maritime Continent (Fig. 10b). The correlation pattern here (Fig. 10d), in SST's feature region 4, is partly resembled by the SST at this point in time. Seeing the model put forward such an emergent feature from a relatively small region, human forecasters can study its trustworthiness. The influence of the feature does remain the second largest factor also for the second part of the summer (Figs. 10g,j)

What does change from the first to the second high-temperature period, according to our model, is the strength of inhibiting drivers, that together decrease forecast probability from 0.74

to 0.68 (Figs. 10a–f). The state of 31-day average sea ice concentration is dominant in this. We see that during the first predicted period (10 June–30 June), a negative sea ice concentration anomaly was present in the Kara Sea (Fig. 10i) which, as discussed in the last section, can project negatively on the target temperatures (Fig. 10l). Accordingly the full model lowered probability slightly, though still kept it elevated with respect to the base model. Exceedance of the threshold did happen in both high-temperature periods of the summer of 2015.

4. Discussion and conclusions

It is understood that there is a large number of climate variables and time scales involved in lead up to high summer temperatures. In this study we have extracted data-driven features from nine variables, varying on eight time scales. Interactions in such a set can source subseasonal predictability, but often at a complexity level that is beyond direct human understanding. We have explored whether an ML model can integrate and discover the sources of predictability for us. To our knowledge this has been the first attempt using such a large set of features.

Relative to the amount of features, the ERA5 dataset provides a low amount of samples. Whereas the use of ERA5 helps to account for processes that are inadequately captured in numerical simulations, a full integration of important subseasonal interactions is a great challenge for machine learning models (He et al. 2021). Combined with the limited samples to train on, a machine learning model does not attain the maximum skill possible. In certain lead time ranges, operational numerical models in combination with statistical postprocessing, will probably do better (e.g., Ferrone et al. 2017; van Straaten et al. 2020). But as this study has demonstrated, even purely statistical forecasts for west and central Europe, can be reliable and valuable in certain windows, like forecasting exceedance of monthly summer temperature with a lead time of 15 days (Fig. 6). We have further shown that this long-term predictability is not due to successful prediction of its short-term (2-week) constituents (Fig. 7).

