Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Office of Meteorology and Climatology MeteoSwiss

MeteoSwiss

Technical Report MeteoSwiss No. 281

Operational setup and skill analysis of a sub-seasonal forecasting system for detecting heat stress

Pascal-Andreas Noti, Christoph Spirig, Ana Casanueva, Jonas Bhend and Mark Liniger

ISSN: 2296-0058

Technical Report MeteoSwiss No. 281

Operational setup and skill analysis of a sub-seasonal forecasting system for de-tecting heat stress

Pascal-Andreas Noti, Christoph Spirig, Ana Casanueva, Jonas Bhend and Mark Liniger

Recommended citation:

Noti P.-A.,Ch. Spirig, A. Casanueva, J. Bhend and M. A. Liniger: 2017, Operational setup and skill analysis of a sub-seasonal forecasting system for detecting heat stress. *Technical Report MeteoSwiss,* **281,** 56 pp.

Editor:

Federal Office of Meteorology and Climatology, MeteoSwiss, © 2017

MeteoSwiss
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

Abstract

Heat waves and heat stress conditions can have dramatic impacts on human health or agriculture but also negatively affect workers' productivity in various industry sectors. In the framework of the Horizon2020 project HEAT-SHIELD, MeteoSwiss has set up an operationally running forecast system for detecting heat stress risk situations. The forecasting system consists of site-specific post-processing of the extended range forecasts of the European Center for Medium Range Weather Forecasts (ECMWF) based on an extensive set of weather stations all over Europe. Air temperature, dew point temperature, wind and radiation forecasts from ECMWF have been downscaled and bias-corrected by a quantile mapping technique using the re-forecasts (hindcasts) over the past 20 years and corresponding observation data. From these bias-corrected variables the wet-bulb globe temperature (WBGT) in the shade (based on air and dew point temperatures) and in the sun (including also wind speed and solar radiation) and probabilities of exceeding relevant WBGT thresholds have been computed. Finally, these WBGT forecasts have been verified with a set of skill scores using the hindcasts and observations of the summers 1996 - 2015. The verification showed that forecasts of summer WBGT are better than naïve predictions based on climatology for lead times up to two to three weeks.

Zusammenfassung

Hitzestress kann schwerwiegende Schäden für die menschliche Gesundheit, für die Agrarwirtschaft, Tourismus, Transportindustrie und für die Produktionsleistung von Handwerkern und Fabriken verursachen. Im Rahmen des Horizon2020 Projektes HEAT-SHIELD hat MeteoSchweiz einen Prototypen eines Vorhersagesystems für Hitzewellen aufgebaut. Die europaweiten Vorhersagen sollen frühzeitig vor den Gefahren des Hitzestresses warnen. Der Hitzestress wird mit einem Hitzeindikator bemessen, welcher aus Temperatur, Taupunkttemperatur, Strahlung und Windgeschwindigkeit berechnet wird. Die Monatsvorhersagen dieser Parameter vom Europäischen Zentrum für mittelfristige Wettervorhersagen (EZMSW) werden von MeteoSchweiz auf Stationsniveau herunterskaliert und anhand langjähriger Beobachtungsdaten statistisch korrigiert. Aus diesen nachbearbeiteten Vorhersagen werden für 1800 Stationen in Europa Ensemble-Vorhersagen des Hitzeindikators erstellt. Daraus lassen sich Wahrscheinlichkeiten für die Überschreitung kritischer Grenzen von Hitzestress für die kommenden vier Wochen ableiten. Mit Nachhersagen der vergangenen 20 Jahre wurde das Vorhersagesystem anhand verschiedener Gütemasse verifiziert. Die Verifikation zeigt, dass gute Prognosen für einen Zeitraum von zwei bis drei Wochen erstellt werden können.

Contents

1 Introduction

The future is like a corridor into which we can see only by the light coming from behind.
Edward Weyer Jr.

The ongoing climate change proceeds at a fast pace and affects nature and society negatively. More frequent and intense heat stress periods may influence biodiversity and various forms of life, leading to extinctions of many species (Kingsolver et al., 2013); reduction of crop production in temperate and sub-tropical agricultural areas (Teixeira et al., 2013) and work productivity especially in tropical and subtropical areas (Lundgren et al., 2013).

