



# Predictability of the European heat and cold waves

Christophe Lavaysse<sup>1,2</sup> · Gustavo Naumann<sup>1</sup> · Lorenzo Alfieri<sup>1</sup> · Peter Salamon<sup>1</sup> · Jürgen Vogt<sup>1</sup>

Received: 11 January 2018 / Accepted: 20 May 2018 / Published online: 5 June 2018  
© The Author(s) 2018

## Abstract

Heat and cold waves may have considerable human and economic impacts in Europe. Recent events, like the heat waves observed in France in 2003 and Russia in 2010, illustrated the major consequences to be expected. Reliable Early Warning Systems for extreme temperatures would, therefore, be of high value for decision makers. However, they require a clear definition and robust forecasts of these events. This study analyzes the predictability of heat and cold waves over Europe, defined as at least three consecutive days of  $T_{\min}$  and  $T_{\max}$  above the quantile Q90 (under Q10), using the extended ensemble system of ECMWF. The results show significant predictability for events within a 2-week lead time, but with a strong decrease of the predictability during the first week of forecasts (from 80 to 40% of observed events correctly forecasted). The scores show a higher predictive skill for the cold waves (in winter) than for the heat waves (in summer). The uncertainties and the sensitivities of the predictability are discussed on the basis of tests conducted with different spatial and temporal resolutions. Results demonstrate the negligible effect of the temporal resolution (very few errors due to bad timing of the forecasts), and a better predictability of large-scale events. The onset and the end of the waves are slightly less predictable with an average of about 35% (30%) of observed heat (cold) waves onsets or ends correctly forecasted with a 5-day lead time. Finally, the forecasted intensities show a correlation of about 0.65 with those observed, revealing the challenge to predict this important characteristic.

**Keywords** Predictability · Heat waves · Cold waves · Ensemble forecasts · Validation

## 1 Introduction

Heat or cold waves (HWs or CWs hereafter) are linked to increased risks of mortality, morbidity, and different cardiovascular and respiratory diseases (Anderson and Bell 2009; Gasparrini and Armstrong 2011; Kovats and Hajat 2008; Huynen et al. 2001; Thakur et al. 1987). These effects are felt more severely among the vulnerable groups such as the elderly, children, and people with pre-existing chronic medical conditions (Anderson and Bell 2009; Braga et al. 2002; Hajat et al. 2007; Rocklöv et al. 2014). During the

last decades, several extreme events around the globe were particularly strong, resulting in important impacts on human health. In Europe, the event of 2003 dramatically demonstrated that HW could pose a significant public health threat. In France, more than 14,000 people died due to the direct and indirect impacts of persistent warm temperatures in 2003 (Fouillet et al. 2006). In Moscow during the summer 2010, a HW combined with widespread forest fires resulted in more than 11,000 excess deaths (Shaposhnikov et al. 2014). During the winter 2009/2010, several severe CWs occurred in Northern and Western Europe due to the persistence of the negative phase of the North Atlantic Oscillation (Cattiaux et al. 2010). The increasing frequency and severity of these events (and more particularly the HWs) during the recent decades and the likelihood for the occurrence of even more intense events demands skillful predictions of these high-impact events beyond the 10-day forecasts range (IPCC 2012). Recent events underscore the need for early warning systems for extreme temperature anomalies, bringing the need for such systems to the attention of many European governments. Public health institutes and weather services

---

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00382-018-4273-5>) contains supplementary material, which is available to authorized users.

---

✉ Christophe Lavaysse  
christophe.lavaysse@ird.fr

<sup>1</sup> European Commission, Joint Research Centre (JRC),  
21027 Ispra, VA, Italy

<sup>2</sup> University of Grenoble Alpes, CNRS, IRD, G-INP, IGE,  
38000 Grenoble, France

around Europe have been responsible for the design of these warning systems and most of the European countries have implemented tailored action plans in affected areas to minimize harm to those most at risk. These plans, which may include early alerts and advisories and a variety of emergency measures to mitigate the heat or cold dangers, are called heat warning systems (HWS) or cold warning system (CWS) (Lowe et al. 2011; Stanojevic et al. 2014; Åström et al. 2014; Matzarakis 2017; Gough et al. 2014; Ghosh et al. 2014; Chalabi et al. 2016; Hajat et al. 2013; Masato et al. 2015). However, these systems require robust and reliable meteorological forecasts that must be linked to a comprehensive assessment of their uncertainties to correct and better predict not only the hazard but also the likely impacts (PJM 2014).

On the sub seasonal timescale, low-frequency intra seasonal tropical phenomena such as the Madden–Julian oscillation or slowly varying boundary conditions, e.g. SST anomalies, are commonly regarded as the primary potential sources of predictability. The mid-latitude circulation, however, is dominated by unpredictable weather noise and its evolution is strongly chaotic making it difficult to predict beyond 10 days. However, this limit of prediction skill does not exclude the possibility that some circulation states can be substantially more predictable than the average, either from having large amplitude or being associated with patterns that are intrinsically of low frequency (Goddard et al. 2001; Naumann and Vargas 2012). Some case studies have suggested that HWs (and associated droughts) may be preceded and accompanied by quasi-stationary large-scale mid-latitude atmospheric Rossby waves. Osman and Alvarez (2017) illustrates the potential benefit and predictability of heat waves by analysing the scores of two models for a specific case study in Southern America.

HW or CW and health warning systems are preparedness plans designed to reduce extreme temperature related adverse health effects, and can help raise awareness among vulnerable populations. Recently, Henderson and Kosatsky (2012) defined heat health emergencies as the coincidence of extreme temperature and extreme mortality days in the Vancouver metropolitan area, Canada. In a second step the predictive ability of heat alerts based on forecast data was assessed showing that at least 19 hours of lead time for preparedness was provided. Kosatsky et al. (2012) found that the accuracy of heat alerts predicted by forecasted temperatures varied with lead time and geographical area when compared with those predicted by observed temperatures. We know that forecast quality varies for different weather parameters. For example, temperature usually has a more accurate forecast than dewpoint temperature (Werth and Garrett 2011). Previous studies (Åström et al. 2014; Zhang et al. 2014) showed discrepancies between estimated risk increases based on direct forecast model output and actual

observed temperatures. This raises the question whether operational HW/CW early warning systems can adjust for this forecast bias. The practice of providing daily predictions of the forecasts is clearly one of a key point.

With a robust estimation of the forecast bias and capacities the HW/CW early warning systems could provide specific indicators with potential bias correction/adjustment to provide accurate warnings sufficiently far in advance.

This study aims at assessing the predictability of heat and cold waves in Europe by using robust statistics during the hindcasts of one of the most robust extended ensemble systems, developed at ECMWF. This study is presented in four sections, after the introduction, the data and methods used in this study are introduced in Sect. 2. Section 3 presents the results and discusses the spatial and temporal sensitivity of the forecasts. Finally Sect. 4 concludes this study.

## 2 Data and methods

### 2.1 Datasets

In this work the ECMWF extended range ensemble forecasting system (ENS, Vitart 2014) is used once a week to forecast HWs and CWs. This model is the up-to-date version of the ECMWF Integrated Forecasting System already tested (Vitart 2005). For lead times up to 10 days, the model is not coupled to the ocean and has a resolution of 32 km (T639). It is forced by persistent sea-surface temperature anomalies. Beyond a lead time of 10 days the resolution of the model becomes coarser (T319, 64 km) and is coupled to an ocean model. In this study, we focus on large-scale HWs and CWs, as these are responsible for the largest share of the impacts, so we upscale all the forecast products to one square degree resolution. The vertical resolution remains unchanged during the entire simulation at 62 vertical levels. Associated with each extended forecast, ECMWF provides hindcasts (or reforecast) for back statistic which is a 11-member ensemble starting on the same day and month of each Thursday's real-time forecast for each of the past 20 years (Vitart 2014). These hindcasts of the last 20 years of each forecast from June 2015 to June 2016 are used to assess the predictability of rare events. Therefore, the period of study spans from June 1995 to June 2015. The variables extracted from the model are the minimum and maximum temperature recorded in the model within 24-h (0000 UTC to 0000 UTC day + 1) windows ( $T_{\min}$  and  $T_{\max}$ ).

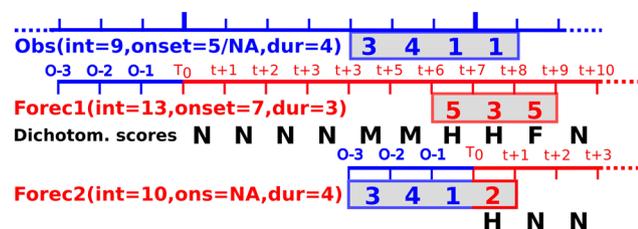
To validate the forecasts, the observed temperatures ( $T_{\min}$  and  $T_{\max}$ ) provided by the E-OBS/ECAD dataset Version 14 (Haylock et al. 2008; Van den Besselaar et al. 2011) are used. This dataset is based on station data and includes daily

gridded data of  $T_{min}$  and  $T_{max}$  over Europe with a spatial resolution of 0.25 square degrees. The original resolution was upscaled for coherence to the same one square degree model resolution.