Limited data makes certain types of statistical association hard to estimate. In feature extraction we correlated climate variables with temperatures (Fig. 2c). This appears suboptimal, as the features are later supposed to predict only the upper tail of the temperature distribution, namely, binary exceedance of a relatively high threshold. Rank correlation remains usable because nonlinear associations specific to the upper tail are accounted for, as long as they are monotonic. Better suited quantities like extremal-dependence metrics (Coles et al. 1999) and Kendall's Tau weighted toward the tail of the distribution, were also tried, but their estimates were found to be too unstable for this amount of data. An interesting alternative to our correlation-based dimension reduction is to directly apply a predictive ML model to the raw input data (e.g., He et al. 2021). In our setting with large domains, and nine climate variables for eight time scales, that will be challenging, as even after dimension reduction a low observation-to-feature ratio was obtained.
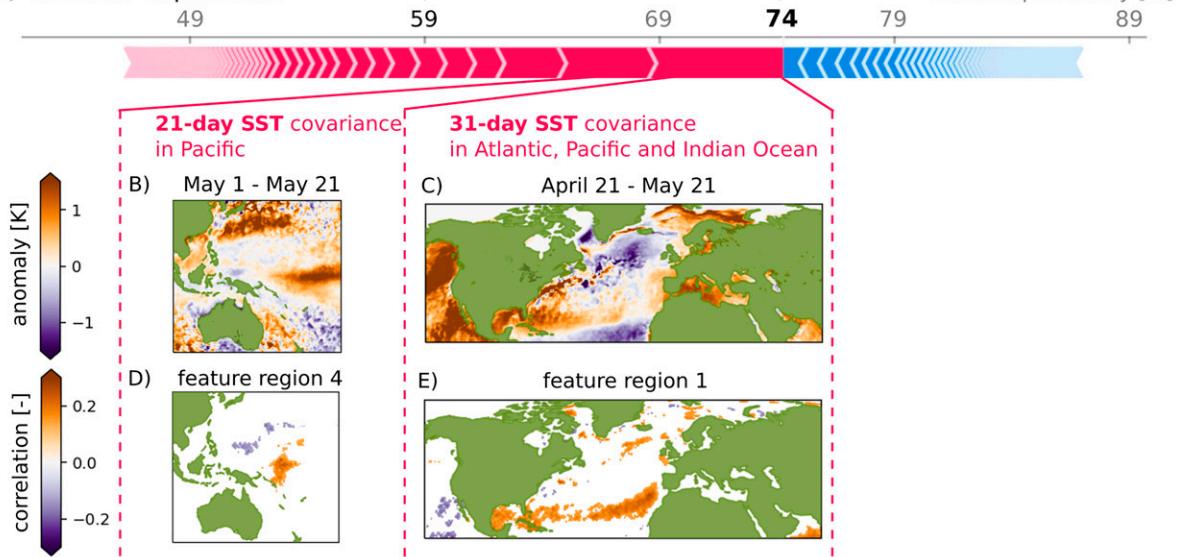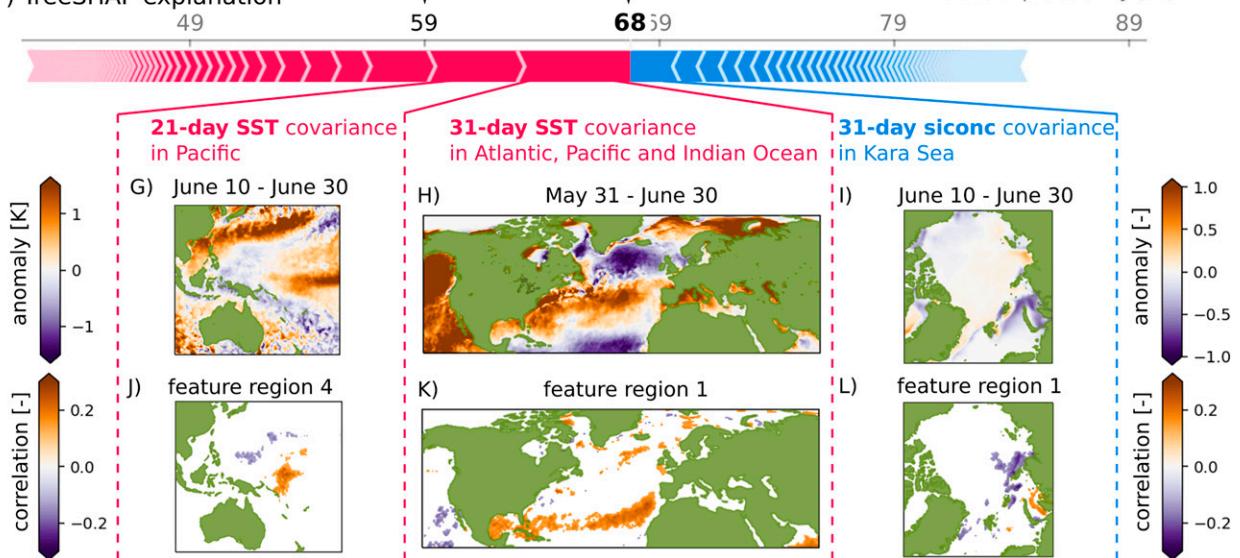
FIG. 10. Summer 2015. Explaining the forecast probability that monthly temperature exceeds the 0.66 quantile (top) from 6 Jun to 6 Jul and (bottom) from 16 Jul to 15 Aug, made with a lead time of 15 days. (a),(f) TreeSHAP explanation of the full model forecast, i.e., the decomposition of the forecast probability (boldface) into additive contributions from all driving features. Features increasing the probability are in pink, features decreasing the probability are in blue. The type (i.e., covariance) and time scale of the largest contributors are annotated. (b),(c) Anomalies leading to the two largest contributions. (d),(e) Correlation patterns underlying the contributing features. (g)–(i) As in (b) and (c), but including the largest negative contributor. (j)–(l) As in (d) and (e), but including the largest negative contributor.