HEAT-SHIELD (https://www.heat-shield.eu/) is a Horizon 2020 research project and aims to protect workers of heat stress and sustain working productivity despite increased heat risk as a consequence of climate change. The project focuses on five strategic European industries: manufacturing, construction, transportation, tourism and agriculture. The assessment and choice of relevant heat stress indices and their prediction on short- and longer timescales are one of the central tasks in HEAT-SHIELD. A major component of this task is the development of a prototype for an operationally running forecast system to predict environmental heat strain, which is presented in this report.

The output of the forecasts should be a measure describing the heat stress on workers. There are more than 170 thermal indices suggested and described in literature (de Freitas and Grigorieva, 2015). Partners of the HEAT-SHIELD project specialised in human physiology have decided to use the Wet Bulb Globe Temperature (WBGT hereafter) index. This index allows to estimate heat stress on human bodies both for shaded and sunny conditions and it can be further interpreted by means of ISO standards.

Numerical models based on physical laws simulate feasible pathways development of future weather and climate. The use of weather and climate forecasting is subject to uncertainties from the following major sources (Weigel et al., 2008a, 2008b): On the one hand, the model initialisation is affected by measurement errors, incompleteness in the observation data and imperfect data assimilation. On the other hand, models are imperfect as a result from parametrization of physical processes, incomplete boundary conditions or unresolved scales (Buizza et al., 2005; Schwierz et al., 2006; Weigel et al. 2008b). Ensemble model forecasting allows to address errors in the model initialisation and uncertainties in parametrization (Buizza et al., 1999; Weigel et al., 2008b; Williams et al., 2014) and ideally can represent the propagation of such errors with increasing forecast lead times. Nevertheless, ensemble predictions may suffer from various shortcomings: There exist systematic model biases; the spread of the ensembles might become too wide or narrow; forecasts may lack skill meaning that the accuracy remains low. In these ways the forecast is not more valuable than naïve climatology (Wilks

and Hamill, 2007; Lerch and Thorarinsdottier, 2003; Schepen and Wang, 2014; Baran and Lerch, 2015; Khajehei and Moradkhani, 2017; Zhao et al., 2017). Model deficiencies are traditionally treated by a post-processing of the ensemble forecasts in order to achieve more accurate and reliable forecasts.

This report describes the post-processing applied to the ECMWF extended range forecast for producing WBGT forecasts. It is followed by a verification analysis of the WBGT forecast skill in summertime. Verification is performed by computing various skill scores from the forecasts and observations during summers 1996-2015. Finally the operational implementation of the forecast prototype is presented. The primary forecast output is an ensemble prediction of WBGT, from which probabilities of exceeding user relevant thresholds can be derived. As a first realization, probabilities of exceeding two WBGT levels are provided to a project partner, the Department of Agrifood Production and Environmental Sciences of the University of Florence. The delivered predictions will be further processed, warning level assigned and uploaded on a web-based platform, which is the major task for a later work package (WP5) of HEAT-SHIELD. The prototype presented in this report will be the basis of the final heat stress forecast system. Furthermore, the current setup will provide a basis for future applications of long range forecasts and outlooks of MeteoSwiss.

2 Data and Methods

This chapter describes the characteristics of the forecast, re-forecast and observational data and presents the methods for downscaling and bias correction. Then we show the details of the computation of the heat stress index and follow up with a description of the verification metrics.

2.1 Extended range forecasts from ECMWF

The European Centre for medium range weather forecasts (ECMWF) runs the Integrated Forecasting System (IFS hereafter) for generating forecasts over timescales from days to several months (https://www.ecmwf.int/). For monthly forecasts, runs of the medium range ensemble prediction system (IFS-ENS) are extended to six weeks, albeit at a coarser spatial resolution than the daily IFS-ENS runs. The extended runs are provided twice a week, initialized on Mondays and Thursdays, and are referred to as ECMWF extended range predictions (ENS-EXT).