### 2.2 HWs and CWs metrics

Following a recent study by Lavaysse et al. (2018), the definition of HWs and CWs proposed by Russo et al. (2015) was adapted to provide operational information for monitoring these extreme temperature events. To focus more on human impacts, both  $T_{min}$  and  $T_{max}$  are used as recommended by the WMO (McGregor et al. 2015). The definition of the events is done in 4 steps:

- Transformation of the daily  $T_{min}$  and  $T_{max}$  into daily quantiles.
- A specific day is defined as hot (cold) when the quantiles of both  $T_{min}$  and  $T_{max}$  are above Q90 (lower than Q10).
- A heat (cold) wave is then defined when at least three consecutive hot (cold) days are recorded in summer (winter).
- Two different methods were proposed by Lavaysse et al. (2018), see their Fig. 1 to assess the intensities of the HWs and CWs. The first method is the anomaly relative to calendar temperature thresholds, defined as the temperature equal to Q90 (or Q10, noted Idev hereafter, for an Intensity related to the temperature deviation). The second is the anomaly related to a yearly constant climatological threshold (noted Ihum hereafter, for an Intensity potentially related to the human impacts). This method aims at highlighting HWs or CWs that occur during the heart of the summer or winter season and which have potentially more impacts on human health.



**Fig. 1** Schematic illustration of the validation methods of HW/CW. The observed (illustrated with the blue box and the blue daily intensities) and the forecasted (red) events are first compared by using daily contingency results (in black, N—correct negative, M—misses, F—false alarms, H—hits). These results are then used to calculate POD/FAR/GSS. To assess the accuracy of the forecasted intensities, sum of daily intensities, the onset and the durations are recorded and compared (respectively int, onset and duration) if at least one hit is recorded

The forecasts of HWs and CWs are computed following the same procedure as for the monitoring with few adjustments. Indeed, to fit the requirements of users and decision makers, a dichotomous solution derived from the probabilistic forecasts is used. Following (Lavaysse et al. 2015) on the prediction of droughts, 10 different methods are tested (see Table 1). Some methods are integrative or based on specific individual members (i.e. ensemble mean, median member, Q25, Q75). These methods have their own thresholds (based on the quantiles Q10 and Q90). This allows to obtain the same number of occurrences of hot/cold days observed and forecasted (but not necessarily the HWs and CWs). Moreover, in order to correct the drift of the forecasts, these thresholds are also dependent on the lead time. A second type of method is based on different percentages (from 20 to 70%) of the ensemble members that forecast values over the threshold (e.g., the event is forecasted if the percentage of members forecasting an extreme event is larger than this percentage threshold). For this type of methods, no correction is applied, so the number of Heat/Cold days are dependent on the percentages and the lead time. Due to the increase of the ensemble spread with lead time, the frequency of occurrence of forecasted events decreases drastically after a lead time of approximately 7 days for each percentage used. All these methods are then compared to identify the best method for each lead time.

### 2.3 Validation tools

The ability of the model to correctly forecast extreme events is assessed using scores for probabilistic and dichotomous predictions. The reliability diagrams and the Relative Operating Characteristic (ROC) curves are exploited with probabilistic forecasts. The reliability diagram plots the observed frequency against the forecast probability, where the range of forecast probabilities is divided into K bins (from 0 to 100%) and allows to assess how well do the predicted probabilities

**Table 1** List of the 10 methods used to provide a Boolean index for HW/CW forecasting using an ensemble system

Name	Definition
Med	Temp. of the ensemble median of the members
Mean	Temp. of the ensemble mean of the members
Q25	Temp. of the member located at the 25% of the CDF
Q75	Temp. of the member located at the 75% of the CDF
20pc	20% of members above the temp. thresholds
30pc	30% of members above the temp. thresholds
40pc	40% of members above the temp. thresholds
50pc	50% of members above the temp. thresholds
60pc	60% of members above the temp. thresholds
70pc	70% of members above the temp. thresholds

of an event correspond to their observed frequencies. The perfect reliability is thus indicated by the proximity of the plotted curve to the diagonal and the deviation from the diagonal gives the conditional bias. The ROC score (Jolliffe and Stephenson 2012), which is conditioned on the observations, is complementary to the reliability diagram. This score is based on the plot of hit rate vs. false alarm rate, using a set of increasing probability thresholds to make the yes/no decision and allows to assess the ability of the forecast to discriminate between events and non-events. To synthesize on the scores depending on the lead time, the ROC area (the surface under the ROC curve) is calculated. Therefore, one value is produced per lead time and is bounded within 0 and 1 where 0.5 indicates no skill and 1 the perfect score.

For the deterministic solutions, the dichotomous scores use the four types of agreement between observed and forecasted fields that are hits, false alarms, misses and correct negatives. The probability of detection (POD) indicates the percentage of observed HWs or CWs that have been correctly forecasted. The false alarm rate (FAR) indicates the percentage of forecasted events that did not occur. Finally, the Gilbert Skill Score (GSS, Jolliffe and Stephenson 2003) takes into account the hits, misses and false alarms and neglects the correct negative forecasts that would artificially improve the score. It indicates how well the forecasted HWs/CWs correspond to the observations. This score is adjusted to account for the hits obtained by chance (i.e. closely related to the climatology). It is calculated as follows:

$$GSS = \frac{(hits - hits_c)}{(hits + misses + false\ alarms - hits_c)} \quad (1)$$

where  $hits_c = \frac{(hits+misses)(hits+false\ alarms)}{total}$ . Based on these equations, a perfect score will be 1, and a score equal to 0 indicates the score of the climatology (no skill).

## 2.4 Validation procedures

The choice of the methodology to assess the quality of HW/CW forecasts is also a key point. Most of previous studies focused on the prediction of extreme temperature anomalies, while no previous study estimated the predictability of HW/CW forecasts. In this study, two types of procedure are adopted. The first one is based on historical observations and forecasts using a recomposed time series of datasets. The purpose is to compare the occurrence frequencies and the intensities of the forecasts versus the observations at climatological level.

The second methodology focuses on the predictability aspects. This is done with two approaches, as illustrated in Fig. 1. The first approach (illustrated with the dichotomous scores in black in Fig. 1) is based on daily scores for each lead time. To do so, the false alarms, misses, correct

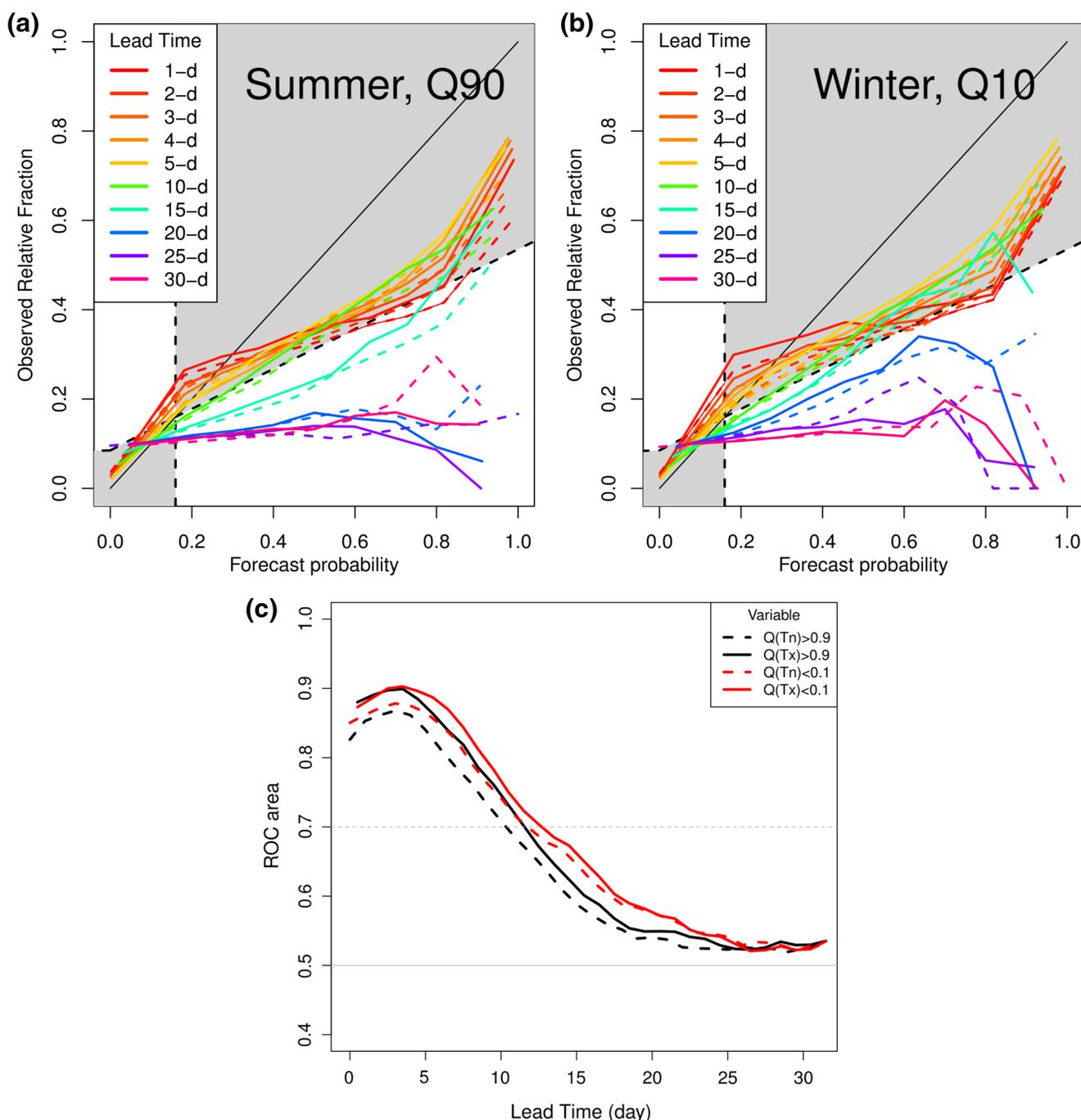
negative and hits are calculated on a daily basis. This method allows to assess the skill of the model to predict a day as part of a heat or cold wave at a certain lead time. The second approach focuses on entire events (entire HW or CW) rather than daily values. To do so, the forecast of a specific event is considered, if one of the days during an observed event is also forecasted (e.g., at least one hit day during the period of an observed event). In such case, three indicators are derived and compared from the observed and forecasted events; (1) the sum of the daily intensities, (2) the onset and (3) the duration. If there is no overlap between the forecasted and the observed events, the event is considered as missed or false alarm. For this approach, the time series is reconstructed by concatenating the three previous days of observation to the forecasts. This is done to take into account the initial conditions (known at the moment of initialising the forecasts) and to avoid bias in the HW/CW detection due to a bounded time series. Once the HW/CWs are detected, the scores are calculated from the first day of the forecasts (e.g., only the forecasted values are used to avoid dumping of the scores).