A suitable ML model to identify sources of predictability in a set with a low observation-to-feature ratio is a random forest (Wei et al. 2015). The resulting picture of importance might, however, be more diffuse than the importance in reality. We had to mitigate the large number of features by drawing 35 at random. This can make single dominant sources of predictability (e.g., from 31-day SST) unavailable at prime splits in the decision trees, a role that is then taken by correlated features (e.g., from 21-day SST). This is also the reason that TreeSHAP contributions have small values. After computation of contributions, even the most important source of predictability alters the forecast probabilities by a mere 5%

on average (Fig. 8). Clear importance patterns have nevertheless emerged. The importance of long-term variability in SST is in line with physical understanding. Surprising is the dominance of deep soil moisture over shallow soil moisture, at all targets and time scales, and that of snow cover over sea ice.

An inherent challenge for subseasonal forecasts is the nonstationarity of driving features. For sea ice it is known that a link to European temperature can exist at certain moments in time (Kolstad and Årthun 2018). When such a potential for predictability is not systematic over all possible samples, it can happen that empirical models are trained on a subset with its presence and will make false forecasts on a set without. TreeSHAP can show in real time whether such features keep contributing to the forecasts (Fig. 10). A later computation of permutation importance can then reveal whether the important TreeSHAP contributions were valid or not. This needs a series of observed outcomes first, but features that have already started to degrade the scores will show no importance after further permutation. We expect that differences between TreeSHAP and permutation importance (Fig. 8) might inform one about potential non-stationarity.

We have treated subseasonal predictability as "deviations from a changing climatology that are predictable at extended lead times." As in other studies (Dole et al. 2014; Prodhomme et al. 2022), this separation from the trend is influential (Fig. 5), because climate change provides a large part of total predictability, and is even valuable to certain users (Fig. 6). Our first attempt at separation involved detrending average temperature before creating the binary (threshold-exceedance) target. This distributes events uniformly over time, and thus supposes that moderate temperatures of the past and current more extreme temperatures comprise a homogeneous class with similar dynamics. Depending on the definition of extremity, this is likely not the case (Vogel et al. 2020). It led to bad performance in the last decade. We therefore modeled the probabilistic deviations as additions to a base model [Eq. (6); Fig. 3b]. More elegant tools than subtraction and addition exist, like logarithmic transformations (e.g., Scheuerer et al. 2020) and Bayesian methods. But unfortunately those were not (yet) compatible with the random forests we needed to handle the low observation-to-feature ratio.

There is good reason to keep applying complex ML-methods to subseasonal prediction. First, we have shown that such a method can reliably increase forecast resolution by leveraging features from reanalysis data. Second, we demonstrated that an explainable method gives conceptual grip on the complexity. It confirmed many physical expectations, like the weighing of time scales and the transfer of information across variables with lead time. It also discovered surprising features like the long-term predictability originating from 850-hPa temperature. Overall, the associative learning of an ML-method will complement research with NWP models (e.g., Quinting and Vitart 2019). It shows which links need further understanding, and which variables need correct representation in NWP models for better future forecasts.

REFERENCES

Albers, J. R., and M. Newman, 2019: A priori identification of skillful extratropical subseasonal forecasts. *Geophys. Res. Lett.*, **46**, 12 527–12 536, https://doi.org/10.1029/2019GL085270.

Ardilouze, C., L. Batté, and M. Déqué, 2017: 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, https://doi.org/10.5194/asr-14-115-2017.

Bakker, K., K. Whan, W. Knap, and M. Schmeits, 2019: Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation. *Sol. Energy*, **191**, 138–150, https://doi.org/10.1016/j.solener.2019.08.044.

Barriopedro, D., E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera, 2011: The hot summer of 2010: Redrawing the temperature record map of Europe. *Science*, **332**, 220–224, https://doi.org/10.1126/science.1201224.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, **57**, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Beverley, J. D., S. J. Woolnough, L. H. Baker, S. J. Johnson, and A. Weisheimer, 2019: The Northern Hemisphere circumglobal teleconnection in a seasonal forecast model and its relationship to European summer forecast skill. *Climate Dyn.*, **52**, 3759–3771, https://doi.org/10.1007/s00382-018-4371-4.

Black, E., and R. Sutton, 2007: The influence of oceanic conditions on the hot European summer of 2003. *Climate Dyn.*, **28**, 53–66, https://doi.org/10.1007/s00382-006-0179-8.