ECMWF updates the operational forecast model about twice a year, the most recent versions are model cycles 43R1 (launched in November 2016) and cycle 43R3 (operational since mid-July 2017). The current ENS-EXT has a horizontal resolution of 0.2° x 0.2° (~18 km) for the first 15 days and 0.4° x 0.4° (~ 36 km) from days 16 to 46. Vertically the ENS-EXT consists of 91 layers with finest in geometric height in the planetary boundary layer and coarsest near the model top. The forecast ensemble consists of 51 members, with the control being the member with initial conditions corresponding to the best estimate of the operational analysis. The other 50 members are slightly perturbed in terms of initial conditions and by including stochastic physics. The atmospheric initial conditions of the real-time forecasts come from the operational analysis. The soil initial conditions are derived from an offline soil reanalysis.

In addition to the operational forecasts, ECMWF provides 20 years of re-forecasts (hindcasts) for calibration and verification purposes (Vitart et al. 2014). Model drift leads to significant changes towards the model's own climatology after the 10th day of model simulation. Thus hindcasts are essential to gain the information about the model drift and to enable feasible long- term predictions. The hindcasts are produced with the same model version as the operational forecast and with almost identical configuration. Only the number of members is reduced to 11 members (one unperturbed and 10 perturbed members) and initial conditions are from ERA-Interim (Dee et al., 2011) rather than the operational analysis. The hindcast ensembles are produced along the operational forecasts. Therefore, the hindcasts of the operational model version are available from the release up to the current date. The extended range forecasts are known to have some predictability in the time range from 10 to 30 days (Vitart 2014; Vitart et al. 2016). This time period is short enough to have still a memory of the initial state of the atmosphere while it is long enough to let the evolving state of the

oceans gain increasing influence on the atmospheric circulation. The extend range forecasts contain an ocean-atmospheric coupling and ocean model. Such configurations allow to model large scale pattern appropriately and lead to higher predictability. The Madden-Julian Oscillation is seen as the major source of predictability over Europe for forecasts beyond the medium range (Ferranti et al. 1990).

For analysing the forecast performance for the summer season, hindcasts (from May to mid-August 2016) of the model version of summer 2016 were used (cycle 41r2, Haiden et al. 2016). This prior version was identical to the current version in terms of resolution and exhibited similar or only slightly lower skill than the latest version according to ECMWF verification analyses (http://www.ecmwf.int/en/forecasts/documentation-and-support/evolution-ifs/cycles/cycle-43r1-summary-changes-latest).

2.2 Observational data

Non-standard parameters such as dew point temperature or radiation are needed to quantify heat stress. Thus, the following four datasets were combined in order to get a ground based observation dataset covering whole Europe. First, the European Climate Assessment and Dataset project (ECA&D; http://www.ecad.eu/; Klein-Tank et al., 2002) was considered as the major source for the observational dataset. The amount of available stations was dramatically reduced for parameters like dew point temperature and wind speed, therefore other station-based products were examined. Second, the Global Surface Summary of the Day (GSOD) dataset from the National Oceanic Atmospheric Administration (NOAA) was selected to improve the spatial coverage of the European continent (https://data.noaa.gov/dataset/global-surface-summary-of-the-day-gsod).Third, 65 stations of the Swiss national observing system SwissMetNet were added to the final dataset. The stations from the three datasets with more than 20% of missing values were removed producing a final observational dataset with measurements from 1799 stations across Europe over the past 20 years. Since ground observations of solar radiation are very sparse, daily measurements in the visible range of the Meteosat First and Second Generation Satellites were used (Posselt et al. 2012; 2014). The satellite-derived data covers the time period from 1983 to 2015 with an horizontal resolution of 0.05°. The closest grid box to each of the 1799 stations was used to approximate the solar radiation values at the stations. Further details in the observational datasets are described by Casanueva et al. (2017).

2.3 Downscaling and bias correction

The ECMWF ensemble forecasts of the relevant parameters were first interpolated to the exact locations of the observation sites and then post-processed using a quantile mapping technique. Out of these bias corrected forecasts the WBGT is computed, yielding an ensemble of WBGT predictions.