## 3 Results

### 3.1 Climatological assessment

Over Europe, the ECMWF ENS model provides skillful forecasts of temperatures with a low bias (not shown, Haiden et al. 2014) and a correlation, using concatenation of 2-week lead time forecasts, of the temperature quantiles of both  $T_{min}$  and  $T_{max}$  above 0.65 (up to 0.78 over the south-eastern Europe, see Fig. S1 in supplementary material). Note that the use of the quantiles of temperature removes the seasonal variability and thus focuses on the high frequency variabilities that are more difficult to predict. Only over Spain, Greece and the UK, the correlations are recorded under 0.6. The temporal variabilities at yearly time scales (Fig. S2a in supplementary material) depicts low differences during the last 20 years. This is mainly due to the use of the hindcasts that are based on the same version of the model during the entire period. Some higher skill is found in 1998, 2002 and 2010 and lower correlation in 1997 and 2009. The explanations of these differences could be related to more intense large scale forcing or more stationary conditions during these years that are more predictable than high-frequency variabilities. The same analysis (see Fig. S2b in supplementary materials) is done at monthly scale to show the variation of correlation depending on the month.

To analyze the predictability of the ensemble for the relative extreme temperatures (defined as above or under the quantiles Q90 or Q10), the reliability diagrams based on different lead times are presented for summer and winter



**Fig. 2** (Top) reliability diagrams for the quantiles of temperature above 0.9 in summer (a) or under 0.1 in winter (b). Color lines indicate the lead time (dashed lines using  $T_{min}$ , continuous lines using

$T_{max}$ ). c ROC area for quantile of temperature above Q90 in summer (black) or lower than Q10 in winter (red) according to the lead time for  $T_{min}$  (dashed lines) and  $T_{max}$  (continuous lines)

in Fig. 2a. Generally, the ensemble model is overconfident and tends to converge too quickly toward the same solution (curves located under the diagonal line on the right part of the graph). By adding the skillful region, between the skill lines (in grey in Fig. 2a, b, Toth et al. 2003; Wilks 2011) it is possible to analyse the reliability of the forecasts at each

lead time. In summer, the reliability diagrams are skillful up to 10-day lead time and up to 15-day lead time for the winter Q10. Tests are not designed to distinguish two lead times. Nevertheless, we can consider there are two groups, one with skillful reliability and the second one that converges to the climatology. The figure allows to highlight a certain

limitation of the predictability to forecast a single hot day, though the HWs/CWs, associated with persistency and large scale features, could be more predictable. It is worth to note that the system is slightly underconfident for low rates of probability forecasts. The ROC area scores for the same variables (Fig. 2c) are in agreement with the reliability diagram showing an abrupt decrease of the score beyond 15-day lead time. Two main characteristics appear: (1) there is a diurnal cycle of the score due to the better forecast of  $T_{\max}$  in relation to  $T_{\min}$ ; (2) the skill score of the ROC area for cold waves is better than the one for the heat waves, and appears skillful up to two weeks according to (Buizza et al. 1999).

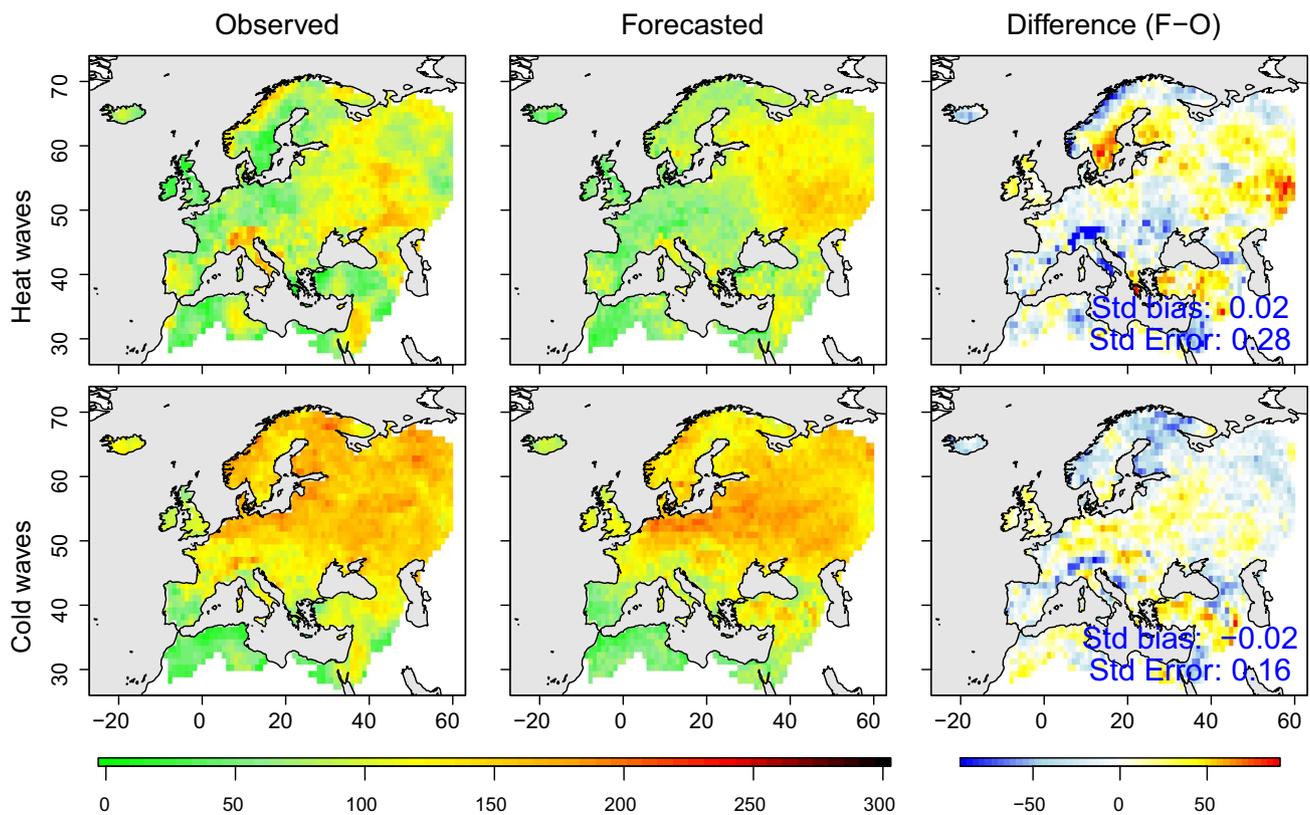
For the next validation, a re-composition of a continuous time series of 20 years of hindcasts is built by concatenating the 15-day lead time forecasts of  $T_{\min}$  and  $T_{\max}$  from runs separated by two weeks. The choice of using the first two weeks is done following the skill scores of the predictability of temperature quantiles above Q90 (or lower than Q10) shown previously. This time window is long enough to ensure a consistent detection of HWs/CWs by the forecast system and decrease the influences of abrupt changes or drifts of the model. Sensitivity tests are conducted to assess the influence of this strategy of climatological reconstruction. To do so, two additional time series are built by merging every week the first week of forecast (e.g., from 1-day to 7-day lead time), or the second week of forecast (e.g. from 8-day to 14-day lead time). Results, shown in the supplementary materials (Figs. S3 and S4), show a decrease of HW/CW detection during the second week using identical methodologies (ensemble mean, ensemble median, same percentage). This is mainly explained by the increase in the ensemble spread. Nevertheless, by using an adapted methodology depending on the lead time (e.g., changing the percentage threshold of ensemble members associated with extreme events) the results allow to conclude that the model does not present important drift during the two first weeks. The occurrences are comparable and the spatial variability is well represented. Because of this stability and since the focus of this study is on long lead times, we use the entire 15-day lead time forecast to build the climatology. The main objective here is to compare spatial and temporal climatologies and variabilities from the observations and the model. Indeed, it is important to evaluate the quality of the climatology used as reference, before assessing the predictability of events that will be done in the following subsection.