Bladé, I., B. Liebmann, D. Fortuny, and G. J. van Oldenborgh, 2012: Observed and simulated impacts of the summer NAO in Europe: Implications for projected drying in the Mediterranean region. *Climate Dyn.*, **39**, 709–727, https://doi.org/10.1007/s00382-011-1195-x.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brunner, L., G. C. Hegerl, and A. K. Steiner, 2017: Connecting atmospheric blocking to European temperature extremes in spring. *J. Climate*, **30**, 585–594, https://doi.org/10.1175/JCLI-D-16-0518.1.

Casanueva, A., and Coauthors, 2019: Overview of existing heat-health warning systems in Europe. *Int. J. Environ. Res. Public Health*, **16**, 2657, https://doi.org/10.3390/ijerph16152657.

Cassou, C., L. Terray, and A. S. Phillips, 2005: Tropical Atlantic influence on European heat waves. *J. Climate*, **18**, 2805–2811, https://doi.org/10.1175/JCLI3506.1.

Cohen, J., D. Coumou, J. Hwang, L. Mackey, P. Orenstein, S. Totz, and E. Tziperman, 2018: S2s reboot: An argument for greater inclusion of machine learning in subseasonal to seasonal forecasts. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e00567, https://doi.org/10.1002/wcc.567.

Coles, S., J. Heffernan, and J. Tawn, 1999: Dependence measures for extreme value analyses. *Extremes*, **2**, 339–365, https://doi.org/10.1023/A:1009963131610.

Coughlan de Perez, E., and Coauthors, 2018: Global predictability of temperature extremes. *Environ. Res. Lett.*, **13**, 054017, https://doi.org/10.1088/1748-9326/aab94a.

Della-Marta, P. M., J. Luterbacher, H. von Weissenfluh, E. Xoplaki, M. Brunet, and H. Wanner, 2007: Summer heat waves over Western Europe 1880–2003, their relationship to large-scale forcings and predictability. *Climate Dyn.*, **29**, 251–275, https://doi.org/10.1007/s00382-007-0233-1.

Dole, R., and Coauthors, 2014: The making of an extreme event: Putting the pieces together. *Bull. Amer. Meteor. Soc.*, **95**, 427–440, https://doi.org/10.1175/BAMS-D-12-00069.1.

Dorrington, J., I. Finney, T. Palmer, and A. Weisheimer, 2020: Beyond skill scores: Exploring sub-seasonal forecast value through a case-study of French month-ahead energy prediction. *Quart. J. Roy. Meteor. Soc.*, **146**, 3623–3637, https://doi.org/10.1002/qj.3863.

Duchez, A., and Coauthors, 2016: Drivers of exceptionally cold North Atlantic Ocean temperatures and their link to the 2015 European heat wave. *Environ. Res. Lett.*, **11**, 074004, https://doi.org/10.1088/1748-9326/11/7/074004.

Ferrone, A., D. Mastrangelo, and P. Malguzzi, 2017: Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale. *Adv. Sci. Res.*, **14**, 123–129, https://doi.org/10.5194/asr-14-123-2017.

Feudale, L., and J. Shukla, 2011: Influence of sea surface temperature on the European heat wave of 2003 summer. Part I: An observational study. *Climate Dyn.*, **36**, 1691–1703, https://doi.org/10.1007/s00382-010-0788-0.

Fischer, E. M., S. I. Seneviratne, P. L. Vidale, D. Lüthi, and C. Schär, 2007: Soil moisture–atmosphere interactions during the 2003 European summer heat wave. *J. Climate*, **20**, 5081–5099, https://doi.org/10.1175/JCLI4288.1.

——, U. Beyerle, and R. Knutti, 2013: Robust spatially aggregated projections of climate extremes. *Nat. Climate Change*, **3**, 1033–1038, https://doi.org/10.1038/nclimate2051.

Folland, C. K., J. Knight, H. W. Linderholm, D. Fereday, S. Ineson, and J. W. Hurrell, 2009: The summer North Atlantic oscillation: Past, present, and future. *J. Climate*, **22**, 1082–1103, https://doi.org/10.1175/2008JCLI2459.1.

Garcia-Serrano, J., and C. Frankignoul, 2014: Retraction note: High predictability of the winter Euro–Atlantic climate from cryospheric variability. *Nat. Geosci.*, **7**, E2, https://doi.org/10.1038/ngeo2164.