Specifically, the gridded model output of raw forecasts and hindcasts were interpolated to the exact coordinates of observational stations with bilinear interpolation (see the chapter of bi-linear interpolation in the User guide to ECMWF forecast products; https://www.ecmwf.int/files/user-guide-ecmwf-forecast-products). A Quantile Mapping (QM hereafter) approach, also called distribution mapping or

quantile-quantile transformation, was applied to remove systematic biases. The QM approach is often used in the post-processing of forecast and climate model data (e.g. Verkade et al., 2013; Gneiting 2014; Rajczak et al., 2016a; Bedia et al., 2017). QM has a high effectiveness in removing bias and slightly enhances the reliability of the ensemble forecast (Crochemore et al., 2016).

A non-parametric empirical implementation of QM was applied to correct the forecasts using the 20 years of hindcasts and corresponding observations. The chosen approach refers to the implementation by Rajczak et al. 2016a, 2016b and is conceptually similar to that by Themessl et al. 2012. The Empirical Cumulative Distribution Function (ECDF hereafter) of the hindcasts and observations at the lead time t and at a specific station is computed. The raw forecast model time series X at the lead time t is matched to the ECDF of the hindcasts. The mapping allows then to apply a correction function (for each quantile), which is derived out of the differences out of the ECDF of the observations and hindcasts.

The 20 time series (20 years) and the 11 hindcast members at the lead time t are pooled for the computation of the ECDF. The sample for ECDF was increased by including hindcast and observational datasets from the previous and the following hindcast starting dates. The whole procedure can be defined as:

$$Y_t = \mathrm{ECDF}^{\mathrm{obs,cal}^{-1}}[\mathrm{ECDF}^{\mathrm{mod,cal}}(X_t)] \quad (1)$$

, where Y corresponds to the targeted and bias corrected forecast model time series, obs stands for observations and mod for the raw forecast model both in a calibration period (cal, 20 years of hindcasts and observations, see figure 17).

The sample sizes for the ECDFs were further increased by using a moving window of 7 days around the lead time t, resulting in a sample size of 420 in case of daily observations (20 years, 7 day moving window, 3 starting dates) and 4620 for the hindcasts (420, 11 members)

2.4 Heat stress indicator WBGT

Heat stress is quantified with the Wet-Bulb Globe Temperature (WBGT), a popular quantity used by human physiologists. It was introduced in the 50s during a campaign to investigate heat illness on soldiers in training camps of the United States Army and Marine Corps (Budd 2008). A special measuring device was constructed to realistically quantify the heat stress for human bodies. The resulting instrument contained a sling psychrometer for measuring the dry and shade wet bulb temperatures, a globe thermometer for assessing the effective radiation, thermo-anemometers for determining the wind velocity and a thermometer for the air temperature. The measured values were integrated and a basic formula was defined.

Heat stress is induced by high temperatures and incoming radiation. The human metabolism tries to adapt by cooling through transpiration, which is again impacted by the humidity and wind speed of the surrounding air. The principle variables affecting heat stress are therefore temperature, humidity, radiation and wind speed. In the context of heat stress affecting workers, it is often distinguished between indoor (or in the shade) and outdoor (or in the sun) working environments. The former is largely determined by temperature and humidity, whereas the latter includes also wind speed and solar radiation. In the following we refer to these two variants by $\mathrm{WBGT}_{\mathrm{sun}}$ (outdoor) and $\mathrm{WBGT}_{\mathrm{shade}}$

(indoor). Lemke and Kjellstrom (2012) compared published methods of computing WBGT from standard weather and climate data. They recommended to use the implementation of Bernard and Pourmoghani (1999) for the calculation of the $\mathrm{WBGT_{shade}}$. The $\mathrm{WBGT_{shade}}$ is defined by Bernard and Pourmoghani (1999) as:

$$\mathrm{WBGT_{shade}} = 0.67 * \mathrm{T_{pwb}} + 0.33 * \mathrm{T_a} \quad (2)$$

, where the psychrometric wet bulb temperature is denoted as $\mathrm{T_{pwb}}$ and $\mathrm{T_a}$ is the air temperature. The psychrometric wet bulb temperature is calculated from the air and dew point temperatures. The $\mathrm{WBGT_{shade}}$ computation is described in more detail in the Appendix 1.