To detect HWs/CWs in the ensemble, ten methods are tested to transform the probabilistic onto a deterministic forecast. These methods are defined in Table 1. For each method, the climatological occurrences are compared to those obtained with the observations. The optimum forecasts (i.e. associated with the lowest bias and standard error) for the climatology are depending on the characteristics of the HW/CW analyzed. For example, to predict the occurrences,

the optimum forecast is when 50% of members for the HWs (40% for the CWs) are associated with an extreme event. Such bias could be optimized by choosing the best percentage threshold. Once the best methods are defined, it is interesting to note the good representation of the spatial distributions with a clear gradient pointing to north-eastern Europe for both HWs and CWs (Fig. 3). The main differences are located over the Alps for HWs (negative bias) and Poland for CWs (positive bias). The same tests were performed to define the optimum forecasts of HW/CW intensities, resulting in the median and the Q25 for HWs and CWs respectively. To assess the forecasts, the strongest intensities recorded over each grid point were compared between the observations and the forecasts for both Idev and Ihum (Fig. 4). The correlations are better for Idev than Ihum and for HWs than CWs. These scores allow to show the proportion of spatial variability that is reasonably well represented, yet highlighting also the specific problem of assessing the wave intensities that need to be evaluated separately.

The inter-annual variability during the entire period is then analysed to assess the temporal stability of the forecast skill (Fig. 5). Note that, since 1995 and 2015 are not complete, these two years are removed from this analysis. The yearly mean occurrences of HWs/CWs (Fig. 5, left panels) reveal the low negative (positive) bias of the forecasted HWs (CWs). The inter annual variability of HWs is well represented except in 1998 and 2007 where the events are underestimated and in 2014 with an overestimation by the forecasts. In this figure, the wide Russian event in 2010 is clearly visible with the increase of HW occurrence. Nevertheless, in 2003, there is no signal related to the extreme event in France and Western Europe. This is mainly due to the not exceptional spread of this HW regarding the climatology. This event was defined as extreme because of its intensity (related to the deviation of the temperature anomalies) and the number of people exposed. The CW evolution is also well represented, despite a small overestimation at the end of the period that creates an underestimation of the linear trends toward a decrease of CWs with time. The Pearson's correlations of the temporal variability of the observations and the forecasts are significant and equal to 0.82 (0.88) for the HWs (CWs) with a 90% confidence interval of [0.72;0.97].

The long-term trend is assessed using the yearly values, Loess and linear regressions. The linear trend provides a global overview with an increase (decrease) of the HW (CW) occurrences in both observations and forecasts. Using the Mann-Kendall trend and the Sens slope tests, the positive trends of the HWs in summer are significant with a confidence level of 90% (Fig. 5, top left). The tests reveal no significant trends for the CWs (bottom left). The Loess regression, which allows to observe with more details the low frequency variations, highlights an increase of CWs



**Fig. 3** Occurrence of heat (top panels) and cold (bottom panels) waves from E-OBS (left), and from the optimum forecasts (up to 2-week lead time, middle, see text for more details) and the difference (right)

from 2010 to 2013 that could limit the significance of the trends.

The last climatological verification done is on the model assessment to represent the interannual variability of the occurrence frequencies at monthly time scale. This analysis allows to consider the reliability of the model to represent occurrence anomalies, and to check if there is no bias for each month along the hindcast period. Despite some errors for specific months, there is no systematic bias for the monthly occurrences, and the model is able to represent correctly the periods with both small or large occurrence frequencies (Fig. 5, right panels). This is true for both the HWs and CWs with Pearson's correlation equal to 0.79. Only September is associated with a non-significant temporal correlation mainly due to the small interannual variability (black dots in Fig. 5 top right).

### 3.2 Predictability of HWs and CWs

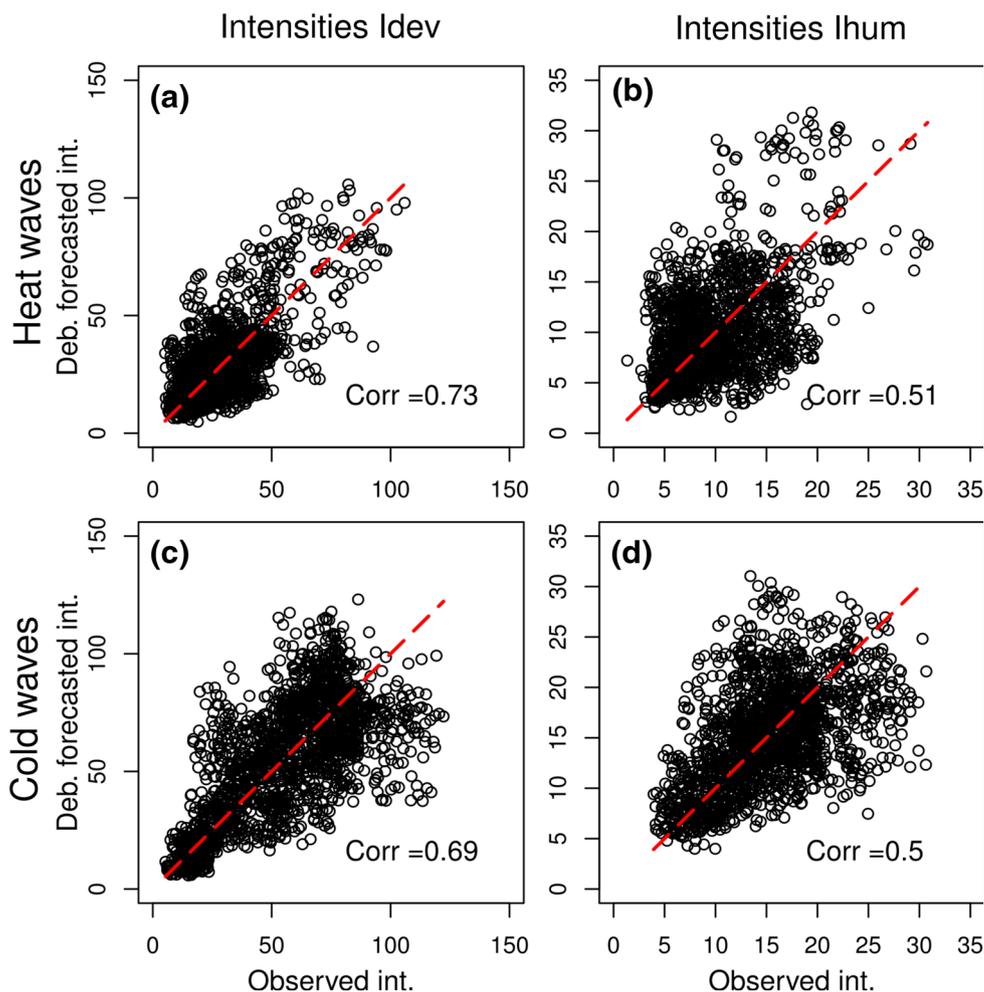
#### 3.2.1 Temporal and spatial distributions

To assess the predictability of the forecasted HWs and CWs, 10 methods to transform the probabilistic forecasts into dichotomous solutions are tested (see Table 1) to find the

most reliable prediction. Figure 6 displays the POD, FAR and GSS scores depending on the lead time and method. It shows that HWs have generally shorter predictability than the CWs. According to the positive values of GSS, the forecasts are skillful up to 15 days for the HWs and up to 21 days for the CWs. This affirmation is validated using two significance tests, one based on the standard error of the threat scores and the second based on the single population proportion test. Both indicate skillful positive values of GSS with a confidence interval up to 0.9. Nevertheless, the two scores display a drastic decrease during the first week of forecast. The importance of the method is also clearly visible with differences that reach 30% of the POD and FAR and 25% of the GSS depending on the method used. Based on these results, it is also possible to define the best method depending on the lead times. Considering only the first 10-day lead time, the best method to forecast the detection of HWs (CWs) is based on the 50% of the members (40% respectively). For longer lead time, missed events increase drastically due to the increase in the ensemble spread.

