# The Pprobabilistic Sskill of Extended-Range Heat Wwave Fforecasts Oover Europe

Natalia Korhonen[1], Otto Hyvärinen[1], Virpi Kollanus[2], Timo Lanki[2,3,4], Juha Jokisalo[5], Risto Kosonen[5,6], David S. Richardson[7], and Kirsti Jylhä[1]

[1] Weather and Climate Change Impact Research, Finnish Meteorological Institute, Helsinki, Finland
[2] Unit of Environmental Health, Department of Health Security, Finnish Institute for Health and Welfare, Kuopio, Finland.
[3] School of Medicine, University of Eastern Finland, Kuopio, Finland.
[4] Department of Environmental and Biological Sciences, University of Eastern, Kuopio, Finland.
[5] Department of Mechanical Engineering, Aalto University, Espoo, Finland
[6] College of Urban Construction, Nanjing Tech University, Nanjing, China
[7] European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

*Correspondence to*: Natalia Korhonen (Natalia.Korhonen@fmi.fi)

**Abstract.**

Severe heat waves lasting for weeks and expanding over hundreds of kilometres in horizontal scale have many harmful impacts on health, ecosystems, societies, and economy. Under the ongoing climate change heat waves are becoming even longer and hotter, and as proactive adaptation, the development of early warning services is essential.

Weather forecasts in the extended range (2 weeks to 1 month) tend to indicate a higher skill in predicting warm extremes than average temperature events in Europe. We verified hindcasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) in forecasting heat wave days, here defined as periods with the 5-day mean temperature being above its 90$^{th}$ percentile. The verification was done in 5° × 2° resolution over Europe, based on the forecast week (1 to 4 weeks). In the first forecast week, it is evident that across Europe, the accuracy of ECMWF heat wave forecasts surpasses that of a mere climatological forecast. Even into the second week, in many places in Europe, the ECMWF forecasts prove to be more reliable than their statistical counterparts. However, if we extend the forecast lead time to 3-4 weeks, predictability begins to lower to such a level that it can no longer be said, with the exception of Southeastern Europe, that the forecasts in general were statistically significantly better than the statistical forecast. Nonetheless, persistence of prolonged heat waves seems to have higher-than-average level of predictability even at a 3-week lead time, offering early warning services an indication of the potential duration of an ongoing heat wave.

## 1 Introduction

The severest heat waves in Europe since the 1950s have lasted from several weeks to even longer than a month, with horizontal spatial ranges exceeding several hundred kilometres, even 1000 km (Russo et al. 2015). In recent decades the number of

extreme heat waves over Europe and across the Northern Hemisphere has increased ~~and in the future,~~and, in the future, due to the ongoing climate change, heat waves are expected to become even more common and intense (IPCC, 2021; Russo et al. 2014; Coumou and Rahmstorf, 2012; Kim et al. 2018, Vogel et al. 2020, Ruosteenoja and Jylhä, 2023). This growing occurrence of heat waves underscores the urgent need to understand their dynamics and improve forecasting methods, especially for prolonged events with severe impacts.

Prolonged heat waves have negative impacts on, e.g., human health and wellbeing (Arsad et al. 2022; Guo et al. 2017, Ruuhela et al. 2021; Gasparrini et al. 2022, Kivimäki et al. 2023); labour productivity (Kjellstrom et al. 2009; Dunne et al. 2013; Orlov et al. 2019); energy and water resources (Añel et al. 2017; Hatvani-Kovacs et al. 2016; van Vliet 2023); transport systems (Mulholland & Feyen, 2021), wildfire safety (Rossiello and Szema, 2019; Ruffault et al., 2020); agriculture (Heino et al. 2023; Vogel et al. 2019), and livestock (Ahmed et al. 2022, ; Morignat et al. 2014). During heat waves, apartments lacking air conditioning gradually begin to overheat, which exacerbates heat stress (Velashjerdi Farahani et al. 2021; Velashjerdi Farahani et al. 2023; Velashjerdi Farahani et al. 2024a). The warm-up time of buildings related to outdoor temperature depends on building properties (U-value, ventilation airflow rate, and thermal mass of buildings). In Northern Europe, where apartments are typically not equipped with mechanical cooling systems, the thermal inertia of buildings plays a critical role. For instance, a Finnish study observed that buildings required 5-6 days to reach overheating conditions, highlighting the importance of the 5-day mean temperature as a predictor for indoor heat stress (Velashjerdi Farahani 2024a). In not well-insulated buildings and/or light structures, such as wooden ones, the warm-up time can be significantly shorter, often only 1–2 days. These findings emphasize the relevance of forecasting tools capable of predicting not only the occurrence but also the persistence of heat waves. ~~In Northern Europe apartments are typically not equipped with mechanical cooling systems. In a Finnish study, heating of buildings has been observed to take 5-6 days, during which the thermal mass of the building was able to slow down the indoor temperature rise (Velashjerdi Farahani 2024a).~~

Prolonged and intensive heat waves occurring over a wide area can lead to significant, and potentially catastrophic, impacts on public health. In Europe, the 2003 heat wave has been estimated to have resulted in over 70 000 (Robine et al. 2008) and the 2022 heat wave in over 60 000 (Ballester et al. 2023) heat-related deaths. As climate change progresses, severe health effects of heat waves are expected to further increase (Guo et al. 2018). Recognizing this, many countries in Europe and other parts of the world have developed heat-health action plans over the past 20 years to mitigate heat-related health risks (Kotharkar et al. 2022; Martinez et al. 2022, Martinez et al. 2019, Matthies et al. 2008). A key element of these preparedness plans consists of heat wave early warning systems, the operation of which is based on weather forecasts and pre-defined threshold criteria for triggering the warning services (Casanueva et al., 2019; Prodhomme, et al. 2021). As health effects of heat exposure occur quickly, at the same day or a few days lag (Baccini et al. 2008), it is imperative that the protection measures are implemented rapidly when a potentially dangerous heat wave is forecasted. However, organization of the response measures requires

**Formatted:** English (United States)

coordination of actions between many stakeholders and distribution of workforce, equipment, and other resources, which take time.

~~During the past 20 years, many countries in Europe and other parts of the world have taken steps to improve heat preparedness by developing heat-health action plans, which aim to reduce the negative health impacts (Kotharkar et al. 2022; Martinez et al. 2022). A key element of these preparedness plans consists of heat wave early warning systems, the operation of which is based on weather forecasts and pre-defined threshold criteria for triggering the warning services (Casanueva et al., 2019; Prodhomme, et al. 2021).~~ Effectiveness of the systems in preventing health effects depends on the ability to accurately forecast the impending heat event, as well as warning lead time. The lead time for heat wave warnings in each European country depends on the respective National Meteorological and Hydrological Services. Currently, heat wave warnings across Europe are typically issued 2-5 days in advance, and in some countries, such as Germany and the U.K., up to 7 days in advance. Extending these lead times could significantly enhance preparedness by allowing for earlier adaptive measures and better resource allocation, particularly for prolonged heat waves.

Sub-seasonal forecasts, which cover the extended range of 2 weeks to 1 month, offer a promising avenue for improving early warning systems.

The skill of these extended range forecasts ~~(2 weeks to 1 month, also called sub-seasonal forecasts)~~ has been found to be atmospheric flow dependent (Frame et al. 2013, Ferranti et al. 2015) and spatially heterogeneous. Vitart & Robertson (2018) highlighted the potential of sub-seasonal predictions in forecasting the progression of prolonged events like heat waves spanning multiple weeks. Moreover, Wulff and Domeisen (2019) and studies by Pyrina and Domeisen (2023) emphasized that extended-range predictions were more successful in forecasting extreme hot summer temperatures in Europe compared to predicting average summer temperatures.

Weather forecasts can be divided into two main categories: deterministic and probability forecasts. Deterministic forecasts provide a single specific scenario for future weather. For example, "tomorrow will be hot" is a deterministic forecast that offers one possible future event. Probability forecasts, on the other hand, provide various possible scenarios and their probabilities, taking into account the uncertainty of the forecast. For instance, "50% chance of heat" is a probability forecast indicating that heat may occur, but it's not certain. As the uncertainty of extended-range forecasts is known to be large, we evaluated their probabilistic rather than deterministic skill. There is a large literature in statistics and decision analysis on the use of probabilistic information in so-called decision making under uncertainty (e.g. Clemen 1996). In theory and practice, probabilistic forecasts have been shown to contain more information and should be more valuable to users than categorical, deterministic forecasts (e.g., Murphy 1977, Richardson 2001), though their practical utility depends on users' ability to

incorporate such information into decisions . ~~A practical question is how well users can use the forecasts in their decision making~~ (e.g., Lopez & Haines 2017; Ramos et al. 2013).

Our objective was to assess the probabilistic skill of the extended-range forecasts made by the European Centre for Medium-Range Weather Forecasts (ECMWF) in predicting *heat wave days*, defined as periods where the local 5-day mean temperature exceeded the 90$^{th}$ percentile of the local summertime 5-day mean temperature distribution. We assessed the reliability of forecasts predicting heat waves surpassing this threshold, as this type of heat waves has~~ve~~ been shown to significantly increase the risk of overheating in apartments in Finland (Velashjerdi Farahani et al. 2024a) and elevate mortality risk among the elderly (Kollanus et al. 2021). Moreover, in an empirical study conducted in Finland, indoor temperatures were found to be more strongly correlated with outdoor 5-day moving average temperature than with average temperatures of a few days only, suggesting impacts of building's thermal inertia (Velashjerdi Farahani et al., 2024b). Our verification process was conducted using a resolution of 5 degrees longitude and 2 degrees latitude (5° × 2°) over Europe for the summers spanning from 2000 to 2019. We examined hindcasts for various lead times, ranging from 1 to 4 weeks. The novelty of the study arises from the verification area encompassing the entirety of the European region allowing to highlight potential regional differences in the forecast skill, as well as from evaluating the model's ability to forecast the life cycle of heat waves, taking into account the forecast initialization (date) relative to the onset of the heat wave ~~the relative time of forecast issuance and heat wave initiation~~.

## 2 Data and Methods

### 2.1 Definition of Heat Wave Days

Formatted: Font color: Auto

In this study, heat wave days were defined as periods where the local 5-day moving average temperature ($T^{5d}$) exceeded its local summertime 90th percentile ($^{90th}T^{5d}$). To calculate $T^{5d}$, local daily mean temperatures over land areas were averaged over a forward-looking 5-day window. The threshold $^{90th}T^{5d}$ was determined using summer data (June-July-August), ensuring that the definition reflected summertime extreme temperatures. By applying this threshold, the continuous variable $T^{5d}$ was converted into a binary variable: days were categorized as either heat wave days ($T^{5d} > {}^{90th}T^{5d}$) or non-heat wave days. In this study, heat wave days served as the forecast target. The choice of the 5-day moving average enables more robust identification of sustained heat wave events by reducing the influence of short-term variability, which is particularly important for extended-range forecasting, since such forecasts are not expected to skilfully predict small-scale day-to-day variability.

Formatted: Font color: Auto

Our definition of heat wave days is meaningful as it aligns with thresholds commonly used in epidemiological studies on heat-related health effects, where heat waves are typically defined as periods when daily temperatures exceed the 90th percentile of the local annual or summertime temperature distribution for two or more consecutive days (Arsad et al., 2022). Such heat waves have been observed to lead to increased mortality and morbidity worldwide (Arsad et al., 2022; Guo et al., 2017). Although high temperature (dry bulb) is the primary variable for assessing heat wave impacts, other factors, such as humidity

and wind speed, also contribute to heat stress. Nevertheless, this study focuses solely on temperature as the key driver of heat stress.

## 2.2 ERA5 data

### 2.2.1 Thresholds for heat wave days

For defining *observed heat wave days* with a horizontal resolution of 5 degrees longitude and 2 degrees latitude (5° × 2°) over Europe (36 to 70° N and -7.5 to 52.5° E) during summers 2000-2019, we used the ERA5 near-surface air temperature reanalysis data (Hersbach et al., 2020). The ERA5 data, with a horizontal resolution of 0.1°, were bilinearly interpolated to a 5° × 2° grid, considering only land grid points. To define heat wave days, we calculated the 5-day moving average temperatures ($T^{5d}_{ERA5}$) for each grid point across Europe during the summers of 2000-2019, and defined periods with $T^{5d}_{ERA5}$ exceeding its 90$^{th}$ percentile ($^{90th}T^{5d}_{ERA5}$) as observed heat wave days. Figure 1a depicts a map of the 90$^{th}$ percentile of the 5-day moving average temperature (in summers 2000-2019) over Europe, in ERA5. Days having ERA5 5-day moving average temperatures above the thresholds, the 90$^{th}$ percentile, were in this study defined as observed heat wave days.