For computing the $\mathrm{WBGT_{sun}}$, Lemke and Kjellstrom (2012) advised to take the formula derived by Liljegren et al. (2008). Wind speed, air temperature, dew point temperature and global radiation are needed for calculating $\mathrm{WBGT_{sun}}$. The basic formula of the $\mathrm{WBGT_{sun}}$ is referred from Yaglou and Minard (1956), which introduced it as:

$$\mathrm{WBGT_{sun}} = 0.7 * \mathrm{T_{nwb}} + 0.2 * \mathrm{T_g} + 0.1 * \mathrm{T_a} \quad (3)$$

, where $\mathrm{T_{nwb}}$ is defined as the natural wet bulb temperature and $\mathrm{T_g}$ stands for the globe temperature and they are calculated from air and dew point temperatures, wind speed and solar radiation. The $\mathrm{WBGT_{sun}}$ is defined and described in more detail in Appendix 1.

Daily data from the input variables were assimilated and used to compute the WBGT forecasts. Daily maximum values were taken for the air temperature and radiation to target the highest heat stress values, whereas daily mean values were chosen for determining the wind speed and dew point temperature. The implementation of the WBGT was defined in the deliverable 1.2 of the HEAT-SHIELD project (Casanueva et al. 2017). The detail code for the computation of the $\mathrm{WBGT_{sun/shade}}$ is available on https://github.com/anacv/HeatStress.

2.5 Verification

The quality of a forecast can be objectively measured by various scores quantifying the difference between past forecasts and their observed counterparts (Jolliffe and Stephenson 2012, Wilks 2011). A skill score relates the quality of a forecast to a reference forecast or baseline. In this study, a 20 year climatology is used as the reference forecast to calculate the skill scores. The skill scores thus show how much better (or worse) the forecasts predict weather than a sample of the climatology. Single verification scores validate only certain aspects of the forecasting quality, a combination of several verification metrics assesses the forecast performance more adequately. Murphy (1993) suggests to use a set of verification scores to assess the association, accuracy, reliability and the discrimination as forecast qualities.

The ensemble correlation can be used to validate the association between the mean ensemble forecast and observational values. The Continuous Ranked Probability Skill Score (CRPSS hereafter) is an adequate measure for validating the accuracy of ensemble forecasts. The reliability of a forecast can be assessed by the Spread to Error Ratio, which validates the agreement of the forecast probabilities to the observed frequencies. The discrimination, as the ability of a forecast to discriminate outcomes from the observations, is achieved by the Receiver Operating Characteristic (ROC hereafter).

The ROC area score is used to validate the discrimination of the forecasts to estimate the probability of exceeding the selected $WBGT_{shade}$ thresholds.

Probabilistic verification scores depend also on the number of members used in the ensemble forecasts. Thus, large ensemble sizes produce lower scores ("better forecasts") than small ensembles (Smith et al., 2015, Müller et al., 2005; Weigel et al., 2007b; Weigel et al., 2012). This is important to consider for the forecasts used in this work, as we verify the forecast system based on the hindcasts exhibiting a smaller ensemble size (11 members) than the operational forecasts (51 members). It has been shown that skill scores can be adjusted for limited ensemble sizes (Weigel et al., 2008).

Ferro et al. (2008) and Ferro et al. (2014) introduced the concept of fair scores accounting for the effect of limited ensemble size. We therefore use "fair" implementations of the scores for all probabilistic skill measures in this work, e.g. the Fair Continuous Ranked Probability Skill Score.

As described in 2.1, the verification analysis was carried out based on summer re-forecasts or hindcasts between 1997 and 2015. For this analysis, the bias correction of the hindcast was performed in a leave-one-out cross-calibration mode (similar to Themessl et al. 2012), in order to avoid artificial skill by including the respective observation years (1 out of 20) in the corrections. Therefore the sample sizes for the QM used in the verification analysis was slightly reduced as compared to the QM applied to operational forecasts, resulting in a sample size of 399 in case of the observations (19 years, 7 day moving window, 3 starting dates) and 4389 for the hindcasts (399, 11 members).

All verification analyses were performed with R (R Core Team, 2016) using the R-packages "easyVerification", version 0.4.09003, and "SpecsVerification" The applied skill metrics are defined and described in more detail in the Appendix 2.