Gibson, P. B., W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser, 2021: Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun. Earth Environ.*, **2**, 159, https://doi.org/10.1038/s43247-021-00225-4.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Ser. B Stat. Methodol.*, **69**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Guigma, K. H., D. MacLeod, M. Todd, and Y. Wang, 2021: Prediction skill of Sahelian heatwaves out to subseasonal lead times and importance of atmospheric tropical modes of variability. *Climate Dyn.*, **57**, 537–556, https://doi.org/10.1007/s00382-021-05726-8.

Haarsma, R. J., F. Selten, B. V. Hurk, W. Hazeleger, and X. Wang, 2009: Drier Mediterranean soils due to greenhouse warming bring easterly winds over summertime Central Europe. *Geophys. Res. Lett.*, **36**, L04705, https://doi.org/10.1029/2008GL036617.

Hall, R. J., J. M. Jones, E. Hanna, A. A. Scaife, and R. Erdélyi, 2017: Drivers and potential predictability of summer time North Atlantic polar front jet variability. *Climate Dyn.*, **48**, 3869–3887, https://doi.org/10.1007/s00382-016-3307-0.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, https://doi.org/10.1256/qj.06.25.

He, S., X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, 2021: Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *Proc. Conf. AAAI Artif. Intell.*, **35**, 169–177.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hooker, G., and L. Mentch, 2019: Please stop permuting features: An explanation and alternatives. https://arxiv.org/abs/1905.03151.

Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: A stimulus for our science. *Quart. J. Roy. Meteor. Soc.*, **139**, 573–584, https://doi.org/10.1002/qj.1991.

Jergensen, G., 2019: Permutationimportance. GitHub, accessed 17 August 2020, https://github.com/gelijergensen/PermutationImportance.

Jézéquel, A., P. Yiou, and S. Radanovics, 2018: Role of circulation in European heatwaves using flow analogues. *Climate Dyn.*, **50**, 1145–1159, https://doi.org/10.1007/s00382-017-3667-0.

Kämäräinen, M., P. Uotila, A. Y. Karpechko, O. Hyvärinen, I. Lehtonen, and J. Räisänen, 2019: Statistical learning methods as a basis for skillful seasonal temperature forecasts in Europe. *J. Climate*, **32**, 5363–5379, https://doi.org/10.1175/JCLI-D-18-0765.1.

Kirkwood, C., T. Economou, H. Odbert, and N. Pugeault, 2021: A framework for probabilistic weather forecast post-processing across models and lead times using machine learning. *Philos. Trans. Roy. Soc. Math., Phys., Eng. Sci.*, **A379**, 20200099, https://doi.org/10.1098/rsta.2020.0099.

Kolstad, E. W., and M. Årthun, 2018: Seasonal prediction from Arctic Sea surface temperatures: Opportunities and pitfalls. *J. Climate*, **31**, 8197–8210, https://doi.org/10.1175/JCLI-D-18-0016.1.

Koster, R. D., and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, **37**, L02402, https://doi.org/10.1029/2009GL041677.

Kretschmer, M., J. Runge, and D. Coumou, 2017: Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophys. Res. Lett.*, **44**, 8592–8600, https://doi.org/10.1002/2017GL074696.

Kueh, M.-T., and C.-Y. Lin, 2020: The 2018 summer heatwaves over northwestern Europe and its extended-range prediction. *Sci. Rep.*, **10**, 19283, https://doi.org/10.1038/s41598-020-76181-4.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol.*, **32**, 1209–1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Lundberg, S. M., and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 2522–5839, https://doi.org/10.1038/s42256-019-0138-9.

Ma, S., A. J. Pitman, R. Lorenz, J. Kala, and J. Srbinovsky, 2016: Earlier green-up and spring warming amplification over Europe. *Geophys. Res. Lett.*, **43**, 2011–2018, https://doi.org/10.1002/2016GL068062.

Mariotti, A., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Mayer, K., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48**, e2020GL092092, https://doi.org/10.1029/2020GL092092.

McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

McInnes, L., J. Healy, and S. Astels, 2017: hdbscan: Hierarchical density based clustering. *J. Open Source Software*, **2**, 205, https://doi.org/10.21105/joss.00205.