**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto
**Formatted:** Font color: Auto

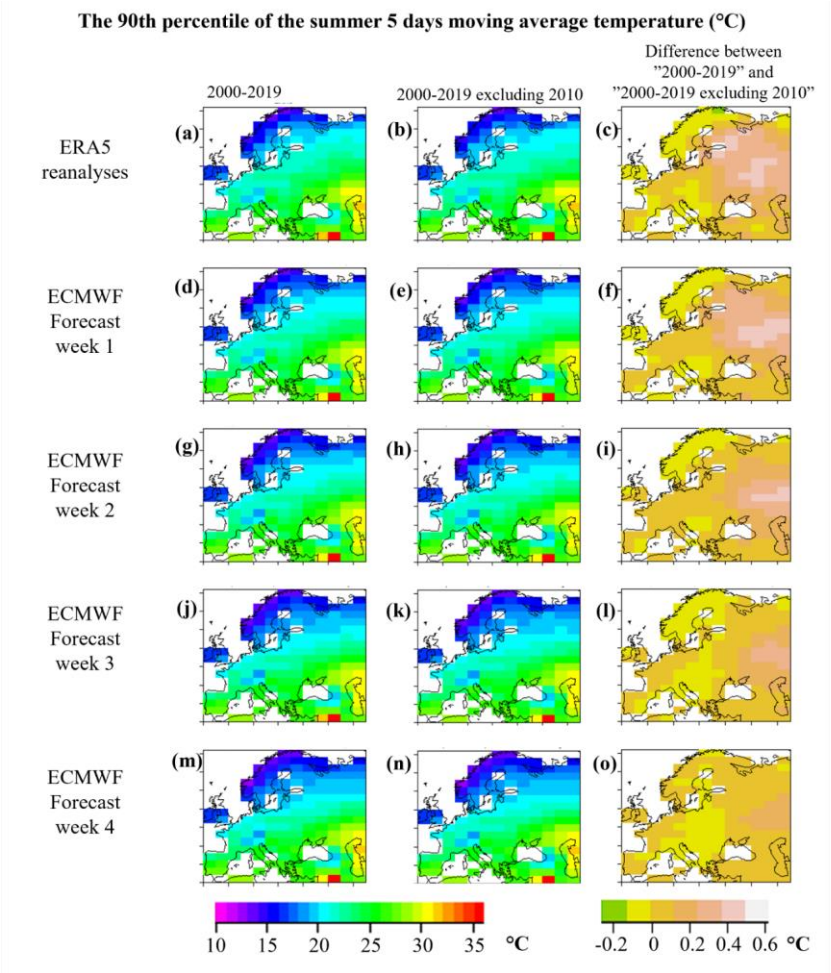



Figure 1: The lower thresholds of heat wave days: the 90th percentile of the 5-day moving average temperature in summers 2000-2019 (first column) and in summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010, middle column) of the ERA5 reanalyses (a and b), and (d,e,g,h,j,k,m, and n) of the ensembles of the ECMWF's hindcasts in different forecast weeks. The last column shows the difference between these two.

Formatted: Font color: Auto

As our definition of heat waves was based on 5-day mean temperatures, rather than daily mean temperatures as commonly used in epidemiological studies on heat-related health effect, we examined the proportion of days where the daily mean temperature, the $T^{1d}_{ERA5}$ exceeded its 90th percentile ($^{90th}T^{1d}_{ERA5}$) within periods with $T^{5d}_{ERA5}$ exceeding its 90$^{th}$ percentile here defined as heat waves. For this we computed daily mean temperatures, the $T^{1d}_{ERA5}$ and their 90th percentiles ($^{90th}T^{1d}_{ERA5}$) across European land areas from 2000 to 2019. For each grid point, we determined the percentage of five-day periods exceeding the $^{90th}T^{5d}_{ERA5}$ that included days where the $T^{1d}_{ERA5}$ exceeded its 90th percentile. Our investigations showed that our definition for heat waves based on exceeding the $^{90th}T^{5d}_{ERA5}$ covered 26% of the one-day heat waves based on exceeding the $^{90th}T^{1d}_{ERA5}$. For the two-days heat waves based on exceeding the $^{90th}T^{1d}_{ERA5}$ our definition covered 61%. The three-days heat waves based on exceeding the $^{90th}T^{1d}_{ERA5}$ were covered 96% by our definition. For four or more consecutive day heat wave events based on exceeding the $^{90th}T^{1d}_{ERA5}$ our definition covered 100%. These statistics show that the 5-day moving average definition covers nearly all longer heat wave events (such as 3- to 4-day heat waves), but only a portion of shorter ones (1- to 2-day heat waves), indicating that the 5-day moving average is particularly useful for identifying sustained heat wave events.

During the period 2000-2019, the summer 2010 was characterized by a particularly long-lasting heat wave over Europe (e.g., Trenberth & Fasullo, 2012). Therefore, we investigated the weight of this event on our results by comparing our results for the period 2000-2019 with and without year 2010. Figure 1b gives a spatial distribution, with 1 °C intervals, for the threshold of the heat wave days for the period 2000-2019 without summer 2010. Figure 1c shows the impacts of including 2010: in most of the western and the southern Europe the difference is ±0.1°C, while in the eastern and north-eastern parts of Europe the impact is mostly between 0 and +0.55 °C, except for the very northern Fennoscandia where the impact is between -0.2 and 0 °C. Compared to the large northwest-southeast gradient of the absolute values of the 90th percentile in the Fig 1a and 1b, these differences are minor.

### 2.2.2 The frequency and duration of the heat wave days

To identify the summer with the longest heat wave, we examined the frequency and duration of heat wave days in the ERA5 reanalysis data. A heat wave was considered to be any period of at least one day where the 5-day moving average temperature remained above the 90$^{th}$ percentile of $T^{5d}_{ERA5}$. The heat wave was considered interrupted when there were two consecutive days with temperatures falling below the 90$^{th}$ percentile of $T^{5d}_{ERA5}$. To clarify, a single day below the threshold did not end the heat wave as long as it continued afterward.

The durations of the longest heat wave events in each grid point over Europe in summers 2000-2019, as derived from ERA5, are depicted in Figure 2(a). The heat wave events were longest in Eastern Europe. Figure 2(a) highlights the extreme heat wave of 2010 in the east, the heat wave of 2018 in the north and parts of Central Europe, and the heat wave of 2003 in parts of south and southwest. Figure 2b indicates that if the summer 2010 is excluded, other years (e.g., 2014) appear in eastern Europe /

Formatted: Font color: Auto

Formatted ...

western Russia, compared to Fig. 2a, and the duration of the longest period of heat wave days get shorter there. Figure 2(c), showing the number of different heat wave events, highlights that in these summers 2000-2019 the heat wave days in Northern Europe and in many parts of Eastern Europe were concentrated within fewer periods, whereas in the Central and Southwestern Europe, the same amount of heat wave days were distributed across a larger number of periods. Figure 2d shows that if the summer 2010 is excluded, especially in those areas where 2010 had the longest period of heat wave days, excluding it leads to an increase in the number of periods with heat wave days, as the 10% of the hottest days are now distributed to a larger number of events.

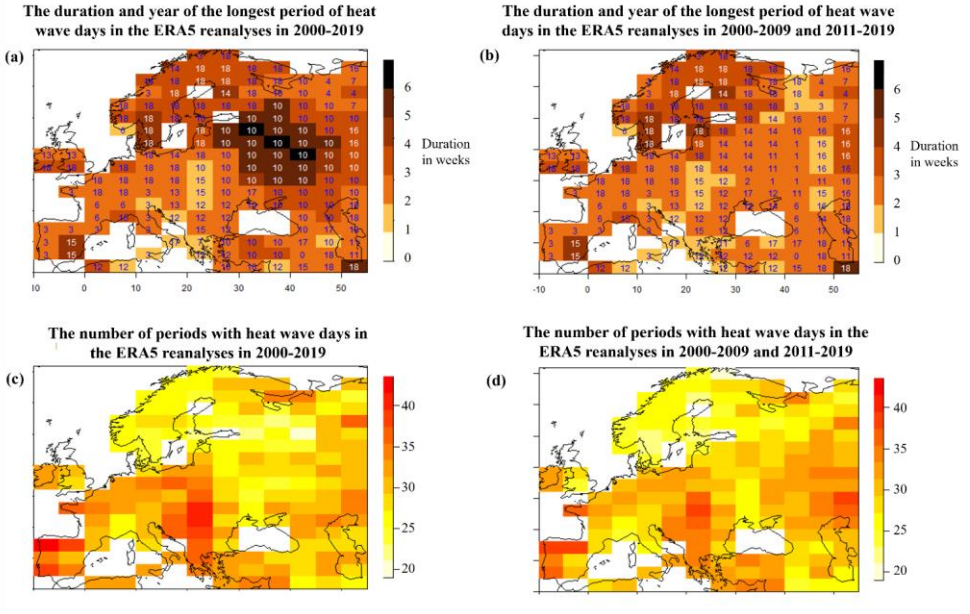



**Figure 2: The duration and the year (marked as 0-19) of the longest period of heat wave days defined from the ERA5 reanalysis data of (a) summers 2000-2019 and (b) summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010), and the number of periods with heat wave days (b) in the ERA5 reanalyses during (c) 2000-2019 and (d) 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010).**

Formatted: Font: Bold, Font color: Auto, English (United Kingdom)

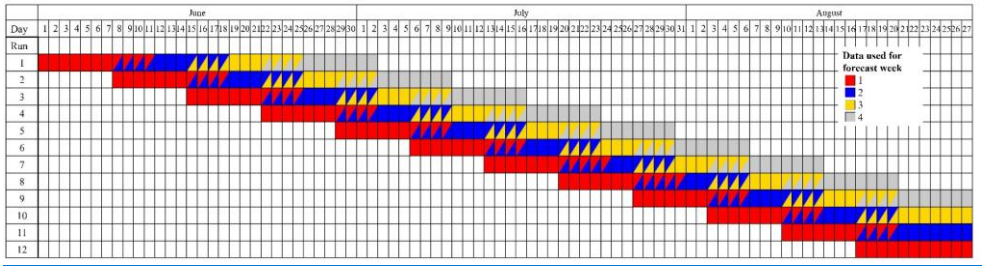
## 2.3 Hindcasts

Formatted: Font color: Auto

Hindcasts, also known as reforecasts, are a type of retrospective weather forecasts. Hindcasts are forecasts of past weather conditions, generated using forecasting models, data assimilation methods, and observational data identical to those used for real-time weather predictions. Here we verified hindcasts of the European Centre for Medium-Range Weather Forecasts

(ECMWF) Integrated Forecasting System (IFS; Cycles 46r1 and 47r1; Vitart, 2014). These hindcasts were run at the ECMWF in 2020 twice a week, on Mondays and Thursdays, initiated using the ERA5 analyses. We investigated 240 hindcasts which were run with a weekly interval for the summers 2000–2019, i.e., 20 years × 12 weeks =240 hindcasts, for details, see Table 1.

**Table 1. Table showing details of the data of the investigated hindcasts. Each row contains one run, altogether 12 runs per year. The colouring of the boxes shows the coverage of the hindcasts' data. The first red boxes on each row show the initialization date of the hindcasts, which are same for all years 2000-2019. The colours of the boxes indicate for which lead time (i.e., forecast week) the data were used: red for 1 week, blue for 2 weeks, yellow for 3 weeks, and grey for 4 weeks. The data used for the different lead times were partially overlapping due to the use of 5-days moving averages with forward-looking window: lead time 1 week used data of days 1 to 11, lead time 2 weeks data of days 8 to 18, lead time 3 weeks data of days 15 to 25, and lead time 4 weeks data of days 22 to 32. The data used for two lead times are here marked with two colours. Note: for lead time 1 week we used data of 12 runs, for lead time 2 weeks we used data of 11 runs, for lead time 3 weeks we used data of 10 runs, and for lead time 4 weeks we used data of 9 runs (of years 2000-2019).**




We examined the 2m temperature (i.e. the near-surface air temperature) from the hindcasts with lead times of 1 to 32 days of the Monday runs. As the 2m temperature has a large temporal autocorrelation, using both the Monday and Thursday initializations would not have added much information and would only have complicated the statistical analysis. We therefore decided to use only the Monday runs. The decision is arbitrary, and we could have chosen to use only the Thursday runs as well. The ECMWF reforecasts were initially run at a horizontal resolution of approximately 18 km for the first 15 days and then re-initialized at a coarser resolution of around 36 km for days 15 to 46. For our verifications, we used ECMWF's hindcasts at a horizontal resolution of 0.4° which were bilinearly interpolated to a 5° × 2° grid, considering only land grid points.

The hindcasts consisted of a control forecast and 10 perturbed ensemble members, making up 11 members in total. It is important to distinguish between the hindcasts and the operational real-time forecasts, which initially had 51 members and now consist of 101 members (IFS Cycle 48r1). Consequently, the results obtained here from the 11-member hindcasts serve as a baseline measure of skill (see, e.g., Richardson 2001, Ferro et al. 2008) and the larger operational ensemble is expected to provide improved estimates of the normal distribution parameters, thereby enhancing skill to some extent.

Formatted: Font color: Auto

Formatted: Font color: Auto

### 2.3.1 Thresholds for forecasted heat wave days

For verification of the hindcasts, we defined the 5-day moving average temperatures in the ECMWF's hindcasts, $T_{\blacktriangle C\blacktriangle}^{5d}$. The calculations of $T_{\blacktriangle C\blacktriangle}^{5d}$ were performed separately for each of the 11 ensemble members, covering each day from June 1st to August 27th (88 days) over the summers of 2000-2019. For each day within this period, we incorporated forecasted mean temperatures for that day and the subsequent four days into the calculations of $T_{\blacktriangle C\blacktriangle}^{5d}$. For each grid point and for each forecast week, ranging from week 1 to week 4, we determined the threshold for *a forecasted heat wave day* by calculating the 90th percentile, the ${}^{90th}T_{\blacktriangle C\blacktriangle}^{5d}$, of the 5-day moving average temperatures, $T_{\blacktriangle C\blacktriangle}^{5d}$ of all days under consideration of the summers 2000-2019. The forecast data used for the forecast weeks were partially overlapping due to the use of 5-days moving averages with forward-looking window: the forecast week 1 used data of days 1 to 11, the forecast week 2 data of days 8 to 18, forecast week 3 data of days 15 to 25, and forecast week 4 data of days 22 to 32 as depicted in Table 1.