The spatial variability of the GSS versus the lead time is assessed in Fig. 7 using the most accurate method for each lead time (indicated at the top of each panel). From 1 to 5-day lead time, the spatial structures and the intensities are

**Fig. 4** Scatterplots of intensities Idev and Ihum forecasted and observed in the 20 years of reconstructed climatology. The forecasted intensities are calculated according to the best forecasts (see text for more details)

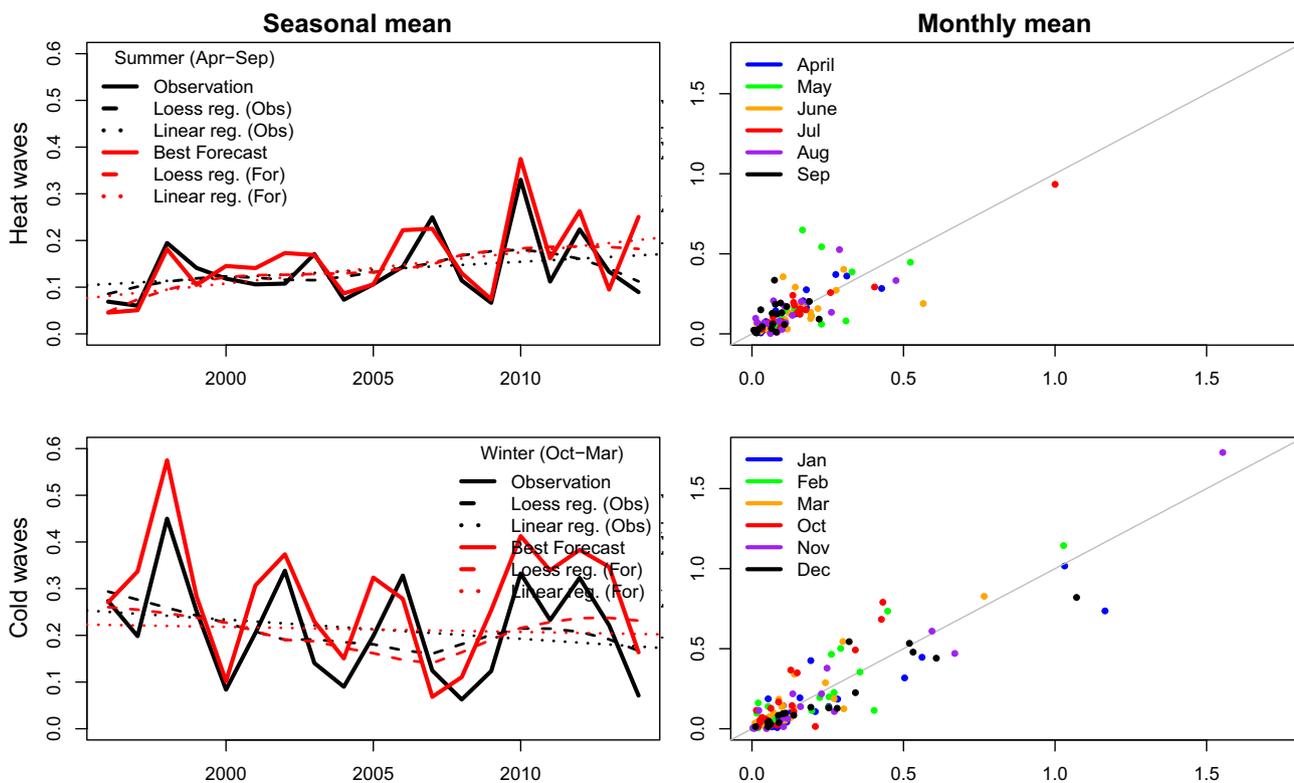


close, with an increase of the score over eastern Europe and over Russia for both HWs and CWs and over a band from northern France to northern Poland for CWs. After 7-day lead time, the GSS scores halve over almost the entire European continent. The maxima of predictability are located in the eastern part of the domain (with predominant temperate continental climates) and along the coast of the North and Baltic Sea. These positive values of GSS are also significant with a confidence interval of 90% by using the test for comparing two proportions. At 2-week lead time, skillful scores (i.e., positive GSS) are found only over northern France and Russia for the HWs and north and central Europe (except Scandinavia) for the CWs. At 3-week lead time, there is no better predictability than the climatology of HWs. For CWs, low and no significant scores are found over central Europe. Finally, regarding the method used to forecast these events, it is possible to notice a decrease in the percentage of members associated with the best forecasts, from 60% (70%) to 40% (40%) for the HWs (CWs respectively).

### 3.2.2 Sensitivity of the temporal and spatial scales

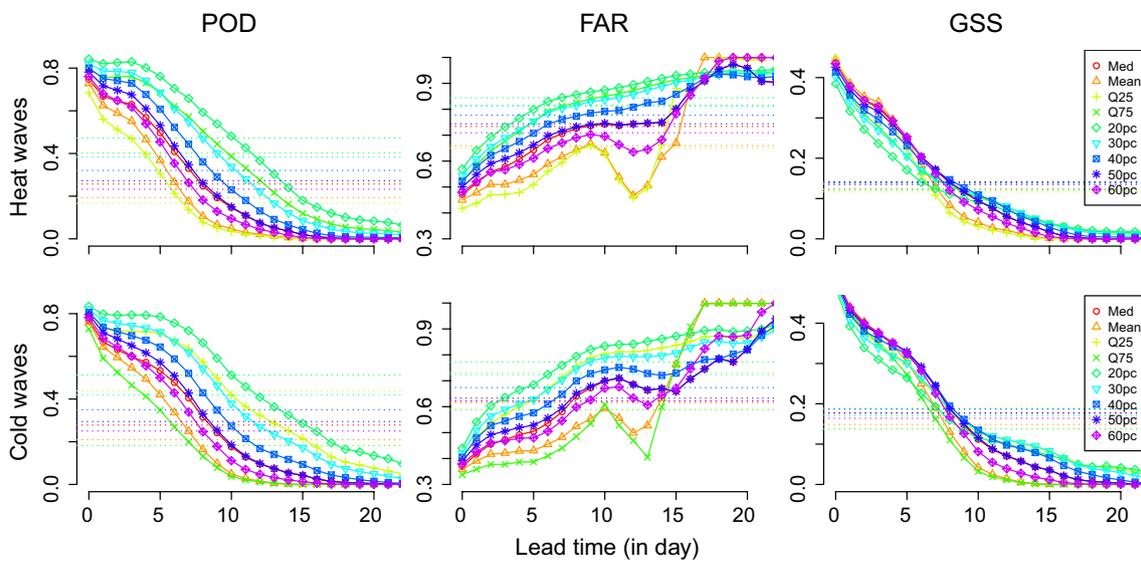
Two sensitivity tests are conducted to measure the influence of the temporal and spatial scales of the HWs/CWs. The temporal sensitivity test is conducted by testing different daily resolutions, from 1 to 6 days. The period is defined (in the observation and forecasts) as affected by HWs/CWs if at least one day within the window is affected. This modifies the time series and decreases the number of time steps (from 32 time steps for the 1-day resolution to 7 time steps for the 6-day resolution). Once the observations and the hind-cast signals are transformed, the same scores are applied. Results for the entire period and using the best forecasts are provided in Fig. S5 in the supplementary material. According to these results, the influence of the time resolution is negligible, demonstrating that most of the forecasted errors are not related to a shift in time of the predicted event but rather to misses or false alarms.

A second test is then conducted to evaluate the influence of the spatial scale of the HWs/CWs on their predictability. The original resolution of the forecasts and the observations