McKinnon, K. A., A. Rhines, M. Tingley, and P. Huybers, 2016: Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nat. Geosci.*, **9**, 389–394, https://doi.org/10.1038/ngeo2687.

Mecikalski, J. R., T. N. Sandmæl, E. M. Murillo, C. R. Homeyer, K. M. Bedka, J. M. Apke, and C. P. Jewett, 2021: A random-forest model to assess predictor importance and nowcast severe storms using high-resolution radar–GOES satellite–lightning observations. *Mon. Wea. Rev.*, **149**, 1725–1746, https://doi.org/10.1175/MWR-D-19-0274.1.

Miralles, D. G., P. Gentine, S. I. Seneviratne, and A. J. Teuling, 2019: Land–atmospheric feedbacks during droughts and heatwaves: State of the science and current challenges. *Ann. N. Y. Acad. Sci.*, **1436**, 19–35, https://doi.org/10.1111/nyas.13912.

Molnar, C., G. Casalicchio, and B. Bischl, 2020: Interpretable machine learning—A brief history, state-of-the-art and challenges. *ECML PKDD 2020: Communications in Computer and Information Science*, Vol. 1323, Springer, 417–431, https://doi.org/10.1007/978-3-030-65965-3_28.

Muñoz-Sabater, J., and Coauthors, 2021: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data*, **13**, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021.

Nakanishi, T., Y. Tachibana, and Y. Ando, 2021: Possible semi-circumglobal teleconnection across Eurasia driven by deep convection over the Sahel. *Climate Dyn.*, **57**, 2287–2299, https://doi.org/10.1007/s00382-021-05804-x.

O'Reilly, C. H., T. Woollings, L. Zanna, and A. Weisheimer, 2018: The impact of tropical precipitation on summertime Euro-Atlantic circulation via a circumglobal wave train. *J. Climate*, **31**, 6481–6504, https://doi.org/10.1175/JCLI-D-17-0451.1.

Osborne, J. M., M. Collins, J. A. Screen, S. I. Thomson, and N. Dunstone, 2020: The North Atlantic as a driver of summer atmospheric circulation. *J. Climate*, **33**, 7335–7351, https://doi.org/10.1175/JCLI-D-19-0423.1.

Ossó, A., R. Sutton, L. Shaffrey, and B. Dong, 2020: Development, amplification and decay of Atlantic/European summer weather patterns linked to spring North Atlantic sea surface temperatures. *J. Climate*, **33**, 5939–5951, https://doi.org/10.1175/JCLI-D-19-0613.1.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Perkins, S. E., 2015: A review on the scientific understanding of heatwaves—Their measurement, driving mechanisms, and changes at the global scale. *Atmos. Res.*, **164**, 242–267, https://doi.org/10.1016/j.atmosres.2015.05.014.

Prodhomme, C., F. Doblas-Reyes, O. Bellprat, and E. Dutra, 2016: Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dyn.*, **47**, 919–935, https://doi.org/10.1007/s00382-015-2879-4.

——, S. Materia, C. Ardilouze, R. H. White, L. Batté, V. Guemas, G. Fragkoulidis, and J. García-Serrano, 2022: Seasonal prediction of European summer heatwaves. *Climate Dyn.*, **58**, 2149–2166, https://doi.org/10.1007/s00382-021-05828-3.

Quesada, B., R. Vautard, P. Yiou, M. Hirschi, and S. I. Seneviratne, 2012: Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nat. Climate Change*, **2**, 736–741, https://doi.org/10.1038/nclimate1536.

Quinting, J., and F. Vitart, 2019: Representation of synoptic-scale Rossby wave packets and blocking in the S2S prediction project database. *Geophys. Res. Lett.*, **46**, 1070–1078, https://doi.org/10.1029/2018GL081381.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Röthlisberger, M., L. Frossard, L. F. Bosart, D. Keyser, and O. Martius, 2019: Recurrent synoptic-scale Rossby wave patterns and their effect on the persistence of cold and hot spells. *J. Climate*, **32**, 3207–3226, https://doi.org/10.1175/JCLI-D-18-0664.1.

Runge, J., V. Petoukhov, and J. Kurths, 2014: Quantifying the strength and delay of climatic interactions: The ambiguities of cross correlation and a novel measure based on graphical models. *J. Climate*, **27**, 720–739, https://doi.org/10.1175/JCLI-D-13-00159.1.

——, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, 2019: Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.*, **5**, eaau4996, https://doi.org/10.1126/sciadv.aau4996.

Russo, S., and Coauthors, 2014: Magnitude of extreme heat waves in present climate and their projection in a warming world. *J. Geophys. Res. Atmos.*, **119**, 12–500, https://doi.org/10.1002/2014JD022098.

Saunders, K., A. Stephenson, and D. Karoly, 2021: A regionalisation approach for rainfall based on extremal dependence. *Extremes*, **24**, 215–240, https://doi.org/10.1007/s10687-020-00395-y.

Schaller, N., J. Sillmann, J. Anstey, E. Fischer, C. Grams, and S. Russo, 2018: Influence of blocking on Northern European and Western Russian heatwaves in large climate model ensembles. *Environ. Res. Lett.*, **13**, 054015, https://doi.org/10.1088/1748-9326/aaba55.

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schneidereit, A., S. Schubert, P. Vargin, F. Lunkeit, X. Zhu, D. H. Peters, and K. Fraedrich, 2012: Large-scale flow and the long-lasting blocking high over Russia: Summer 2010. *Mon. Wea. Rev.*, **140**, 2967–2981, https://doi.org/10.1175/MWR-D-11-00249.1.

Schubert, S., H. Wang, and M. Suarez, 2011: Warm season subseasonal variability and climate extremes in the Northern Hemisphere: The role of stationary Rossby waves. *J. Climate*, **24**, 4773–4792, https://doi.org/10.1175/JCLI-D-10-05035.1.

Schumacher, D. L., J. Keune, C. C. Van Heerwaarden, J. V.-G. de Arellano, A. J. Teuling, and D. G. Miralles, 2019: Amplification of mega-heatwaves through heat torrents fuelled by upwind drought. *Nat. Geosci.*, **12**, 712–717, https://doi.org/10.1038/s41561-019-0431-6.

Segal, M. R., 2004: Machine learning benchmarks and random forest regression. Tech. Rep., University of California, 1 pp., https://escholarship.org/uc/item/35x3v9t4.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Sci. Rev.*, **99**, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004.

Sillmann, J., and Coauthors, 2017: Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Wea. Climate Extremes*, **18**, 65–74, https://doi.org/10.1016/j.wace.2017.10.003.

Sousa, P. M., R. M. Trigo, D. Barriopedro, P. M. Soares, and J. A. Santos, 2018: European temperature responses to blocking and ridge regional patterns. *Climate Dyn.*, **50**, 457–477, https://doi.org/10.1007/s00382-017-3620-2.

Stéfanon, M., P. Drobinski, F. D'Andrea, and N. de Noblet-Ducoudré, 2012a: Effects of interactive vegetation phenology on the 2003 summer heat waves. *J. Geophys. Res.*, **117**, D24103, https://doi.org/10.1029/2012JD018187.

——, F. D'Andrea, and P. Drobinski, 2012b: Heatwave classification over Europe and the Mediterranean region. *Environ. Res. Lett.*, **7**, 014023, https://doi.org/10.1088/1748-9326/7/1/014023.

Suarez-Gutierrez, L., W. A. Müller, C. Li, and J. Marotzke, 2020: Dynamical and thermodynamical drivers of variability in European summer heat extremes. *Climate Dyn.*, **54**, 4351–4366, https://doi.org/10.1007/s00382-020-05233-2.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

Tilloy, A., B. Malamud, and A. Joly-Laugel, 2021: A methodology for the spatiotemporal identification of compound hazards: Wind and precipitation extremes in Great Britain (1979–2019). *Earth Syst. Dyn.*, https://doi.org/10.5194/esd-2021-52, in press.

van den Hurk, B., F. Doblas-Reyes, G. Balsamo, R. D. Koster, S. I. Seneviratne, and H. Camargo, 2012: Soil moisture effects on seasonal temperature and precipitation forecast scores in Europe. *Climate Dyn.*, **38**, 349–362, https://doi.org/10.1007/s00382-010-0956-2.

van Oldenborgh, G. J., F. D. Reyes, S. Drijfhout, and E. Hawkins, 2013: Reliability of regional climate model trends. *Environ. Res. Lett.*, **8**, 014055, https://doi.org/10.1088/1748-9326/8/1/014055.

van Straaten, C., K. Whan, and M. Schmeits, 2018: Statistical post-processing and multivariate structuring of high-resolution ensemble precipitation forecasts. *J. Hydrometeor.*, **19**, 1815–1833, https://doi.org/10.1175/JHM-D-18-0105.1.