In Figure 1 the first column depicts maps of the 90$^{th}$ percentile of the 5-day moving average temperature (in summers 2000-2019) over Europe, in ERA5 (Figure. 1a) and in the ECMWF hindcasts for forecast weeks 1-4 (Figs.1d, 1g, 1j, and 1m). As stated earlier, days having ERA5 5-day moving average temperatures above the thresholds, the 90$^{th}$ percentile, were in this study defined as observed heat wave days. The ECMWF hindcasts capture the northwest-southeast gradient in the threshold of the heat wave days, even though the absolute values are somewhat lower in the hindcasts than in ERA5, and this difference is growing with the lead time.

Summer 2010 was marked by an unusually prolonged heat wave over Europe. In Figure 1, the middle column depicts the spatial distribution of the thresholds for observed and forecasted heat wave days over the period 2000–2019, excluding summer 2010. The last column of Figure 1 (Figures 1c, 1f, 1i, 1l, and 1o) illustrates the impact of including 2010. Compared to the large northwest-southeast gradient of the absolute heat wave day thresholds in the first two columns, the differences in the last column are minor. For assessing the impact of the summer of 2010 on the probabilistic skill of heat wave forecasts, the threshold values in the middle column are used.

### 2.3.2. Probability forecasts

The forecasted probability of a heat wave day, *p*, was here based on fitting a normal distribution to the $T_{\blacktriangle C\blacktriangle}^{5d}$ forecasts of the 11-member ensemble (practically a set of deterministic forecasts) and defining the probability of the forecasted $T_{\blacktriangle C\blacktriangle}^{5d}$ being above the ${}^{90th}T_{\blacktriangle C\blacktriangle}^{5d}$ on each day. Hence, a heat wave in the forecast is defined relative to the forecast model's climatology. Moreover, the comparison of the hindcasts to the lead time dependent model climatology is expected to remove the systematic frequency bias resulting from the forecast model drift (Manzanas, 2020).

Formatted: Heading 3

Formatted: Font color: Auto

In the verification, the forecast model-based probability of a heat wave day, $p$, was compared to the observed heat wave days (Section 2.2.1) derived from the ERA5 dataset. Since we used the data from the entire period (years 2000–2019) to define the heat wave day thresholds, we may achieve an overestimation of the forecast skill in the verification compared to using a leave-one-out method (in which one year is excluded at a time from the dataset when defining the threshold). However, as shown in the last column of Figure 1, excluding even the most extreme year has only a minimal impact on the threshold definition. Therefore, it is reasonable to assume that the effect on the skill is not substantial.

Formatted: Font: Italic

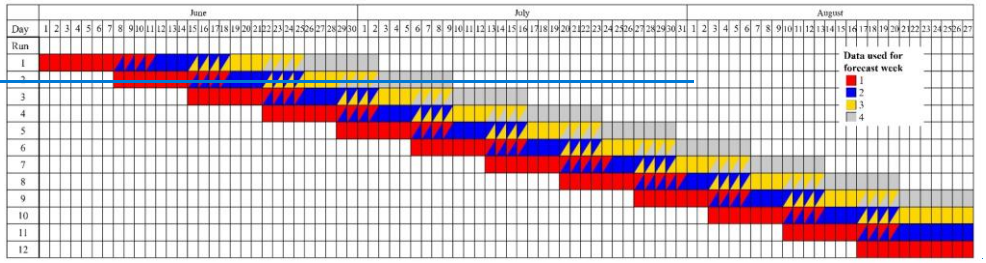
Formatted: Normal

~~**2.1 Hindcasts and verification data**~~

~~Hindcasts, also known as reforecasts, are a type of retrospective weather forecasts. Hindcasts are forecasts of past weather conditions, generated using forecasting models, data assimilation methods, and observational data identical to those used for real-time weather predictions. Here we verified hindcasts of the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS; Cycles 46r1 and 47r1; Vitart, 2014). These hindcasts were run at the ECMWF in 2020 twice a week, on Mondays and Thursdays. We investigated 240 hindcasts which were run with a weekly interval for the summers 2000–2019, i.e., 20 years × 12 weeks =240 hindcasts, for details, see Table 1.~~

~~We examined the 2m temperature (i.e. the near-surface air temperature) from the hindcasts with lead times of 1 to 32 days of the Monday runs. As the 2m temperature has a large temporal autocorrelation, using both the Monday and Thursday initializations would not have added much information and would only have complicated the statistical analysis. We therefore decided to use only the Monday runs. The decision is arbitrary and we could have chosen to use only the Thursday runs as well. Since these hindcasts were initiated using the ERA5 analyses, we used the ERA5 2000-2019 near-surface air temperature reanalyses (Hersbach et al., 2020) for verification. Here the area of interest was Europe (36 to 70° N and -7.5 to 52.5° E). The ECMWF reforecasts were initially run at a horizontal resolution of approximately 18 km for the first 15 days and then re-initialized at a coarser resolution of around 36 km for days 15 to 46. For our verifications, we used ECMWF's hindcasts at a horizontal resolution of 0.4° and ERA5 reanalysis data at 0.1°, both of which were bilinearly interpolated to a coarser 5° × 2° grid, considering only land grid points.~~

~~**Table 1. Table showing details of the data of the investigated hindcasts. Each row contains one run, altogether 12 runs per year. The colouring of the boxes shows the coverage of the hindcasts' data. The first red boxes on each row show the initiation date of the hindcasts, which are same for all years 2000–2019. The colours of the boxes indicate for which lead time (i.e., forecast week) the data were used: red for 1 week, blue for 2 weeks, yellow for 3 weeks, and grey for 4 weeks. The data used for the different lead times were partially overlapping due to the use of 5-days moving averages with forward-looking window: lead time 1 week used data of days 1 to 11, lead time 2 weeks data of days 8 to 18, lead time 3 weeks data of days 15 to 25, and lead time 4 weeks data of days 22 to 32. The data used for two lead times are here marked with two colours. Note: for lead time 1 week we used data of 12 runs, for lead time 2 weeks we used data of 11 runs, for lead time 3 weeks we used data of 10 runs, and for lead time 4 weeks we used data of 9 runs (of years 2000–2019).~~




The hindcasts consisted of a control forecast and 10 perturbed ensemble members, making up 11 members in total. It is important to distinguish between the hindcasts and the operational real-time forecasts, which initially had 51 members and now consist of 101 members (IFS Cycle 48r1). Consequently, the results obtained here from the 11-member hindcasts serve as a baseline measure of skill (see, e.g., Richardson 2001, Ferro et al. 2008) and the larger operational ensemble is expected to provide improved estimates of the normal distribution parameters, thereby enhancing skill to some extent.

**2.2 Definition of heat wave days**

As extended-range forecasts are not expected to capture small-scale day-to-day variability, we applied a 5-day moving average to the temperature forecasts for evaluation. More specifically, for the ECMWF's hindcasts, we calculated the 5-day moving average temperatures, $T_{EC}^{5d}$. This was done separately for each ($5^{\circ} \times 2^{\circ}$) grid point. The calculations of $T_{EC}^{5d}$ were performed separately for each of the 11 ensemble members, covering each day from June 1st to August 27th (88 days) over the summers of 2000-2019. For each day within this period, we incorporated forecasted mean temperatures for that day and the subsequent four days into the calculations of $T_{EC}^{5d}$. For each grid point and for each forecast week, ranging from week 1 to week 4, we determined the threshold for *a heat wave day* by calculating the 90th percentile, the ${}^{90th}T_{EC}^{5d}$, of the 5-day moving average temperatures, $T_{EC}^{5d}$, of all days under consideration of the summers 2000-2019. The forecast data used for the forecast weeks were partially overlapping due to the use of 5-days moving averages with forward-looking window: the forecast week 1 used data of days 1 to 11, the forecast week 2 data of days 8 to 18, forecast week 3 data of days 15 to 25, and forecast week 4 data of days 22 to 32 as depicted in Table 1.

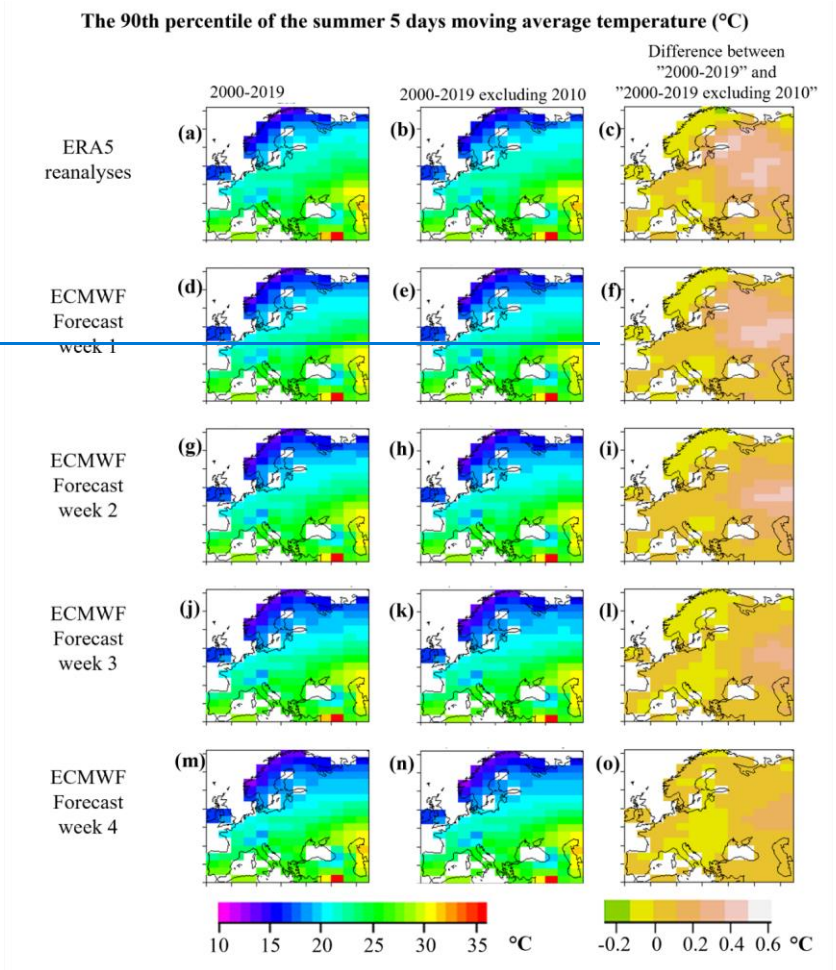
Weather forecasts can be divided into two main categories: deterministic and probability forecasts. Deterministic forecasts provide a single specific scenario for future weather. For example, "tomorrow will be hot" is a deterministic forecast that offers one possible future event. Probability forecasts, on the other hand, provide various possible scenarios and their probabilities, taking into account the uncertainty of the forecast. For instance, "50% chance of heat" is a probability forecast indicating that heat may occur, but it's not certain. Given the substantial uncertainty in extended-range forecasts, we evaluate the forecasts'

probabilistic skill rather than their deterministic accuracy. The forecasted probability of a heat wave day, *p*, was here based on fitting a normal distribution to the $T_{EC}^{5d}$ forecasts of the 11-member ensemble and defining the probability of the forecasted $T_{EC}^{5d}$ being above the ${}^{90th}T_{EC}^{5d}$ on each day. The use of percentiles for the hindcasts leads to forecasting in the model's climatology. Moreover, the comparison of the hindcasts to the lead time dependent model climatology is expected to remove the systematic frequency bias resulting from the forecast model drift (Manzanas, 2020).

For verification of the hindcasts, we separately defined the 5-day moving average temperature in the ERA5 ($T_{ERA5}^{5d}$) data in the corresponding grid points over Europe, in summers 2000-2019, and defined periods with $T_{ERA5}^{5d}$ exceeding its 90$^{th}$ percentile (${}^{90th}T_{ERA5}^{5d}$) as observed heat wave days.

### 2.2.1 Thresholds for the heat wave days

In Figure 1 the first column depicts maps of the 90$^{th}$ percentile of the 5-day moving average temperature (in summers 2000-2019) over Europe, in ERA5 (Figure. 1a) and in the ECMWF hindcasts for forecast weeks 1-4 (Figs.1d, 1g, 1j, and 1m-1e). Days having ERA5 5-day moving average temperatures above the thresholds, the 90$^{th}$ percentile, were in this study defined as heat wave days. The ECMWF hindcasts capture the northwest-southeast gradient in the threshold of the heat wave days, even though the absolute values are somewhat lower in the hindcasts than in ERA5, and this difference is growing with the lead time.