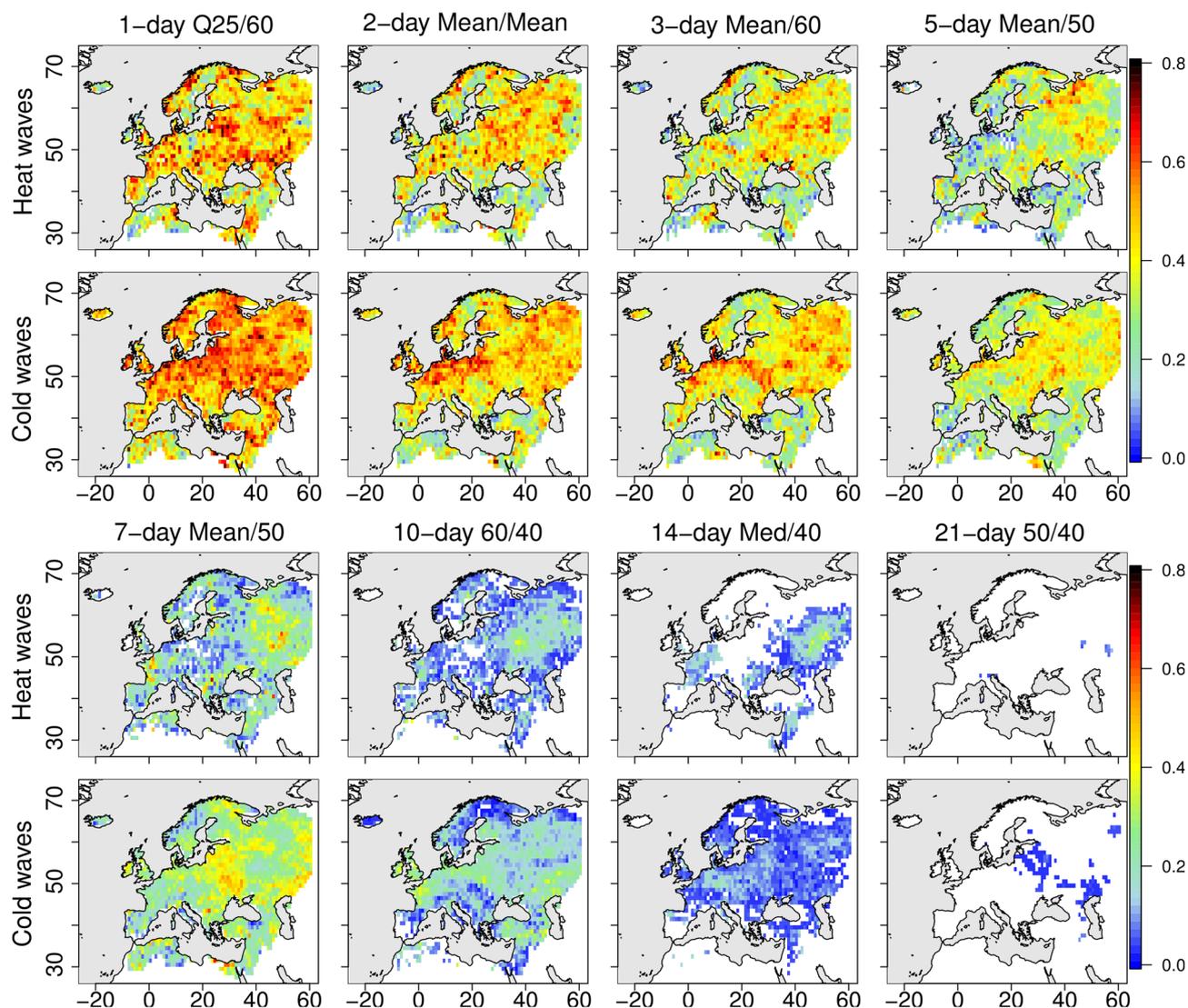


**Fig. 5** (Left panels) inter-annual variabilities of the seasonal mean number of days associated with heat waves (top) and cold waves (bottom) using EOBS and best forecasts (see text for more details). Loess and linear regressions are added in the left panels. (Right panels)

scatterplots of the monthly mean number of days for each month for heat (top) and cold (bottom) waves. The x-axis represent the observed events and the y-axis the forecasted ones



**Fig. 6** POD (left), FAR (center) and GSS (right panel) scores of the forecasts over Europe of heat (top) and cold (bottom) waves in summer and winter respectively, following the lead time (x-axis) and the method used (color lines)



**Fig. 7** Spatial variability of the GSS calculated at different lead-time (1, 2, 3, 5, 7, 10, 14 and 21-day lead time). The methods to forecasts these events are defined following the best method according to

the lead time and are all related to the percentage of members (see Fig. 5). These values are indicated in the title of each column for HWs/CWs

is 1 square degree. Our algorithm to define events requires only a minimum of 3-day duration. That means the possibility to detect isolated, small scale events (i.e. one isolated grid point), which could be more complex to forecast than large scale events. To analyze the spatial sensitivity, four methods, keeping the same resolution but smoothing the signals, are tested in order to distinguish small and large scale events. The boolean files (with values equal to 0 for normal conditions and 1 when HWs/CWs are detected) of each forecast are used as input files to these methods. The two first methods are using the surrounding grid cells with different matrix sizes ( $3 \times 3$  and  $5 \times 5$  respectively). The central grid cell of the moving matrix is equal to the matrix mean. Thus, the values indicate the percentage of the matrix

affected by the HWs/CWs. A threshold (65%) considers if the central grid cell is affected or not. The choice means that  $2/3$  of the domain is affected and allows to detect large scale events by keeping a robust number of events. Note that this method allows also detection on the coastal regions since the undefined values are not taken into account. With these two methods the weight of each grid cell is equal to 1. The third method is a Gaussian smoothing applied with a 2-D convolution operator defined in a  $5 \times 5$  matrix. This method is similar to the mean filter, but it uses different weightings that represents the shape of a Gaussian hump. This method better represents the large scale features of the waves but may be too strict over coastal regions since there is no compensation for undefined values. Finally, the last method is based

on a Nagao–Matsuyama filter (Nagao 2000) based on the lower spatial spread of a moving window. Nevertheless, this method, more adapted for smoothing satellite images, creates odd results (with HWs/CWs detected in wrong places) and tends to minimize the number of cases. For these reasons, this method was discarded. The results of the three first methods (see Fig. S6 in the supplementary materials) reveal a small but significant improvement of the forecasts when only the widest events are studied. The best results are found by using the Gaussian filter that tends to reduce the number of cases over the coastal regions. Nevertheless, for lead times longer than 10 days, there is no difference in the forecast skill depending on the spatial resolution.

As a conclusion of these sensitivity tests, the skill scores of the forecasts of HWs/CWs are not sensitive to the temporal scales. Regarding the spatial resolution, the skill scores appear significantly better when the HWs/CWs possess large scale features. Nevertheless, this is only true for lead times smaller than 7 days. Therefore, the forecast errors are mainly due to misses or false alarms associated with forecast errors of the large scale patterns and not due to uncertainties related to a spatial shift or a temporal delay.

### 3.2.3 Onset and end of HWs and CWs

The section aims at evaluating the skill scores of the forecasts on the onset and end of the HWs and CWs. Indeed, due to the link between the HWs/CWs and blocking situations (Matsueda 2011; Trigo et al. 2005), the predictability may be explained by the correct forecast of duration of an existing event. Also, for human health and economic impacts, users and decision makers need to know these specific onset and end dates to trigger the proper mitigation measures. To assess these scores, the dates of onsets and ends are derived from the boolean files of events. The starting (ending) date is defined as the first day with (without) a HW/CW detected after a non-event (event respectively). Thanks to the concatenation of the three previous daily observation with the forecasts mentioned earlier, the starting and ending dates can be detected from the first day of forecast.

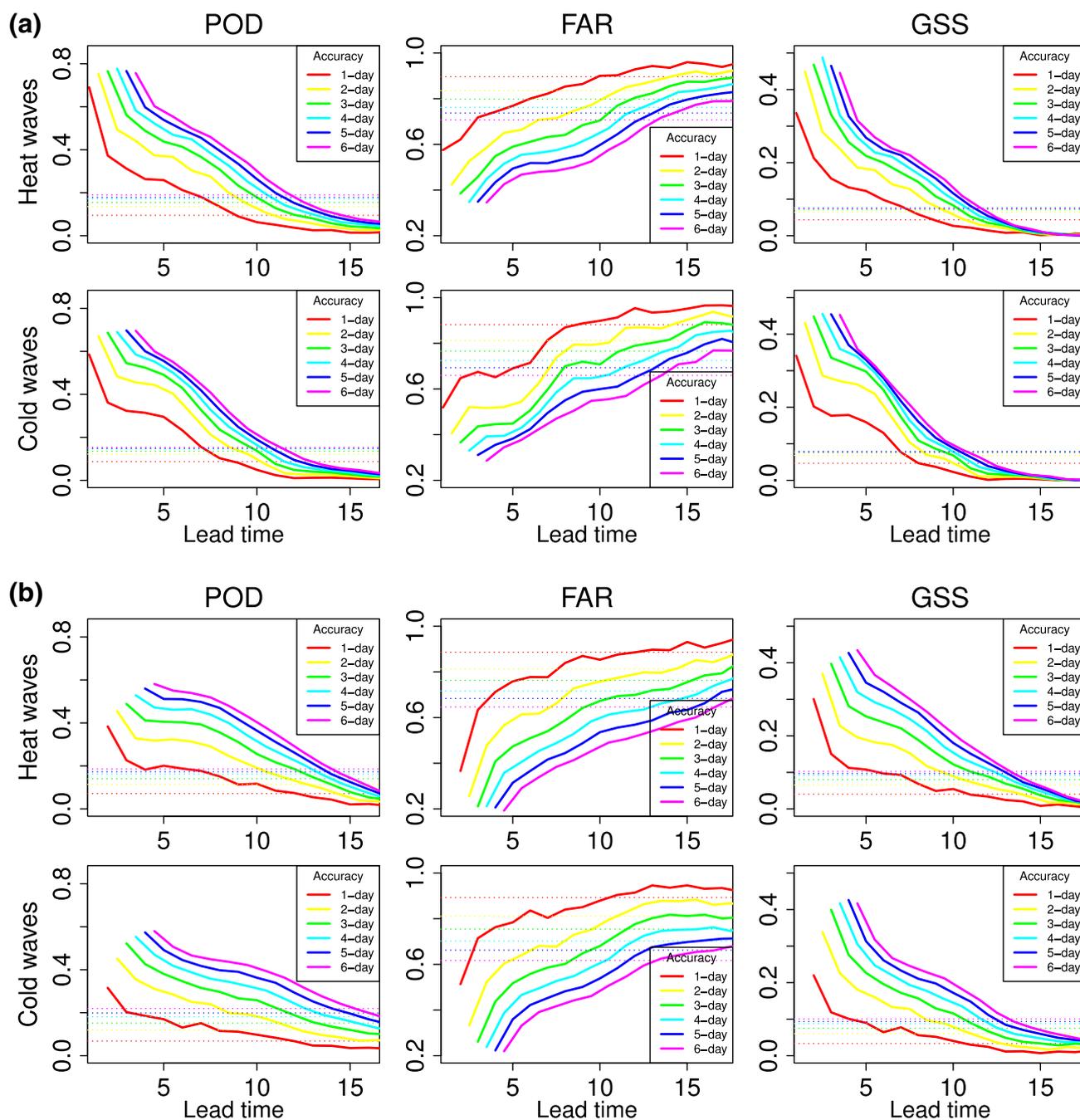
The best method to define forecasted onsets/ends from the ensemble are based on the GSS scores of the first couple of weeks (not shown) and are defined as follows: 20% (30%) of members for the HW (CW respectively) to detect the onset and 40% (30%) for the HW (CW respectively) to forecast the end of the event. According to these methods, the POD of the onsets is set to 70% for 1-day lead time and decreases abruptly to 25% after 5-day lead time. The time resolution (different line colors in Fig. 8) has a significant impact up to a 3-day window (POD larger than 80%). After that the improvements are not significant. The conclusion is similar for the FAR, while the GSS outlines these behaviors with large improvements in the accuracy up to a 4-day window