——, ——, D. Coumou, B. van den Hurk, and M. Schmeits, 2020: The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Quart. J. Roy. Meteor. Soc.*, **146**, 2654–2670, https://doi.org/10.1002/qj.3810.

Vogel, M. M., J. Zscheischler, E. M. Fischer, and S. Seneviratne, 2020: Development of future heatwaves for different hazard

thresholds. *J. Geophys. Res. Atmos.*, **125**, e2019JD032070, https://doi.org/10.1029/2019JD032070.

Wehrli, K., B. P. Guillod, M. Hauser, M. Leclair, and S. I. Seneviratne, 2019: Identifying key driving processes of major recent heatwaves. *J. Geophys. Res. Atmos.*, **124**, 11 746–11 765, https://doi.org/10.1029/2019JD030635.

Wei, P., Z. Lu, and J. Song, 2015: Variable importance analysis: A comprehensive review. *Reliab. Eng. Syst. Saf.*, **142**, 399–432, https://doi.org/10.1016/j.ress.2015.05.018.

Weisheimer, A., F. J. Doblas-Reyes, T. Jung, and T. Palmer, 2011: On the predictability of the extreme summer 2003 over Europe. *Geophys. Res. Lett.*, **38**, L05704, https://doi.org/10.1029/2010GL046455.

Whan, K., and M. Schmeits, 2018: Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Mon. Wea. Rev.*, **146**, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1.

——, J. Zscheischler, R. Orth, M. Shongwe, M. Rahimi, E. O. Asare, and S. I. Seneviratne, 2015: Impact of soil moisture on extreme maximum temperatures in Europe. *Wea. Climate Extremes*, **9**, 57–67, https://doi.org/10.1016/j.wace.2015.05.001.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wirth, V., M. Riemer, E. K. Chang, and O. Martius, 2018: Rossby wave packets on the midlatitude waveguide—A review. *Mon. Wea. Rev.*, **146**, 1965–2001, https://doi.org/10.1175/MWR-D-16-0483.1.

Wolf, G., D. J. Brayshaw, N. P. Klingaman, and A. Czaja, 2018: Quasi-stationary waves and their impact on European weather and extreme events. *Quart. J. Roy. Meteor. Soc.*, **144**, 2431–2448, https://doi.org/10.1002/qj.3310.

——, A. Czaja, D. Brayshaw, and N. Klingaman, 2020: Connection between sea surface anomalies and atmospheric quasi-stationary waves. *J. Climate*, **33**, 201–212, https://doi.org/10.1175/JCLI-D-18-0751.1.

Wulff, C. O., and D. I. Domeisen, 2019: Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. *Geophys. Res. Lett.*, **46**, 11 520–11 529, https://doi.org/10.1029/2019GL084314.

Zampieri, M., F. D'Andrea, R. Vautard, P. Ciais, N. de Noblet-Ducoudré, and P. Yiou, 2009: Hot European summers and the role of soil moisture in the propagation of Mediterranean drought. *J. Climate*, **22**, 4747–4758, https://doi.org/10.1175/2009JCLI2568.1.

Zhang, R., C. Sun, J. Zhu, R. Zhang, and W. Li, 2020: Increased European heat waves in recent decades in response to shrinking Arctic sea ice and Eurasian snow cover. *npj Climate Atmos. Sci.*, **3**, 7, https://doi.org/10.1038/s41612-020-0110-8.

Zhang, Y., W. Huang, and D. Zhong, 2019: Major moisture pathways and their importance to rainy season precipitation over the Sanjiangyuan region of the Tibetan Plateau. *J. Climate*, **32**, 6837–6857, https://doi.org/10.1175/JCLI-D-19-0196.1.

Zschenderlein, P., A. H. Fink, S. Pfahl, and H. Wernli, 2019: Processes determining heat waves across different European climates. *Quart. J. Roy. Meteor. Soc.*, **145**, 2973–2989, https://doi.org/10.1002/qj.3599.