~~**Figure 1: The lower thresholds of heat wave days: the 90th percentile of the 5-day moving average temperature in summers 2000-2019 (first column) and in summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010, middle column) of the ERA5 reanalyses (a and b), and (d,e,g,h,j,k,m, and n) of the ensembles of the ECMWF's hindcasts in different forecast weeks. The last column shows the difference between these two.**~~

In epidemiological studies on heat-related health effects, heat waves are typically defined as periods when *daily temperatures* exceed the 90th percentile of the local annual or summertime temperature distribution for *two or more consecutive days* (Arsad et al. 2022). Such heat waves have been observed to lead to increased mortality and morbidity all around the world (Arsad et al. 2022, Guo et al. 2017). Although high temperature (dry bulb) is a primary variable for assessing heat wave impacts, it is important to note that heat stress can also be influenced by other factors, such as humidity and wind speed. Nevertheless, in this study, we focus solely on the key driver of heat stress, the temperature. As our definition for the heat waves was based on 5-day mean temperatures, we examined the proportion of days where the daily mean temperature, the $T^{1d}_{ERA5}$, exceeded its 90th percentile ($^{90th}T^{1d}_{ERA5}$) within periods with $T^{5d}_{ERA5}$ exceeding its 90$^{th}$ percentile here defined as heat waves. For this we computed daily mean temperatures, the $T^{1d}_{ERA5}$, and their 90th percentiles ($^{90th}T^{1d}_{ERA5}$) across European land areas from 2000 to 2019. For each grid point, we determined the percentage of five-day periods exceeding the $^{90th}T^{5d}_{ERA5}$ that included days where the $T^{1d}_{ERA5}$ exceeded its 90th percentile. Our investigations showed that our definition for heat waves based on exceeding the $^{90th}T^{5d}_{ERA5}$, covered 26% of the one-day heat waves based on exceeding the $^{90th}T^{1d}_{ERA5}$. For the two-days heat waves based on exceeding the $^{90th}T^{1d}_{ERA5}$, our definition covered 61%. The three-days heat waves based on exceeding the $^{90th}T^{1d}_{ERA5}$ were covered 96% by our definition. For four or more consecutive day heat wave events based on exceeding the $^{90th}T^{1d}_{ERA5}$, our definition covered 100%.

As during the period 2000-2019, the summer 2010 was characterized by a particularly long-lasting heat wave over Europe, we investigated the weight of this event on our results by comparing our results for the period 2000-2019 with and without year 2010. In Figure 1 the middle column gives a spatial distribution, with 1 °C intervals, for the threshold of the heat wave days for the period 2000-2019 without summer 2010. The last column in Figure 1 (Figures 1c, 1f, 1i, 1l, and 1o) shows the impacts of including 2010: in most of the western and the southern Europe the difference is ±0.1°C, while in the eastern and north-eastern parts of Europe the impact is mostly between 0 and +0.55 °C, except for the very northern Fennoscandia where the impact is between -0.2 and 0 °C. Compared to the large northwest-southeast gradient of the absolute values of the 90th percentile in the first two columns, these differences are minor. When assessing the impact of the summer of 2010 (with the long heat wave in Europe) on the probabilistic skill of the heat wave forecasts, the threshold values of the middle column are utilized.

**2.2.2 The frequency and duration of the heat wave days**

To identify the summer with the longest heat wave, we examined the frequency and duration of heat wave days in the ERA5 reanalysis data. A heat wave was considered to be any period of at least one day where the 5-day moving average temperature remained above the 90$^{th}$ percentile of $T^{5d}_{ERA5}$. The heat wave was considered interrupted when there were two consecutive days with temperatures falling below the 90$^{th}$ percentile of $T^{5d}_{ERA5}$.To clarify, a single day below the threshold did not end the heat wave as long as it continued afterward.

The durations of the longest heat wave events in each grid point over Europe in summers 2000-2019, as derived from ERA5, are depicted in Figure 2(a). The heat wave events were longest in Eastern Europe. Figure 2(a) highlights the extreme heat wave of 2010 in the east, the heat wave of 2018 in the north and parts of Central Europe, and the heat wave of 2003 in parts of south and southwest. Figure 2b indicates that if the summer 2010 is excluded, other years (e.g., 2014) appear in eastern Europe / western Russia, compared to Fig. 2a, and the duration of the longest period of heat wave days get shorter there. Figure 2(c), showing the number of different heat wave events, highlights that in these summers 2000-2019 the heat wave days in Northern Europe and in many parts of Eastern Europe were concentrated within fewer periods, whereas in the Central and Southwestern Europe, the same amount of heat wave days were distributed across a larger number of periods. Figure 2d shows that if the summer 2010 is excluded, especially in those areas where 2010 had the longest period of heat wave days, excluding it leads to an increase in the number of periods with heat wave days, as the 10% of the hottest days are now distributed to a larger number of events.

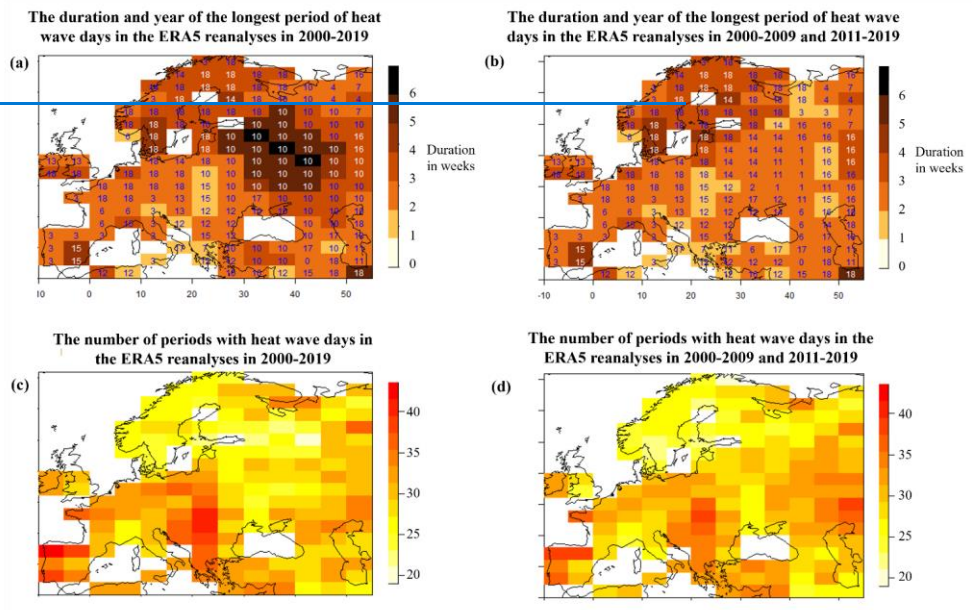



**Figure 2: The duration and the year (marked as 0-19) of the longest period of heat wave days defined from the ERA5 reanalysis data of (a) summers 2000-2019 and (b) summers 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010), and the number of periods with heat wave days (b) in the ERA5 reanalyses during (c) 2000-2019 and (d) 2000-2009 and 2011-2019 (i.e., 2000-2019 excluding 2010).**

### 2.43 Skill scores

The Brier Scores (*BS*, Brier, 1950) of the probabilistic forecasts, *p*, were calculated separately for each grid point and for forecast weeks 1 to 4 as follows:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(p_t - o_t)^2 \,, \quad (1)$$

where $p_t$ is the forecasted probability of a heat wave day, *p*, ranging from 0 to 1, $o_t$ is the actual outcome (based on ERA5 reanalysis) of the heat wave day at instance *t* (0 if there is no heat wave day and 1 if there is a heat wave day), and *N* is the number of forecasting instances. The *BS* is thus here equivalent to the mean squared error of the probability (of the heat wave day), and ranges from 0 to 1. The lower the *BS*, the better the predictions.

It follows from the use of the 90th percentile to define a heat wave day (Sec. 2.2) that the expected probability $p_c$ of a heat wave day is 0.1. This value, also referred as the climatological base rate ~~of~~ $p_c$ was used in Eq. (1) to calculate $BS_{ref}$, i.e., the Brier Score of the reference forecast. The Brier Skill Score (*BSS*) can now be defined as

$$BSS = 1 - \frac{BS}{BS_{ref}} \,, \quad (2)$$

The value of the *BSS* ranges from -∞ to +1: positive values indicate better skill than that of the reference forecasts, and *BSS* value of 1 represents the best possible score.

Initially, we calculated *BSS* for each grid point using data from all 240 hindcasts. To demonstrate the impact of long heat waves on the overall *BSS* of all hindcasts, we also determined the *BSS* while excluding the data from the summer with the longest heat wave (as detailed in Section 2.2.2). Importantly, this analysis was conducted separately for each grid point, acknowledging that the summer with the longest heat wave may vary from one grid point to another. Further, to demonstrate the impact of summer 2010 (with the long heat wave in Europe) on the probabilistic skill of the heat wave forecasts, we also determined the *BSS* while excluding data from summer 2010. Importantly, for this test, we excluded the 2010 data already when defining the thresholds for the heat wave days from the ERA5 and hindcast data, hence the thresholds were as in Fig. 1 in the middle column.

For each grid point and lead time, we determined whether the hindcasts were considered more skilful than the reference forecasts by assessing the *BSS* using a bootstrap resampling procedure. First, we calculated the *BSS* ~~n~~ 5000 times ~~(here n=5000)~~, each time sampling the original data with replacement (i.e., the data points could be selected multiple times). The *BSS* was required to be statistically significantly above zero for the hindcasts to be considered more skilful than the reference forecasts. To assess this issue, we calculated the statistical significance level, i.e., the p-value under the null hypothesis that the *BSS* is zero. The p-value is then the proportion of the bootstrap samples greater the zero. However, because the statistical

Formatted: Subscript

Formatted: Subscript

Formatted: Font: Italic

test on the map is repeated many times, small p-values are bound to occur by chance alone and the null hypothesis is rejected too often. Unadjusted p-values therefore overestimate the results (Wilks, 2016). We adjusted the p-values ~~using~~ following the false discovery rate (FDR) ~~method~~concept. T~~echnically, t~~he FDR ~~-~~ control~~s~~ling procedures limit ~~for~~ the expected proportion of false discoveries (hypotheses that should not have been rejected) among the rejected hypotheses. By setting this threshold $q$ to 0.1 (twice the conventional 0.05, as suggested by Wilks 2016), and using the Benjamini-Hochberg (B-H) procedure (e.g., Benjamini and Hochberg, 1995), we ensured that on average no more than 10% of the rejected null hypotheses are false discoveries. In the B-H procedure, we first ordered the p-values from the smallest to the largest. Then we rejected the null hypothesis if $p_i < q * i/m$, where i was the position and m was the number of p-values. In practice, we can use readily available p-value adjustment functions (such as *p.adjust* in R) that change p-values to the smallest threshold $q$ at which we would reject a particular null hypothesis.

## 3 Results

### 3.1 Reliability of probabilistic heat wave days forecasts

First, we examined the proportion of heat wave forecasts in each category of forecasted probabilities ($p<0.1$, $0.1\leq p<0.2$, …, $p\geq 0.9$) (sharpness diagram in Fig. 3a), and how often heat wave days occurred following a forecast in each category of forecasted probabilities (reliability diagram Fig. 3b). If all the forecasts were perfect, then in Figure 3(a) 90% of the forecasts would have $p=0$ and 10% would have $p=1$, and in Figure 3(b) there would be only two points [0,0] and [1,1] for each forecast week. However, for the first week, in Figure 3(a) roughly 80% of the forecasts belong to the lowest probability class and 5% to the highest one. As the lead time increases, both these portions decrease, while the share of forecasts with $0.1\leq p<0.9$ increases. The sharpness of forecasts drops as the lead time increases.

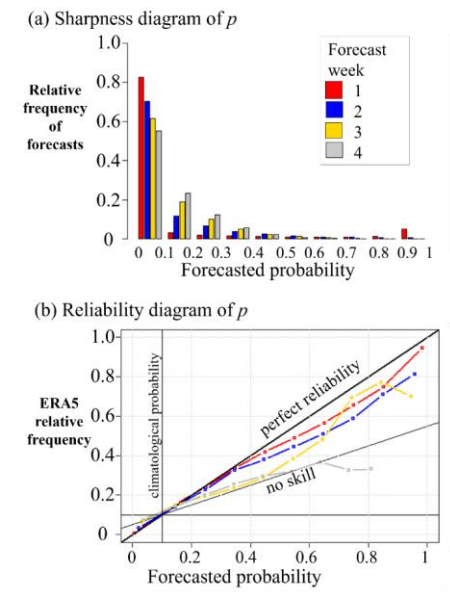



**Figure 3: (a) sharpness diagram and (b) reliability diagram of the 1-4 weeks probabilistic heat wave days forecasts, *p*, over Europe (all land grid points) in summers 2000-2019.**

In Figure 3(b) the forecasted probabilities are displayed on the x-axis and observed frequencies on the y-axis. In a perfectly calibrated forecast, the points on the reliability diagram would fall along a 45-degree diagonal line from the bottom left to the upper right corner. This line represents *perfect reliability*, where the forecasted probabilities equal the observed frequencies. The *climatological probability* line in the reliability diagram represents the expected frequency of ~~a~~ heat wave days (0.1) based on climatology. The points above the *no skill* line contribute positively to the BSS with climatology as reference. The points on the reliability diagram above the *perfect reliability* line indicate underforecasting, meaning that the forecasted probabilities are too low compared to the observed frequency. Conversely, the points on the reliability diagram below the perfect reliability line indicate overforecasting, meaning that the forecasted probabilities are too high compared to the observed frequency.