resolution. Finally, there is no significant advantage in using forecasts for the onset of HWs/CWs beyond a 2-week lead time, even at coarse temporal resolutions. For the end of the HWs/CWs (Fig. 8b), the scores are slightly better (except for the 1-day resolution) than for the onsets for the same lead time. This is due to the ability of the model to predict relatively well the correct durations of the HWs/CWs. The forecast skill for the end of HWs and CWs could go up to an 18-day lead time for HWs and a 21-day lead time for CWs.

### 3.2.4 Forecast skills of HW and CW intensities

The intensity of the HWs/CWs is one of the most important characteristic of the events to predict since it is closely related to the impacts on human health, especially for the intensity calculation  $I_{hum}$ . To assess the forecast ability, an adapted method of predictability assessment is needed and described in Fig. 1. First, the ability to correctly predict the events is verified. This requires at least one day of overlap in-between the forecasted and observed events. In total, about 25% of all the observed cases are well forecasted and are used for this validation. Based on this approach, the intensities, onset and durations are compared in Fig. 9. The delays on the onset and durations are first compared (Fig. 9 first and second columns), to detect any bias. The study of the onset delay reveals a good ability of the model to predict it. These scores are mainly explained by the short-term forecasts and it is important to notice that peaks in the histogram represent about 9% of all the observed events for the HWs and about 12% of the CWs. No specific positive or negative bias is found in the delays. Finally, the evaluation reveals longer durations in the forecasts highlighted by a positive bias of values in the HWs and CWs. The consequence of that is the slight positive delay in the forecast of wave ends in the two cases (not shown).

Then, the observed and forecasted intensities are compared. The method to calculate the intensity of the forecasted event is based on the median value of the member. This forecasted intensity is first unbiased using a quantile-quantile matching method. The correlation scores are 0.61 (0.65) for the HWs (CWs respectively) and the scatter plots in the right panels of Fig. 9 reveal the relative good agreements for intensities up to 5 and 10 for HWs and CWs respectively, while stronger intensities could not be evaluated due to large uncertainties. Nevertheless, these cases represent less than 1000 out of the 160,000 observed cases. Due to the relative short period of study to apply robust statistics, waves with intensities higher than 5 are considered as extreme events and no distinction will be done.



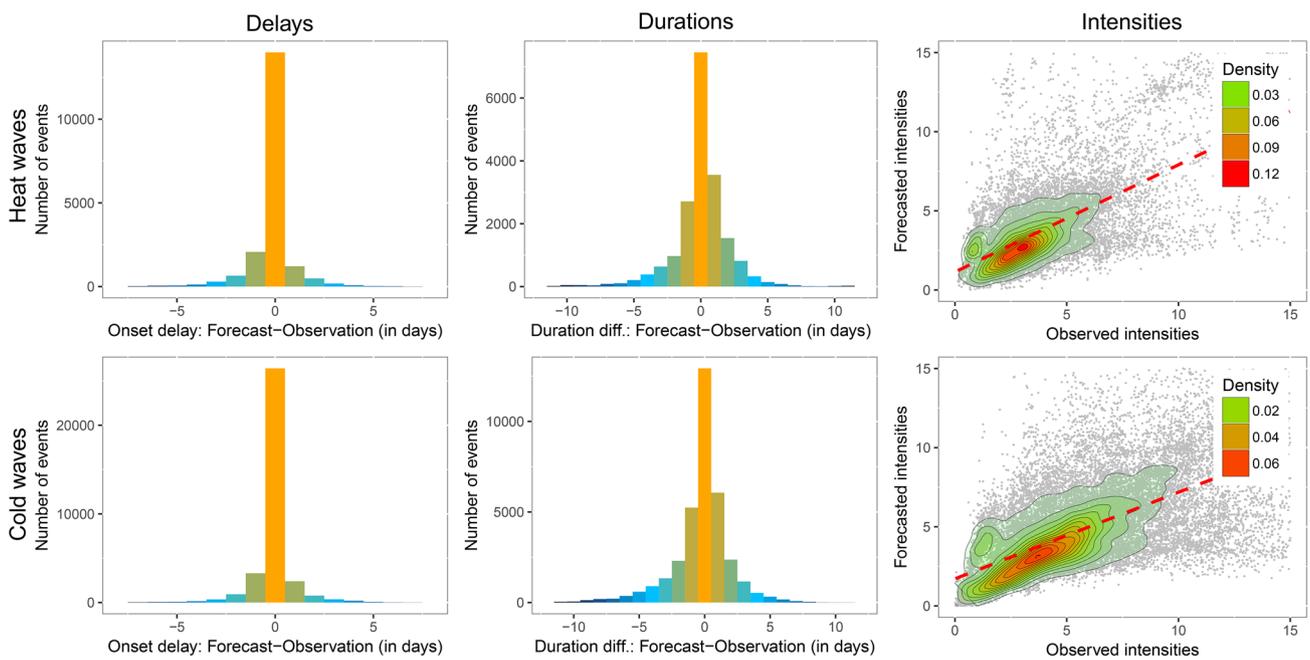
**Fig. 8** POD (left panels), FAR (middle panels) and ETS (right panels) of wave onsets (**a**) and of wave ends (**b**) for heat (top panels) and cold (bottom panels) events. The color lines indicate the accuracy of the forecasts, from 1 to 6-day window resolution

## 4 Conclusion

This study is the first statistical assessment of the predictability of HW and CW in Europe. It is based on 20 years of hindcast, from 1995 to 2015, using probabilistic ensemble forecasts provided by ECMWF and gridded station datasets.

Probabilistic scores, such as reliability and ROC, of predicting extreme temperature anomalies (defined above

and under Q90 and Q10) showed significant benefits up to a 2-week lead time. Based on that result, a reconstructed time series of 20 years was created by concatenating 15-day lead time forecasts every two weeks to assess the climatological evolution of the observed and forecasted events. Based on an algorithm defined in a previous study, the forecasted HWs/CWs are detected using 50% of forecast members for the HWs and 40% for the



**Fig. 9** (Left panels) histogram of onset delays (forecast-observation) in days of the same events, colors indicate the absolute values of the delays.(central panels) Histogram of durations anomalies (forecast-observation) in days of the same events, colors indicate the absolute

values of the duration differences. (Right panels) scatter plots of HWs (top) and CWs (bottom) intensities (I<sub>hum</sub>) for observed and forecasted events. For zone with high density of dots, the density is added in shaded. Red lines indicate the linear regressions

CWs and compared to the observed events. Results show a good representation of the spatial and temporal evolution of the events. The prediction of the most intense waves is more challenging with lower correlation values (from 0.5 to 0.73, depending on the methods and the heat or cold waves). Finally the trends reveal a good agreement between the forecasted and observed datasets, with a global and significant increase during all the summer months for the HWs.

The second main result of this study is the predictability of the HW/CW events in an operational context using all the hindcasts. To do so, two innovative approaches were developed. The first one is based on a daily assessment depending on the lead time, the second one, developed more specifically to validate the forecasted intensities, considers the entire HWs and CWs events and compares the forecasted and observed durations, intensities and onsets when they have, at least, one day overlap. Results show the sensitivity of the method to extract a dichotomous solution from the ensemble (variation for about 20% of the scores). For lead-times lower than one week, the best methods are generally those based on the ensemble median or a high percentage of ensemble members (forecasted event defined when 50% or more of members are associated with a HW/CW). Then, for longer lead-times, methods using less members (30 or 20%) appear the most reliable. According to the evolution of the best forecasts available, we could consider that forecasts at

1-week lead time stay robust and the limit of predictability is completely lost beyond 2-weeks lead time with a low spatial variability over Europe. Sensitivity tests are conducted and reveal a low sensitivity of the forecasts related to the temporal resolution and a small improvement of the scores for wider events. The predictability of the onsets (ends) of HWs and CWs appears more challenging, without significant information beyond a 1-week (10-day, respectively) lead time. Finally, the comparison of the intensities displays a significant agreement for the intensities up to 5 and 7. for the heat and cold waves, respectively, with a slight overestimation of the forecasted events due to a low bias in the HW and CW durations.