The reliability of the heat wave day forecasts was best for shorter lead times~~forecast weeks~~ and dropped with growing lead times (Figure 3b). During forecast weeks 1 and 2, the overall reliability of heat wave day forecasts across Europe was nearly flawless when $p<0.4$. Subsequently, for $p>0.4$, the forecasted probabilities tended to be slightly elevated compared to the

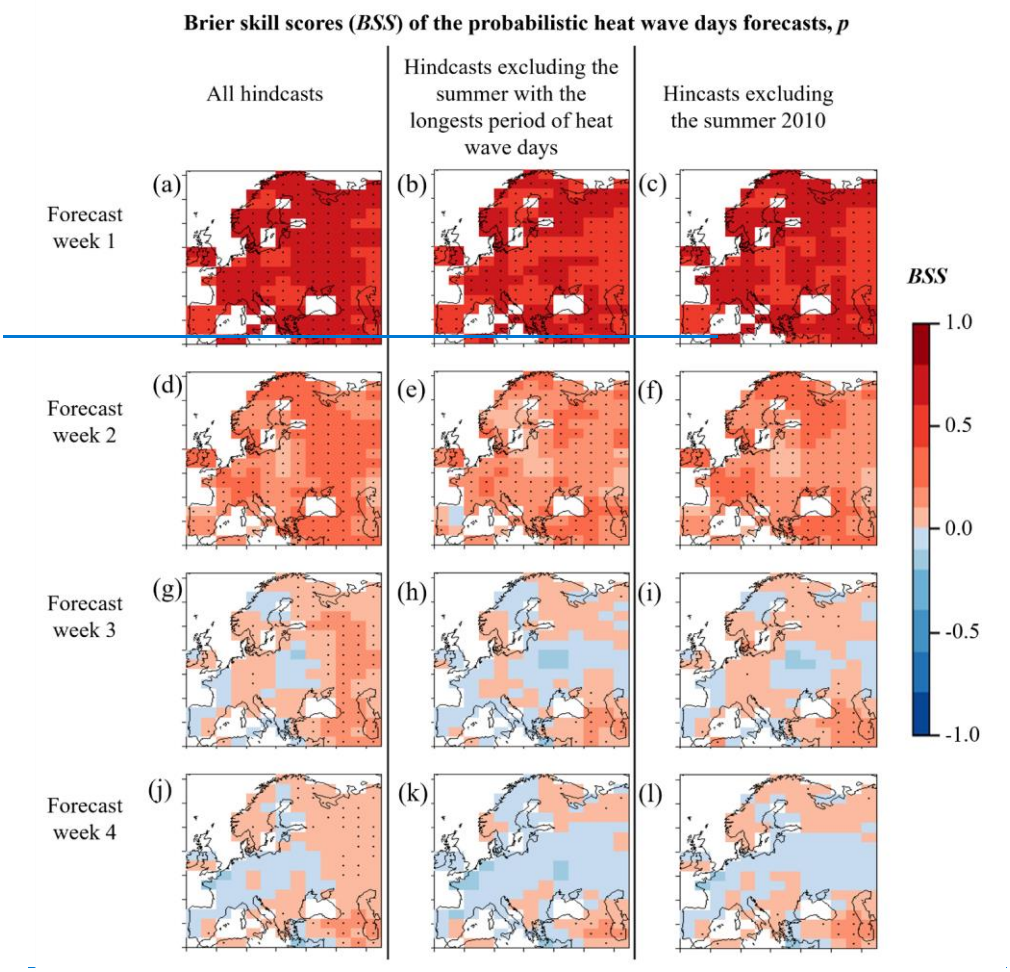
observed frequencies, suggesting a tendency toward overforecasting; however, it should be noted that for lead times of 2 weeks (and longer) there are far fewer samples in the higher probability bins, making these points considerably more uncertain.

### 3.2 Probabilistic forecast skill scores for heat wave days

Figure 3b depicted the average reliability of the heat wave days forecasts over the whole of Europe. Next, we will take a look at the forecast skill across different regions over Europe to find out how the accuracy varies in different regions. First, we assess the performance of ~~the~~ all the hindcasts of all summers from 2000 to 2019. Second, we examine hindcasts of summers from 2000 to 2019 excluding the hindcasts of the summer with the longest heat wave, and third, we focus on the hindcasts excluding summer 2010. In the first column of Figure 4, we present the *BSS* of all hindcasts of the summers 2000-2019. During the first forecast week, the forecasts of heat wave days in Europe demonstrate strong performance, with *BSS* values ranging between 0.5 and 0.8. Based on the adjusted p-values, these values of *BSS* are statistically significantly greater than 0 at every grid point. However, in later forecast weeks, the skill diminishes. In the second forecast week, BSS ranges from 0.1 to 0.4 in Europe, the forecasts still remaining better than the reference forecast in most grid points across the continent. The exceptions include certain grid points over the northern parts of the Iberian Peninsula, eastern central Europe, and northeast of the Caspian Sea. Moving to forecast weeks 3 and 4, BSS values in Europe range between -0.1 and 0.2, exhibiting statistical significance only in specific grid points across Eastern and South-Eastern Europe.

In the middle column in Figure 4, we illustrate the *BSS* for each grid point of all hindcasts excluding the summer with the longest heat wave (as defined in Section 2.2.2). The *BSS* excluding such a heat wave summer differs mostly only +/-0.05 from the *BSS* of all summers, except in Eastern Europe where the *BSS* is even 0.1 lower in forecast weeks 2-4. In more detail: in the first forecast week, the *BSS*s of the hindcasts excluding the summer with the longest heat wave are between 0.4 and 0.7 and in all grid points statistically significantly higher than 0, i.e., better than the reference forecast. In the second week, the *BSS* of the hindcasts excluding the summer with the longest period of heat wave days are between 0 and 0.4 and still statistically significantly higher than 0 in the majority of the grid points. In the third and the fourth week, however, the *BSS* is statistically significantly higher than 0 only in some grid points in southeastern parts of the map.

In ~~the~~ Figure 4, the last column shows the *BSS* of the hindcasts excluding the summer 2010. In some areas, leaving out 2010 seems to have less impact on the probabilistic skill of heat wave forecasts than leaving out, in each grid point, the summer with the longest heat wave (the middle column). For example, in Finland the skill remains for the third week, and ~~.~~ ~~A~~also the southeast parts of the study domain seem to remain with skill. These results suggest that the skill in forecasting heat waves decreased when excluding the longest period of heat wave days, whether it was the 2010 heat wave or a heat wave from another year.


Brier skill scores (BSS) of the probabilistic heat wave days forecasts, p
All hindcasts
Hindcasts excluding the summer with the longests period of heat wave days
Hincasts excluding the summer 2010
Forecast week 1
Forecast week 2
Forecast week 3
Forecast week 4
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
BSS
1.0
0.5
0.0
-0.5
-1.0

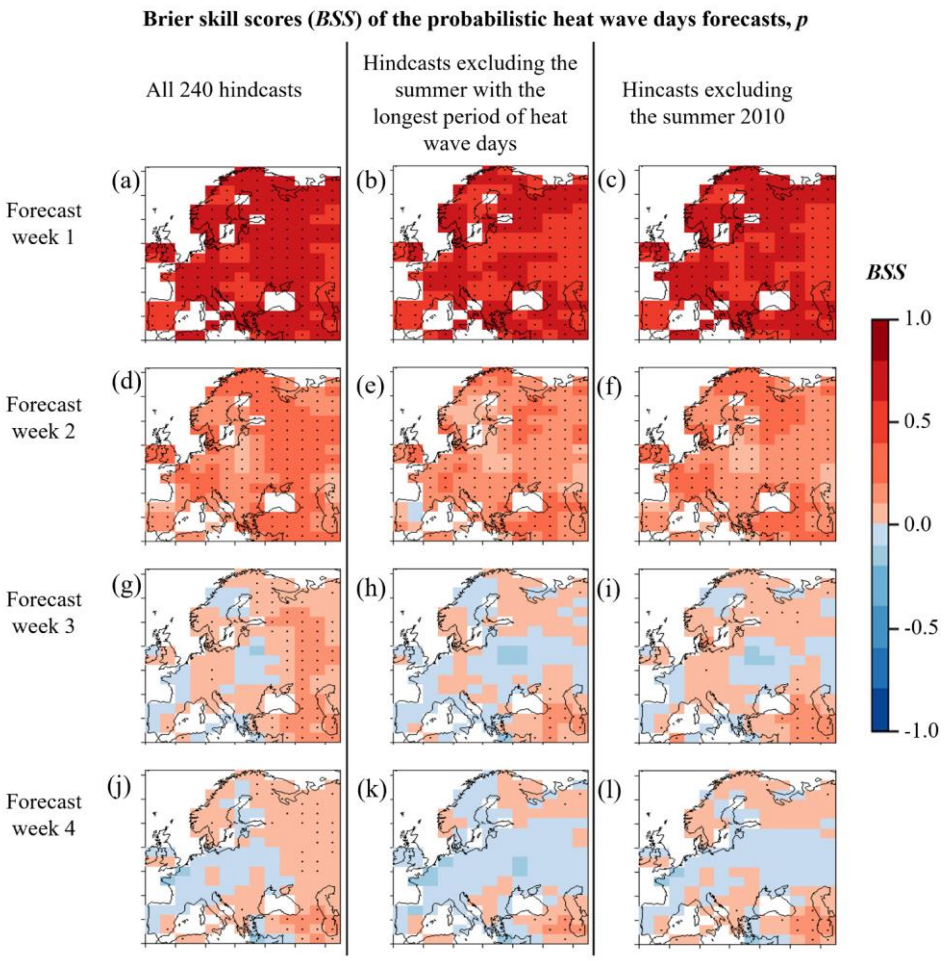



**Figure 4: Brier Skill Scores (*BSS*) of the probabilistic heat wave days forecasts, *p*, during all summers 2000-2019 (first column), in hindcasts excluding the summer with the longest period of heat wave days (middle column), and in hindcasts excluding the summer 2010 (last column). The statistical occurrence *p*=0.1 for heat wave days were used as the reference forecasts. The dotted areas show where *BSS* is greater than zero with the false discovery rate no more than 10%.**

### 3.3 Verification by probability ranges

In the reliability diagram (Fig. 3b) the ERA5-based temperature data is used only as either no hot day (0) or hot day (1). Next, we will examine how close or far the ERA5-based temperatures were from the threshold of a hot day across different levels of *p*, to assess the severity of the over- or under-forecasts. For this, we conducted verification of heat wave day forecasts based on forecasted probabilities falling within the ranges of here defined as low: $p < 0.33$, intermediate: $0.33 \leq p \leq 0.66$, and high: $p > 0.66$. In Figure 5, boxplots depict all the observed ERA5 temperatures (as percentiles) across different levels of *p*. The parts of the boxes above the 90th percentile (grey horizontal line) indicate heat wave days in the ERA5 temperature reanalysis. It is important to note that each box has a different amount of data marked as *n* above each box. Due to the different amounts of hindcast data in each forecast week, as depicted in Table 1, the total amount of data differs for each lead time. The category with the most forecasts is within the low ($p < 0.33$) range, which was also visible in Figure 3a.

If all the heat wave day forecasts were perfect, in Figure 5 the boxes

- for $p < 0.33$ would be totally below the grey line, i.e., heat wave days would occur in 0% of cases,
- for $0.33 \leq p \leq 0.66$ category would be empty, and
- for $p > 0.66$ would be totally above the grey line, i.e., heat wave days would occur in 100% of cases.

All in all, the forecast skill improves when more of the data points in $p < 0.33$ fall below the grey line, and those in $p > 0.66$ are above the grey line. At a glance, forecast week 1 (Fig. 5a) appears to have good skill, while forecast week 4 (Fig. 5d) shows relatively poor skill.

Furthe~~However~~r, ~~as we see~~ in Figure 5, in occasions the forecasted probability for heat wave days was low ($p < 0.33$), heat wave days occurred in 2% (lead time one week), 7% (lead time two weeks), 10% (lead time three~~four~~ weeks), or 11% (lead time four weeks) of cases. Moreover, in occasions the forecasted probability for heat wave days was intermediate ($0.33 \leq p \leq 0.66$), heat wave days occurred in 45% (lead time one week), 39% (lead time two weeks), 30% (lead time four weeks), or 28% (lead time four weeks) of cases. In occasions the forecasted probability for heat wave days was high ($p > 0.66$), heat wave days occurred in 86% (lead time one week), 68% (lead time two weeks), 67% (lead time four weeks), or 38% (lead time four weeks) of cases. Hence, higher probabilities ($p > 0.66$) show that a heat wave event is more likely, however for forecast weeks 3 and 4 the forecasting signal is not very strong due to the relatively low proportion of *n* (amount of data) in group $p > 0.66$. Additionally, $p < 0.33$ provides a good indication that a heat wave is unlikely. Based on the data, the lower the *p* (below 0.33), the less likely a heat wave is to occur, as, e.g., in occasions the $p < 0.1$ (no figure), heat wave days occurred only in 1% (lead time one week), 4% (lead time two weeks), 6% (lead time three weeks), or 8% (lead time four weeks) of cases.

It should be noted that Figure 5 also shows how often forecasts were followed by a heat wave or *near-heat wave* conditions (e.g., temperatures exceeding the 85th percentile) in the ERA5 dataset. For instance, in situations where $p > 0.66$, temperatures

**Formatted:** Normal, No bullets or numbering

**Formatted:** Font: Italic

**Formatted:** Font: Italic

**Formatted:** Font: Italic

**Formatted:** Font: Italic

**Formatted:** Font: Italic

surpassing the 85th percentile (rather than the 90th percentile) occurred even in 95% (lead time one week), 78% (lead time two weeks), 74% (lead time four weeks), or 44% (lead time four weeks) of cases.

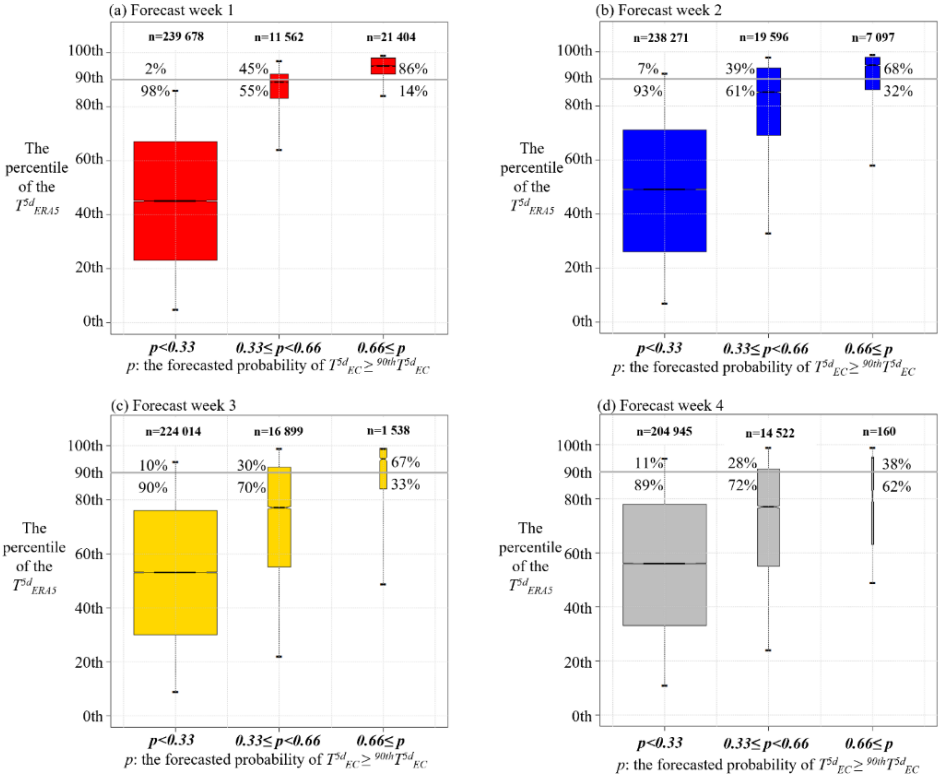



**Figure 5: Boxplots of the ERA5 5-day moving average temperature over Europe in each grid point across different levels of *p* (the forecasted probability of a heat wave day) with lead times of a) one week, b) two weeks, c) three weeks, and d) four weeks. The horizontal line dividing each box into two parts shows the median of the data; the ends of the box show the lower and upper quartiles; and the whiskers indicate the 5th and 95th percentiles of the ERA5 data in each group. The width of each box and the *n* written above each box indicate the number of observations in each group. The grey horizontal line indicates the 90th percentile, i.e., the threshold of a heat wave day, and the percentiles above (and below) the grey line depict the fraction~~amount~~ of observed heat wave days (and no-heat wave days) after the different levels of forecasted probability.**

### 3.4 Predicting the lifecycle of a heat wave

Next, we shall evaluate the capacity of the probabilistic heat wave day forecasts (*p*) to predict the life cycle of heat waves, taking into account the forecast initialization (date) relative~~relative time of forecast issuance and~~ to the onset of the heat wave

~~initiation~~. In Figure 6, *p* are shown for days categorized according to the correspondingsive ERA5 data as: "before the heat wave," "during the heat wave," and "after the heat wave", across the entire European region at each land-grid point. If there were no heat wave days during the entire summer at that grid point according to the ERA5 data, the temporal distance to the nearest heat wave day during all the heatless days of that summer were classified as "over 21 days before the heat wave." Dashed green boxes delineate forecasts where, at the time of issuance, a heat wave in that grid point is about to begin within a week. Solid green boxes indicate forecasts where, at the time of issuance, a heat wave is ongoing in that grid point. If the forecasts were perfectly aligned with reality, *p* should be zero in the categories "before the heat wave" and "after the heat wave," and in the category "during the heat wave," *p* should be 1 (i.e., 100%).

~~In heat wave day forecasts one week in advance (Figure 6a, Forecast week 1), considerably higher *p* for all lengths of heat waves and throughout the heat wave are evident compared to non-heat wave days. Additionally, there is some overestimation, particularly 1-2 days before or after the heat waves indicating slight inaccuracy in forecasting the exact day of the start and ending of the heat wave.~~

~~In heat wave day forecasts two weeks in advance (Figure 6b, Forecast week 2), forecasts show clearly higher *p* for days within the heat wave than outside. Especially as the heat wave extends to the second, third, fourth, fifth, or more weeks, forecasts show significantly higher probabilities. Additionally, there is some slight overestimation for heat wave days in the days just before and after the heat wave, indicating slight inaccuracy in forecasting the exact day of the start and ending of the heat wave.~~

In heat wave day forecasts both one week in advance (Figure 6a) and two weeks in advance (Figure 6b), the forecasts show clearly higher *p* for days within the heat wave than outside, especially for the forecasts which are in the green boxes indicating that the heat wave was just starting or already underway when these forecasts were issued. Additionally, there is some overestimation, particularly 1-2 days before or after the heat waves indicating slight inaccuracy in forecasting the exact day of the start and ending of the heat wave.

Formatted: Font: Italic

In heat wave day forecasts three weeks in advance (Figure 6c~~, Forecast week 3~~), higher *p* for days within the heat wave are still more apparent than for days outside the heat wave. Especially for the third, fourth, and fifth weeks of heat wave days, higher *p* are evident compared to non-heat wave days. These forecasts are in the green box ~~area~~, indicating that the heat wave wa~~i~~s just starting or already underway when the forecast is issued.

In heat wave day forecasts four weeks in advance (Figure 6d~~, Forecast week 4~~), there are only slightly higher *p* during the heat wave than before and after. Particularly, a small portion of the data where the fifth week of the heat wave (days 29..35) is in progress, shows higher *p*. These forecasts are in the green box, indicating an ongoing heat wave when the forecast was issued.

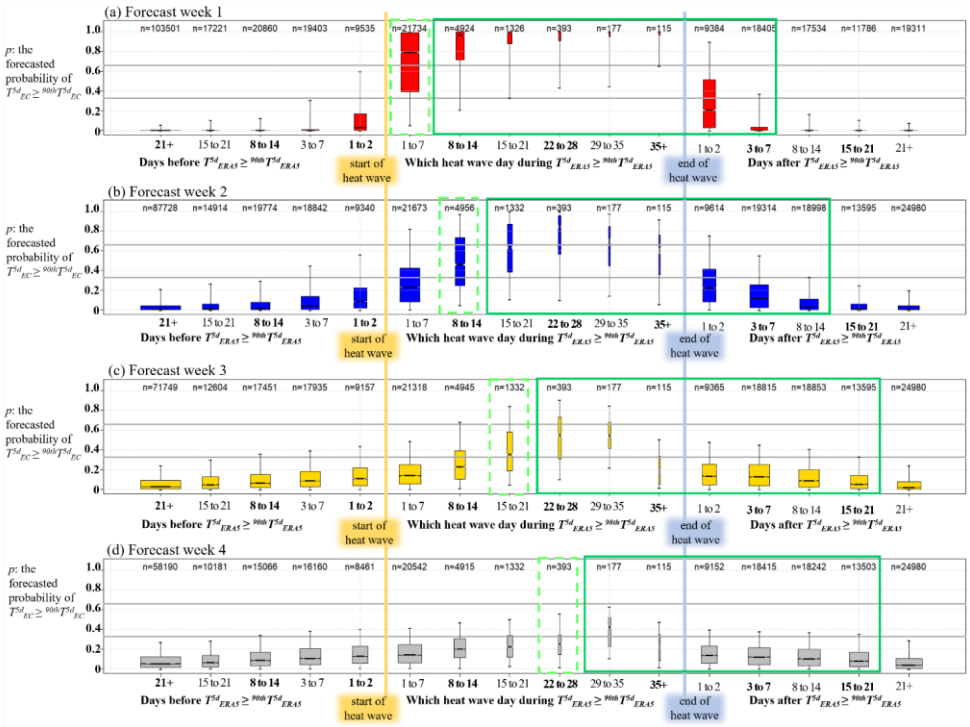



**Figure 6: The forecasted probabilities of heat wave days shown for days that (in ERA5) were 21 to 1 days before the heat wave, the 1st to 35th heat wave day during the heat wave, and 1 to 21 days after the heat wave with lead times of a) one week, b) two weeks, c) three weeks, and d) four weeks. Dashed green boxes indicate forecasts where, at the time of issuance, a heat wave in that grid point was about to begin within a week. Solid green boxes indicate forecasts where, at the time of issuance, a heat wave was already ongoing in that grid point. In the boxplots, a horizontal line box into two parts shows the median of the data; the ends of the box show the lower and upper quartiles; and the whiskers indicate the 5th and 95th percentiles of the data in each group. The boxplots include all forecast data across the European region at each land-grid point. The width of each box and the *n* written above each box indicate the number of observations in each group.**

We also plotted the heat wave life cycle figure without year 2010, here shown as Figure S1 in the Supplementary Material 17. Leaving out year 2010 removes most of the very longest heat waves, i.e., with lengths above 28 days. However, in the same

way as when including the year 2010 (Figure 6), in the forecast weeks 1-3 there is still a signal of enhanced accuracy in forecasting prolonged (here several weeks long) heat waves at the time that the heat wave had initiated prior to the forecast issuance. Thus, the differences remain negligible.

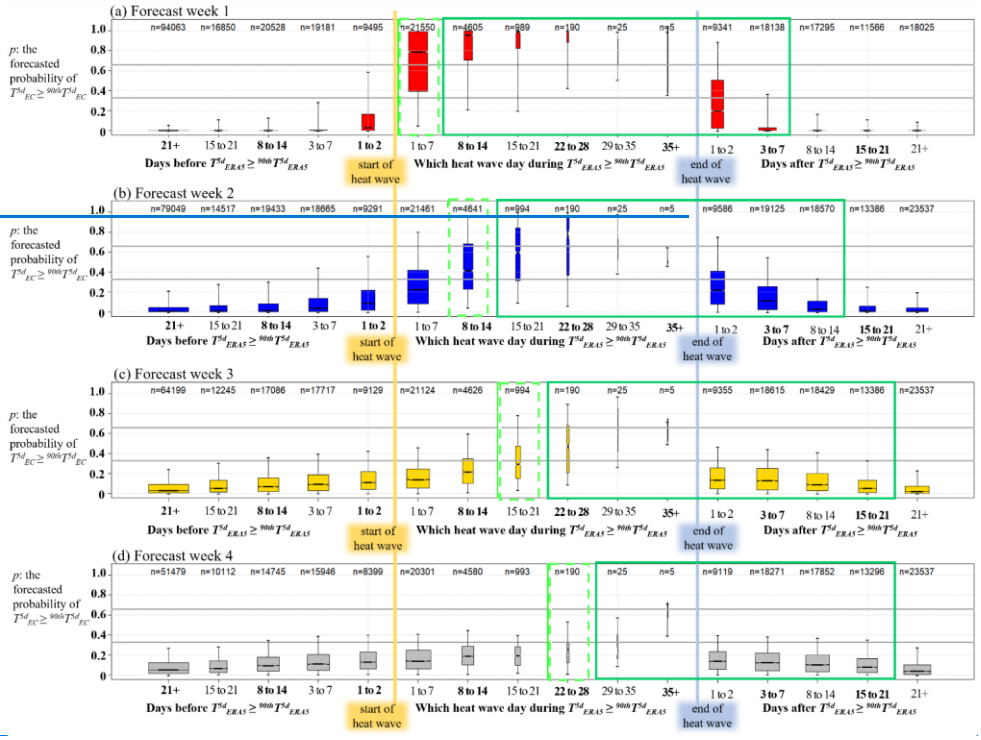



**Figure 7: Same as Figure 6 but excluding summer 2010.**

## 4 Discussion

### 4.1 The skill of the verified probabilistic heat wave forecasts

We examined the skill of hindcasts of the ECMWF in forecasting the probability of heat wave days over Europe 1 to 4 weeks ahead. The verification was done against ERA5 reanalysis. Heat wave days were here defined as days with the 5-day moving average temperature being above its summer 90th percentile. To address potential systematic bias resulting from forecast

~~model drift, we opted to use percentiles separately for ERA5 and the ECMWF's hindcasts.~~ The assessed hindcasts demonstrated varying levels of accuracy across different regions, and decreasing levels with increasing forecasting lead times, which is in line with many earlier studies, e.g.~~,~~ ~~Ferranti et al. (2015),~~ Wulff and Domeisen (2019), and Pyrina and Domeisen (2023). This outcome could be seen as expected, as we employed the same forecasting model and verification region as in these previous works. However, our method for determining the probability of heat wave days was novel, providing a fresh perspective that sets our study apart from earlier research using the same model and verification region.