Thanks to this study, the most robust forecasts of HWs and CWs, including their uncertainties as a function of lead-time and location, can be provided to users and decisions makers. The main perspective of this work is to provide an extension of HW/CW forecasts towards a prediction of the related risks by taking into account the likely impacts.

**Acknowledgements** The authors would like to thank the anonymous reviewers for their fruitful and constructive comments.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate

credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Anderson BG, Bell ML (2009) Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States. *Epidemiology (Cambridge, Mass.)* 20(2):205
- Åström C, Ebi K, Langner J, Forsberg B (2014) Developing a heatwave early warning system for sweden: evaluating sensitivity of different epidemiological modelling approaches to forecast temperatures. *Int J Environ Res Public Health* 12(1):254–267
- Braga AL, Zanobetti A, Schwartz J (2002) The effect of weather on respiratory and cardiovascular deaths in 12 US cities. *Environ Health Perspect* 110(9):859
- Buizza R, Hollingsworth A, Lalauette F, Ghelli A (1999) Probabilistic predictions of precipitation using the ecmwf ensemble prediction system. *Weather Forecast* 14(2):168–189
- Cattiaux J, Vautard R, Cassou C, Yiou P, Masson-Delmotte V, Codron F (2010) Winter 2010 in Europe: a cold extreme in a warming climate. *Geophys Res Lett* 37(20):899–907
- Chalabi Z, Hajat S, Wilkinson P, Erens B, Jones L, Mays N (2016) Evaluation of the cold weather plan for england: modelling of cost-effectiveness. *Public health* 137:13–19
- Fouillet A et al (2006) Excess mortality related to the August 2003 heat wave in France. *Int Arch Occup Environ Health* 80(1):16–24
- Gasparrini A, Armstrong B (2011) The impact of heat waves on mortality. *Epidemiology (Cambridge, Mass.)* 22(1):68
- Ghosh A, Carmichael C, Elliot A, Green H, Murray V, Petrokofsky C (2014) The cold weather plan evaluation: an example of pragmatic evidence-based policy making? *Public health* 128(7):619–627
- Goddard L, Mason SJ, Zebiak SE, Ropelewski CF, Basher R, Cane MA (2001) Current approaches to seasonal to interannual climate predictions. *Int J Climatol* 21(9):1111–1152
- Gough WA, Tam BY, Mohsin T, Allen SM (2014) Extreme cold weather alerts in toronto, ontario, canada and the impact of a changing climate. *Urban Clim* 8:21–29
- Haiden T, Bauer P, Bidlot JR, Ferranti L, Hewson T, Prates F, Richardson D, Vitart F (2014) Evaluation of ECMWF forecasts, including 2013–2014 upgrades. Technical Memorandum, No. 742
- Hajat S et al (2013) Assessment of the implementation of the national cold weather plan for England. *Lancet* 382:S39
- Hajat S, Kovats RS, Lachowycz K (2007) Heat-related and cold-related deaths in england and wales: who is at risk? *Occup Environ Med* 64(2):93–100
- Haylock MR, Hofstra N, Klein Tank AMG, Klok EJ, Jones PD, New M (2008) A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J Geophys Res* 113:D20119. <https://doi.org/10.1029/2008JD010201>
- Henderson SB, Kosatsky T (2012) A data-driven approach to setting trigger temperatures for heat health emergencies. *Can J Public Health/Revue Canadienne de Sante'e Publique* 103(3):227–230
- Huynen M-M, Martens P, Schram D, Weijenberg MP, Kunst AE (2001) The impact of heat waves and cold spells on mortality rates in the dutch population. *Environ Health Perspect* 109(5):463
- IPCC (2012) Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the inter-governmental panel on climate change. Cambridge University Press, New York
- Jolliffe IT, Stephenson DB (2003) Forecast verification: a practitioner's guide in atmospheric science. Wiley, Oxford
- Jolliffe IT, Stephenson DB (2012) Forecast verification. Wiley, Oxford
- Kosatsky T, Henderson SB, Pollock SL (2012) Shifts in mortality during a hot weather event in Vancouver, British Columbia: rapid assessment with case-only analysis. *Am J Public Health* 102(12):2367–2371
- Kovats RS, Hajat S (2008) Heat stress and public health: a critical review. *Annu Rev Public Health* 29:41–55
- Lavaysse C, Vogt J, Pappenberger F (2015) Early warning of drought in Europe using the monthly ensemble system from ECMWF. *Hydrol Earth Syst Sci* 19(7):3273–3286
- Lavaysse C, Cammalleri C, Dosio A, van der Schrier G, Toreti A, Vogt J (2018) Towards a monitoring system of temperature extremes in Europe. *Natl Hazards Earth Syst Sci* 18(1):91–104. <https://doi.org/10.5194/nhess-18-91-2018>
- Lowe D, Ebi KL, Forsberg B (2011) Heatwave early warning systems and adaptation advice to reduce human health consequences of heatwaves. *Int J Environ Res Public Health* 8(12):4623–4648
- Masato G et al (2015) Improving the health forecasting alert system for cold weather and heat-waves in England: a proof-of-concept using temperature-mortality relationships. *PLoS One* 10(10):e0137804
- Matsueda M (2011) Predictability of Euro-Russian blocking in summer of 2010. *Geophys Res Lett* 38:L06801. <https://doi.org/10.1029/2010GL046557>
- Matzarakis A (2017) The heat health warning system of dwd-concept and lessons learned. *Perspectives on atmospheric sciences*. Springer, Berlin, pp 191–196
- McGregor GR, Bessemoulin P, Ebi KL, Menne B (2015) Heatwaves and health: guidance on warning-system development. World Meteorological Organization, Geneva
- Nagao K (2000) Method and apparatus for processing digital images to suppress their noise and enhancing their sharpness. Google Patents, US Patent 6,055,340
- Naumann G, Vargas WM (2012) A study of intraseasonal temperature variability in southeastern South America. *J Clim* 25(17):5892–5903
- Osman M, Alvarez MS (2017) Subseasonal prediction of the heat wave of December 2013 in southern South America by the Poama and BCC-CPS models. *Clim Dyn*. <https://doi.org/10.1007/s00382-017-3582-4>. ISSN:1432-0894
- PJM Interconnection (2014) Analysis of operational events and market impacts during the January 2014 cold weather events. <https://www.pjm.com/~media/documents/reports/20140509-analysis-ofoperational-events-and-market-impacts-during-the-jan-2014-cold-weather-events.aspx>
- Rocklöv J, Forsberg B, Ebi K, Bellander T (2014) Susceptibility to mortality related to temperature and heat and cold wave duration in the population of Stockholm County, Sweden. *Glob Health Act* 7(1):22 737
- Russo S, Sillmann J, Fischer EM (2015) Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environ Res Lett* 10(12):124 003
- Shaposhnikov D et al. (2014) Mortality related to air pollution with the Moscow heat wave and wildfire of 2010. *Epidemiology (Cambridge, Mass.)* 25(3):359.
- Stanojevic GB, Spalevic A, Kokotovic V, Stojilkovic J (2014) Does Belgrade (Serbia) need heat health warning system? *Disaster Prev Manag* 23(5):494–507
- Thakur C, Anand M, Shahi M (1987) Cold weather and myocardial infarction. *Int J Cardiol* 16(1):19–25
- Toth Z, Talagrand O, Candille G, Zhu Y (2003) Forecast verification: a practitioners guide in atmospheric science. Wiley, Oxford, pp 137–163
- Trigo RM, García-Herrera R, Díaz J, Trigo IF, Valente MA (2005) How exceptional was the early August 2003 heat wave in France? *Geophys Res Lett* 32:L10701. <https://doi.org/10.1029/2005GL022410>
- van den Besselaar EJM, Haylock MR, van der Schrier G, Klein Tank AMG (2011) A European daily high-resolution observational

- gridded data set of sea level pressure. *J Geophys Res* 116:D11110. <https://doi.org/10.1029/2010JD015468>
- Vitart F (2005) Monthly forecast and the summer 2003 heat wave over Europe: a case study. *Atmos Sci Lett* 6(2):112–117
- Vitart F (2014) Evolution of ECMWF sub-seasonal forecast skill scores. *Q J R Meteorol Soc* 140(683):1889–1899
- Werth D, Garrett A (2011) Patterns of land surface errors and biases in the global forecast system. *Mon Weather Rev* 139(5):1569–1582
- Wilks D (2011) *Statistical methods in the atmospheric sciences*. Academic, New York, p 100
- Zhang K, Chen Y-H, Schwartz JD, Rood RB, O’Neill MS (2014) Using forecast and observed weather data to assess performance of forecast products in identifying heat waves and estimating heat wave effects on mortality. *Environ Health Perspect* 122(9):912–918