We investigated the impact of the longest heatwaves on the forecast skill (*BSS*) in two ways: i) by excluding the summer with the longest heatwave observed at each grid point, and ii) by excluding the summer of 2010, which saw a prolonged and widespread heatwave in Europe. We found that the skill in forecasting heat waves decreased when excluding the longest period of heat wave days~~the best of the forecast skill seems to come from the longest period of heat wave days~~, whether it was the 2010 heat wave or a heat wave of some other year.

Figures 6 and S1~~7~~ present a novel way for evaluating the ability of probabilistic heat wave day forecasts to capture the life cycle of heat waves, taking into account the timing of forecast issuance relative to heat wave onset. This approach could be developed further by adding information about the spread of the ensemble to the figure, and it could be applied to the verification of other extended-range models' heat wave forecasts in future studies.

**4.2 Potential added value of probabilistic heat wave forecasts**

~~Heat waves cause severe health risks and have considerable impacts on public health, particularly if the high temperatures last for many days or weeks and occur over a wide area (Ballester et al. 2023, Robine et al. 2008). Prolonged hot periods may pose concurrent risks that challenge functioning of society, including health care system and other critical infrastructure. The negative health impacts of heat waves can be reduced by timely short-term measures to protect the vulnerable population groups (Martinez et al. 2019, Tooloo et al. 2013). The needed measures are preferably outlined in national or sub-national heath-health actions plans, as recommended by the World Health Organization (WHO) (Martinez et al. 2019, Matthies et al. 2008).~~

~~Heat-health action plans rely on early-warning systems to inform all stakeholders and citizens on the impending heat wave and to activate the public health interventions in time. As health effects of heat exposure occur quickly, at the same day or a few days lag (Baccini et al. 2008), it is imperative that the protection measures are implemented rapidly when a potentially dangerous heat wave is forecasted. However, organisation of the response measures requires coordination of actions between many stakeholders and distribution of workforce, equipment, and other resources, which take time.~~ Currently, most heat warning systems in Europe have lead~~-~~ times of only a few days (Casanueva et al. 2019). However, in this study the probabilistic heat wave days forecasts seem to have high potential in warning of heat risk in 1-2 weeks in advance, as for lead times for 1-

2 weeks, there is signal that lower probability (probabilities below 0.33) forecasts could be valuable for indicating periods when it is unlikely that a heat wave will occur. And the higher probability (probabilities above 0.66) forecasts could be valuable for indicating periods when a heat wave could occur. Further, the persistence of heat waves seems to have higher level of predictability up to 3 weeks, offering early warning services an indication of the potential duration of an ongoing heat wave.

To the knowledge of the authors, there has been no published research on ~~There has been no research on~~ how warning lead time contributes to the effectiveness of heat-health warning systems. However, considering the short lag between heat exposure and worsening of health conditions, extending warning lead times from the current level of few days is acknowledged to be valuable to public health, as prevention and emergency measures need to be in place and operational at the onset of a hazardous heat event (WHO, 2021). Organization of the measures, such as communication campaigns, establishing cooling centers, arrangements to protect vulnerable population groups, and ensuring adequate supply and distribution of workforce, equipment, and other resources, require time and would benefit from receiving early warnings 1–2 weeks ahead, particularly because heat waves often occur at times when organizations and services are already short-staffed due to summer holiday season. Longer lead time is especially important regarding exceptionally severe and prolonged hot periods, which challenge the functioning of ~~the~~ society on a wider scale and may require large-scale interagency and even transboundary response. The likelihood for these types of events can be expected to increase in Europe as climate change progresses.

**Formatted:** Normal
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt
**Formatted:** Font: (Default) Times New Roman, 10 pt

## 5 Conclusions

Our examination of ECMWF hindcasts for predicting heat wave days (periods where the local 5-day mean temperature exceeded the 90th percentile of the local summertime 5-day mean temperature distribution) of summers 2000-2019 across Europe, 1 to 4 weeks in advance, showed varying accuracy levels across forecast lead times and regions, aligning with previous research. The examined ECMWF's hindcasts showed:

- in the first forecasts week (1 to 7 days in advance): strong forecast skill in predicting heat wave days,
- in the second forecast week (8 to 14 days in advance): statistically significantly better skill than the reference forecast in most grid points over Europe,
- in the forecast weeks 3-4: statistically significantly better skill than the reference forecast only in some grid points across ~~Eastern and~~ South-Eastern Europe, and
- in the forecast weeks 1-3: enhanced accuracy in forecasting prolonged (here several weeks long) heat waves at the time that the heat wave had initiated prior to the forecast issuance.

These findings underscore the potential of these ECMWF's heat wave days forecasts to serve as early warnings for impending heat risks 1-2 weeks in advance. Notably, the higher-than-average predictability for intense and prolonged heat waves (at the

time they have already started), offers a potential to early warnings even at a 3-week lead time. However, it is crucial to highlight the known uncertainty in the 3-week lead time forecast. Building on these insights, future research could investigate at which stage of the heat wave development extended-range weather forecast models in general, not only the specific model system considered here, begin to predict heat wave~~its~~ occurrence, potentially enhancing early warning capabilities ~~further~~.

## Data availability

The ERF data of the ECMWF's IFS cycles 46r1 and 47r1 were retrieved from the ECMWF's MARS archive at https://apps.ecmwf.int/mars-catalogue/ (MARS, 2024). The ERA5 reanalysis data were retrieved from the Copernicus Climate Change Service Climate Data Store (CDS) at https://cds.climate.copernicus.eu/cdsapp#!/home (CDS, 2024).: date of access 11th April 2024. The data of Figs. 1–6 and S1~~7~~ are available at https://doi.org/10.57707/FMI-B2SHARE.372EC54BE8014B399AF3900DD253925A (Korhonen, 2024).

## Author contribution

NK led the writing of the paper and produced figures 1-6 and S1~~7~~. KJ: Conceptualization, Funding acquisition, Project administration, Writing – review & editing. All authors contributed to writing the manuscript.

## Competing interests

The authors declare that they have no conflict of interest.


## Acknowledgements

We acknowledge the ECMWF for the forecast data and all those involved in producing ERA5 data. We acknowledge Matti Kämäräinen from the Finnish Meteorological Institute for valuable comments during this research. We acknowledge our two anonymous referees, whose excellent comments helped us to improve the quality of the manuscript.

## Financial support

This study is part of the following projects: HEATCLIM (Heat and health in the changing climate, Grant Numbers 329304, 329305, 329306, and 329307) within the CLIHE (Climate change and health) program, and ACCC (Atmosphere and Climate Competence Center, Flagship Grant No. 337552) both funded by the Research Council of Finland.

**References**


Ahmed, H., Tamminen, LM., and Emanuelson, U.: Temperature, productivity, and heat tolerance: Evidence from Swedish dairy production, Climatic Change 175, 10, doi.org/10.1007/s10584-022-03461-5, 2022.

Añel, J.A., Fernández-González, M., Labandeira, X., López-Otero, X., and De la Torre, L.: Impact of Cold Waves and Heat Waves on the Energy Production Sector, Atmosphere, 8(11):209, doi.org/10.3390/atmos8110209, 2017.

Arsad, F.S., Hod, R., Ahmad, N., Ismail, R., Mohamed, N., Baharom, M., Osman, Y., Radi, M.F.M., and Tangang, F.: The Impact of Heatwaves on Mortality and Morbidity and the Associated Vulnerability Factors: A Systematic Review, Int. J. Environ. Res. Public Health., 19(23):16356, doi:10.3390/ijerph192316356, 2022.

Baccini, M., Biggeri, A., Accetta, G., Kosatsky, T., Katsouyanni, K., Analitis, A., Anderson, H.R., Bisanti, L., D'Ippoliti, D., Danova, J., Forsberg, B., Medina, S., Paldy, A., Rabczenko, D., Schindler, C., and Michelozzi, P.: Heat effects on mortality in 15 European cities, Epidemiology, 19(5):711–719, doi:10.1097/EDE.0b013e318176bfcd, 2008.

Ballester, J., Quijal-Zamorano, M., Méndez Turrubiates, R.F., Pegenaute, F., Herrmann, F.R., Robine, J.M., Basagaña, X., Tonne, C., Antó, J.M., and Achebak, H.: Heat-related mortality in Europe during the summer of 2022, Nat Med., 29(7):1857-1866. doi:10.1038/s41591-023-02419-z, 2023.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300, doi.org/10.1111/J.2517-6161.1995.TB02031.X, 1995.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3, 1950.

Casanueva, A., Burgstall, A., Kotlarski, S., Messeri, A., Morabito, M., Flouris, A. D., Nybo L, Spirig C, and Schwierz C.: Overview of Existing Heat-Health Warning Systems in Europe, Int. J. Environ. Res. Public Health, 16(15):2657, doi: 10.3390/ijerph16152657, 2019.

CDS: ERA5 reanalysis data, available at https://cds.climate.copernicus.eu/cdsapp#!/home, last access: 29 April 2024.

Clemen, R.T., 1996. Making Hard Decisions: an Introduction to Decision Analysis . Duxbury, 664pp.

Coumou, D., and Rahmstorf, S.: A decade of weather extremes, Nature Climate Change, 2(7), 491–496, doi.org/10.1038/nclimate1452, 2012.

Dunne, J. P., Stouffer, R. J., and John, J. G.: Reductions in labour capacity from heat stress under climate warming, Nature Climate Change, 3, 563–566, doi.org/10.1038/nclimate1827, 2013.

Ferranti, L., Corti, S., and Janousek, M.: Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector, Quart. J. Roy. Meteor. Soc., 141, 916–924, doi.org/10.1002/qj.2411, 2015.

Ferro, C. A. T., Richardson, D. S., and Weigel, A. P.: On the effect of ensemble size on the discrete and continuous ranked probability scores, Meteorological Applications, 15(1), https://doi.org/10.1002/met.45, 2008.

Frame, T. H. A., Methven, J., Gray, S. L., and Ambaum, M. H. P.: Flow-dependent predictability of the North Atlantic jet, Geophys. Res. Lett., 40, 2411–2416, doi.org/10.1002/grl.50454, 2013.

Gasparrini, A., Masselot, P., Scortichini, M., Schneider, R., Mistry, M. N., Sera, F., Macintyre, H. L., Phalkey, R., and Vicedo-Cabrera, A. M.: Small-area assessment of temperature-related mortality risks in England and Wales: a case time series analysis, The Lancet Planetary Health, 6, e557-64, doi.org/10.1016/S2542-5196(22)00138-3, 2022.

Guo, Y., Gasparrini, A., Armstrong, B.G., Tawatsupa, B., Tobias, A., Lavigne, E., de Sousa Zanotti Stagliorio Coelho, M., Pan, X., Kim, H., Hashizume, M., Honda, Y., Guo, Y.L., Wu, C., Zanobetti, A., Schwartz, J.D., Bell, M.L., Scortichini, M., Michelozzi, P., Punnasiri, K., Li, S., Tian, L., Osorio Garcia, S.D., Seposo, X., Overcenco, A., Zeka, A., Goodman, P., Dang, T.N., Van, D.D., Mayvaneh, F., Saldiva, P.H.N., Williams, G., and Tong, S.: Heat wave and mortality: A multicountry, multicommunity study, Environ. Health Perspect, 125(8):087006, doi.org/10.1289/EHP1026, 2017.

Guo, Y., Gasparrini, A., Li, S., Sera, F., Vicedo-Cabrera, A.M., de Sousa Zanotti Stagliorio Coelho, M., Saldiva, P.H.N., Lavigne, E., Tawatsupa, B., Punnasiri, K., Overcenco, A., Correa, P.M., Ortega, N.V., Kan, H., Osorio, S., Jaakkola, J.J.K., Ryti, N.R.I., Goodman, P.G., Zeka, A., Michelozzi, P., Scortichini, M., Hashizume, M., Honda, Y., Seposo, X., Kim, H., Tobias, A., Iniguez, C., Forsberg, B., Astrom, D.O., Guo, Y.L., Chen, B., Zanobetti, A., Schwartz, J., Dang, T.N., Van, D.D., Bell, M.L., Armstrong, B., Ebi, K.L., and Tong, S.: Quantifying excess deaths related to heatwaves under climate change scenarios: A multi-country time series modelling study. PLoS Med., 15(7), e1002629, doi.org/10.1371/journal.pmed.1002629, 2018.

Hatvani-Kovacs, G., Belusko, M., Pockett, J., and Boland, J.: Assessment of Heatwave Impacts, Procedia Engineering, 169, 316–323, doi.org/10.1016/j.proeng.2016.10.039, 2016.

Heino, M., Kinnunen, P., Anderson, W., Ray, D. K., Puma, M. J., Varis, O., Siebert, S., and Kummu, M.: Increased probability of hot and dry weather extremes during the growing season threatens global crop yields, Sci. Rep. 13, 3583, doi:10.1038/s41598-023-29378-2, 2023.

Hersbach, H., Bell, B., Berrisford, P., et al.: The ERA5 Global Reanalysis, Quart. J. Roy. Meteor. Soc., 146, 1999-2049, doi:10.1002/qj.3803, 2020.

IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, doi:10.1017/9781009157896, 2021.

Kim, S., Sinclair, V. A., Räisänen, J., and Ruuhela, R.: Heat waves in Finland: present and projected summertime extreme temperatures and their associated circulation patterns, International Journal of Climatology, 38 (3), 1393–1408, doi.org/10.1002/joc.5253, 2018.

Kivimäki, M., Batty, G. D., Pentti, J., Suomi, J., Nyberg, S. T. Merikanto, J., Nordling, K., Ervasti, J., Suominen, S. B., Partanen, A.-I., Stenholm, S., Käyhkö, J., and Vahtera, J.: Climate Change, Summer Temperature, and Heat-Related Mortality in Finland: Multicohort Study with Projections for a Sustainable vs. Fossil-Fueled Future to 2050, Environmental Health Perspectives, 131 (12), 127020, doi.org/10.1289/EHP12080, 2023.

Kjellstrom, T., Kovats, R. S., Lloyd, S. J., Holt, T., and Tol, R. S. J.: The direct impact of climate change on regional labor productivity, Archives of Environmental & Occupational Health, 64(4), 217–227, doi.org/10.1080/19338240903352776, 2009.

Kollanus, V., Tiittanen, P., and Lanki T.: Mortality risk related to heatwaves in Finland —Factors affecting vulnerability, Environmental Research, 201, 111503, doi:10.1016/j.envres.2021.111503, 2021.

Korhonen, N.: Files containing data in Figures 1-7 in the manuscript Korhonen N. et al. "The probabilistic skill of Extended-Range Heat wave forecasts over Europe", Finnish Meteorological Institute, available at: https://doi.org/10.57707/FMI-B2SHARE.372EC54BE8014B399AF3900DD253925A, last access: 08 October 2024.

Kotharkar, R. and Ghosh, A.: Progress in extreme heat management and warning systems: A systematic review of heat-health action plans (1995–2020), Sustain. Cities Soc., 76, 103487, doi:10.1016/j.scs.2021.103487, 2022.

Lopez, A., & Haines, S. (2017). Exploring the Usability of Probabilistic Weather Forecasts for Water Resources Decision-Making in the United Kingdom. Weather, Climate, and Society, 9(4), 701–715. https://doi.org/10.1175/WCAS-D-16-0072.1

Manzanas, R.: Assessment of model drifts in seasonal forecasting: sensitivity to ensemble size and implications for bias correction, Journal of Advances in Modeling Earth Systems, 12, e2019MS001751, 2020.

MARS: ERF data of the European Centre for Medium-Range Weather Forecasts' Integrated Forecasting System cycles 46r1 and 47r1, available at: https://apps.ecmwf.int/mars-catalogue/, last access: 29 February 2024.

Martinez, G. S., Kendrovski, V., Salazar, M. A., de'Donato, F., and Boeckmann, M.: Heat-health action planning in the WHO European Region: Status and policy implications, Environ. Res., 214 (Pt1), 113709, doi:10.1016/j.envres.2022.113709, 2022.

Martinez, G. S., Linares, C., Ayuso, A., Kendrovski, V., Boeckmann, M., and Diaz, J.: Heat-health action plans in Europe: Challenges ahead and how to tackle them, Environ. Res., 176, 108548, doi:10.1016/j.envres.2019.108548, 2019.

Matthies, F., Bickler, G., Marin, N.C., and Hales, S. (eds.).: Heat-health action plans: guidance, World Health Organization, Copenhagen, Denmark, 2008.

Morignat, E., Perrin, J.B., Gay, E., Vinard, J.L., Calavas, D., and Hénaux, V.: Assessment of the impact of the 2003 and 2006 heat waves on cattle mortality in France, PLoS One, 9 (3), e93176, doi:10.1371/journal.pone.0093176, 2014.

Mulholland, E. and Feyen, L.: Increased risk of extreme heat to European roads and railways with global warming, Clim. Risk. Manag., 34, 100365, doi:10.1016/j.crm.2021.100365, 2021.

Murphy, A. H. (1977). The Value of Climatological, Categorical and Probabilistic Forecasts in the Cost-Loss Ratio Situation. Monthly Weather Review, 105(7), 803–816. https://doi.org/10.1175/1520-0493(1977)105<0803:TVOCCA>2.0.CO;2

Murphy, A. H. (1998). The Early History of Probability Forecasts: Some Extensions and Clarifications. Weather and Forecasting, 13(1), 5–15. https://doi.org/10.1175/1520-0434(1998)013<0005:TEHOPF>2.0.CO;2

Orlov, A., Sillmann, J., Aaheim, A., and Aunan, K.: Economic Losses of Heat-Induced Reductions in Outdoor Worker Productivity: a Case Study of Europe, Economics of Disasters and Climate Change, 3, 191–211, https://doi.org/10.1007/s41885-019-00044-0, 2019.

Prodhomme, C., Materia, S., Ardilouze, C. White, R. H., Batté, L., Guemas, V., Fragkoulidis, G., and García-Serrano, J.: Seasonal prediction of European summer heatwaves, Clim Dyn. 58, 2149-2166, doi.org/10.1007/s00382-021-05828-3, 2021.

Pyrina, M. and Domeisen, D. I. V.: Subseasonal predictability of onset, duration, and intensity of European heat extremes, Quart. J. Roy. Meteor. Soc., 149(750), 84–101, doi.org/10.1002/qj.4394, 2023.

Ramos, M. H., van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? Hydrology and Earth System Sciences, 17(6), 2219–2232. https://doi.org/10.5194/HESS-17-2219-2013

Richardson, D. S.: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size, Quart. J. Roy. Meteor. Soc., 127(577), 2473-2489, https://doi.org/10.1256/smsqj.57714, 2001.

Robine, J. M., Cheung, S. L. K., Le Roy, S., Van Oyen, H., Griffiths, C., Michel, J. P., and Herrmann, F. R.: Death toll exceeded 70,000 in Europe during the summer of 2003, Comptes Rendus Biologies, 331(2), 171–178, doi.org/10.1016/J.CRVI.2007.12.001, 2008.

Rossiello, M. R. and Szema, A.: Health effects of climate change-induced wildfires and heatwaves, Cureus, 11(5), e4771, doi:10.7759/cureus.4771, 2019.

Ruffault, J., Curt, T., Moron, V., Trigo, R. M., Mouillot, F., Koutsias, N., Pimont, F., Martin-StPaul, N., Barbero, R., Dupuy, JL., Russo, A. and Belhadj-Khedher, C.: Increased likelihood of heat-induced large wildfires in the Mediterranean Basin, Sci. Rep. 10, 13790, doi.org/10.1038/s41598-020-70069-z, 2020.

Ruosteenoja, K. and Jylhä K.: Average and extreme heatwaves in Europe at 0.5–2.0°C global warming levels in CMIP6 model simulations, Climate Dynamics, 61, 4259–4281, doi.org/10.1007/s00382-023-06798-4, 2023.

Russo, S., Dosio, A., Graversen, R. G., Sillmann, J., Carrao, H., Dunbar, M. B., Singleton, A., Montagna, P., Barbola, P., and Vogt, J. V.: Magnitude of extreme heat waves in present climate and their projection in a warming world, Journal of Geophysical Research: Atmospheres, 119, 500–512, doi.org/10.1002/2014JD022098, 2014.

Russo, S., Sillmann, J., and Fischer, E.: Top ten European heatwaves since 1950 and their occurrence in the coming decades, Environ. Res. Lett., 10(12), 124003, doi:10.1088/1748-9326/10/12/124003, 2015.

Ruuhela, R., Votsis, A., Kukkonen, J., Jylhä, K., Kankaanpää, S., and Perrels, A.: Temperature-Related Mortality in Helsinki Compared to Its Surrounding Region Over Two Decades, with Special Emphasis on Intensive Heatwaves, Atmosphere, 12, 46, doi.org/10.3390/atmos12010046, 2021.

Trenberth, K. E. and Fasullo, J. T.: Climate extremes and climate change: The Russian heat wave and other climate extremes of 2010, J. Geophys. Res., 117, D17103, doi:10.1029/2012JD018020, 2012.

van Vliet, M.T.H.: Complex interplay of water quality and water use affects water scarcity under droughts and heatwaves, Nat. Water, 1, 902–904, doi:10.1038/s44221-023-00158-6, 2023.

Velashjerdi Farahani, A., Jokisalo, J., Korhonen, N., Jylhä, K., Ruosteenoja, K., and Kosonen, R.: Overheating Risk and Energy Demand of Nordic Old and New Apartment Buildings during Average and Extreme Weather Conditions under a Changing Climate, Appl. Sci., 11, 3972, doi.org/10.3390/app11093972, 2021.

Velashjerdi Farahani, A., Kravchenko, I., Jokisalo, J., Korhonen, N., Jylhä, K., and Kosonen, R.: Overheating assessment for apartments during average and hot summers in the Nordic climate, Building Research & Information, doi:10.1080/09613218.2023.2253338, 2023.

Velashjerdi Farahani, A.V., Jokisalo, J., Korhonen, N., Jylhä, K., and Kosonen, R.: Simulation analysis of Finnish residential buildings' resilience to hot summers under a changing climate, Journal of Building Engineering, 82, 108348, doi.org/10.1016/j.jobe.2023.108348, 2024a.

Velashjerdi Farahani, A.V., Jokisalo, J., Korhonen, N., Jylhä, K., and Kosonen, R.: Hot summers in Nordic apartments: Exploring the correlation between outdoor weather conditions and indoor temperature, Buildings, 14, 1053, doi.org/10.3390/buildings14041053, 2024b.

Vitart F.: Evolution of ECMWF sub-seasonal forecast skill scores, Q. J. R. Meteorol. Soc., 140, 1889–1899, doi: 10.1002/qj.2256, 2014.

Vitart, F. and Robertson, A.W.: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events, npj Clim. Atmos. Sci., 1, 3, https://doi.org/10.1038/s41612-018-0013-0, 2018.

Vogel, E., Donat, M. G., Alexander, L. V., Meinshausen M., Ray D. K., Karoly, D., Meinshausen N., and Frieler, K.: The effects of climate extremes on global agricultural yields, Environ. Res. Lett., 14, 054010, doi:10.1088/1748-9326/ab154b, 2019.

Vogel, M.M., Zscheischler, J., Fischer, E.M., and Seneviratne, S.I.: Development of future heatwaves for different hazard thresholds, J. Geophys. Res. Atmos., 125(9), e2019JD032070, doi.org/10.1029/2019JD032070, 2020.

WHO (World Health Organization). Heat and health in the WHO European Region: updated evidence for effective prevention. Copenhagen: WHO Regional Office for Europe, 2021.

Wilks, D. S.: "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, Bulletin of the American Meteorological Society, 97(12), 2263–2273, doi.org/10.1175/BAMS-D-15-00267.1, 2016.

Wulff, C. O., and Domeisen, D. I.: Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers, Geophys. Res. Lett., 46(20), 11520–11529, doi.org/10.1029/2019GL084314, 2019.

## Supplementary Material 1

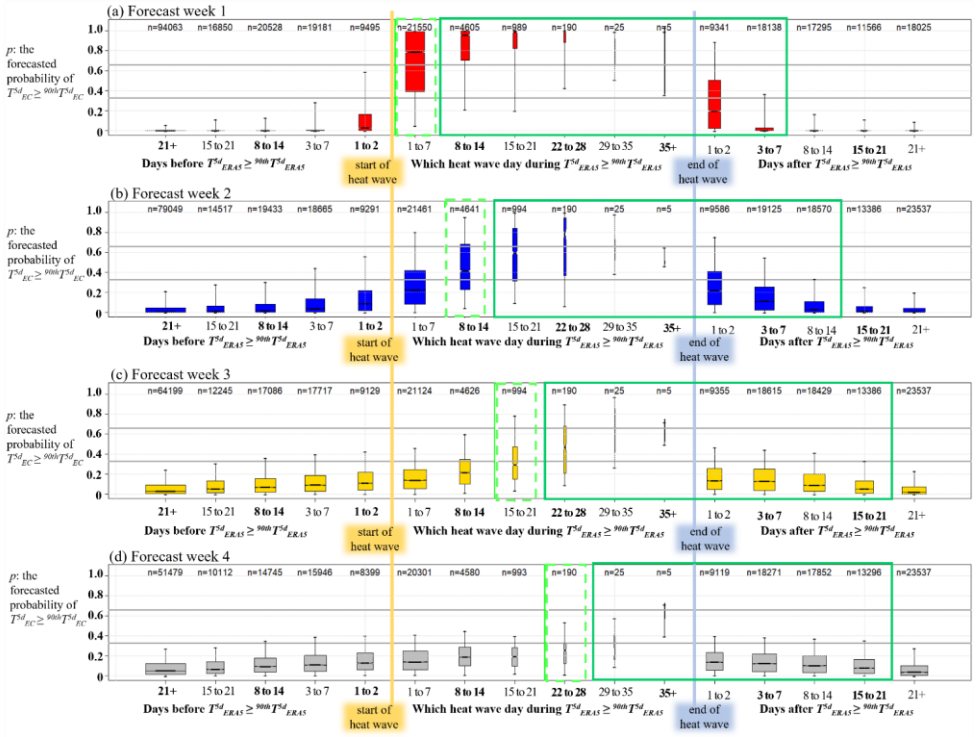



Figure S17: Same as Figure 6 but excluding summer 2010.