3 Verification of heat stress forecasts

This chapter presents the performance of $WBGT_{shade}$ forecasts and those of its underlying variables based on summer hindcasts. The skill of $WBGT_{sun}$ (not analysed) is expected to be roughly similar, given the superiority of the temperature and humidity influence on WBGT, as investigated by Casanueva et al. (2017). The skill analysis includes 1) skill scores evaluated at different lead times over the full forecast range (i.e. day 1 to day 45), presented as averaged scores over all available stations across Europe and 2) maps showing the spatial distribution of scores at selected lead times of 5, 10, 15 and 20 days.

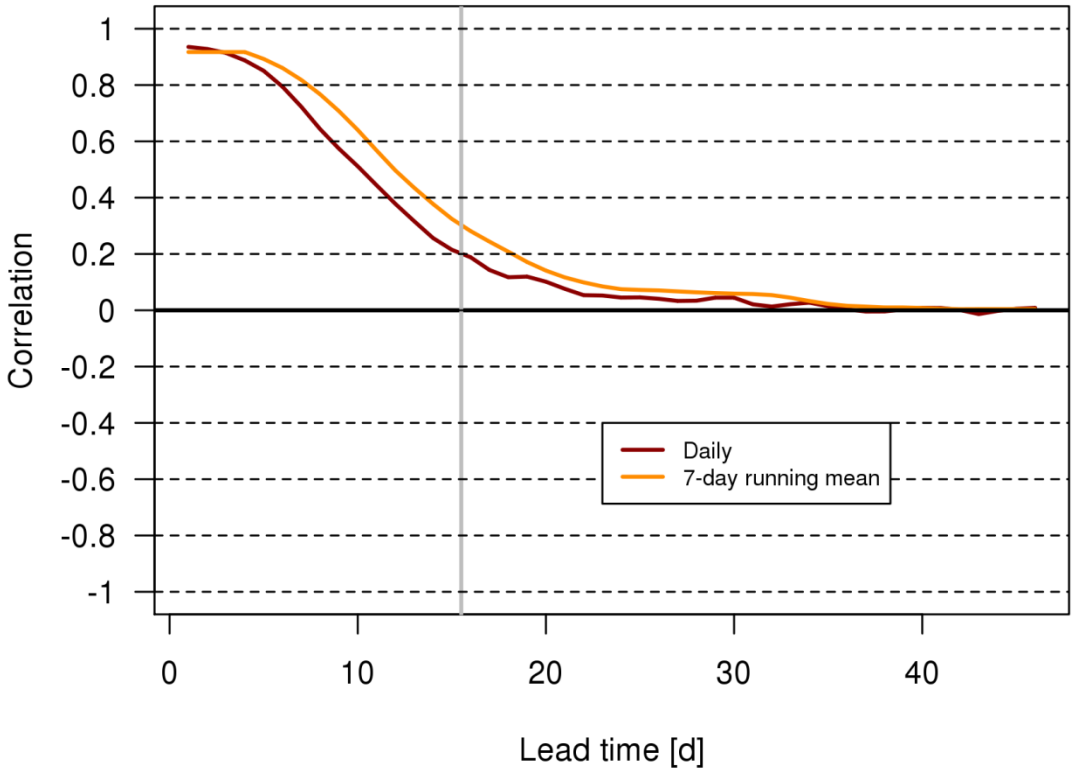

Figure 1: Correlation between the daily and 7-day running mean summer $WBGT_{shade}$ hindcasts (emsemble mean) and observations averaged over the 1799 stations. The vertical grey line between the 15th and 16th day indicates the change in model horizontal resolution.

Figure 1 shows the average correlation between hindcast ensemble means and observations at different forecast lead times, considering all available observation sites throughout Europe and includ-

ing all 1997-2015 hindcasts initialized in the months May to July. The red curve represents the correlation from daily values, whereas the orange curve is the correlation of the weekly running mean of the hindcasts and observations. In the first days, the correlations of the daily values reaches values above 0.95. Around the 10th day the correlation falls below 0.5. The correlation comes close to zero around the 20th day of the forecast and remains around zero thereafter. The correlation for the 7-day running mean is higher than that of the daily values from day three on. The difference is most pronounced at lead times up to 20 days, where the weekly averaging results in correlation gains of about 3-4 days.

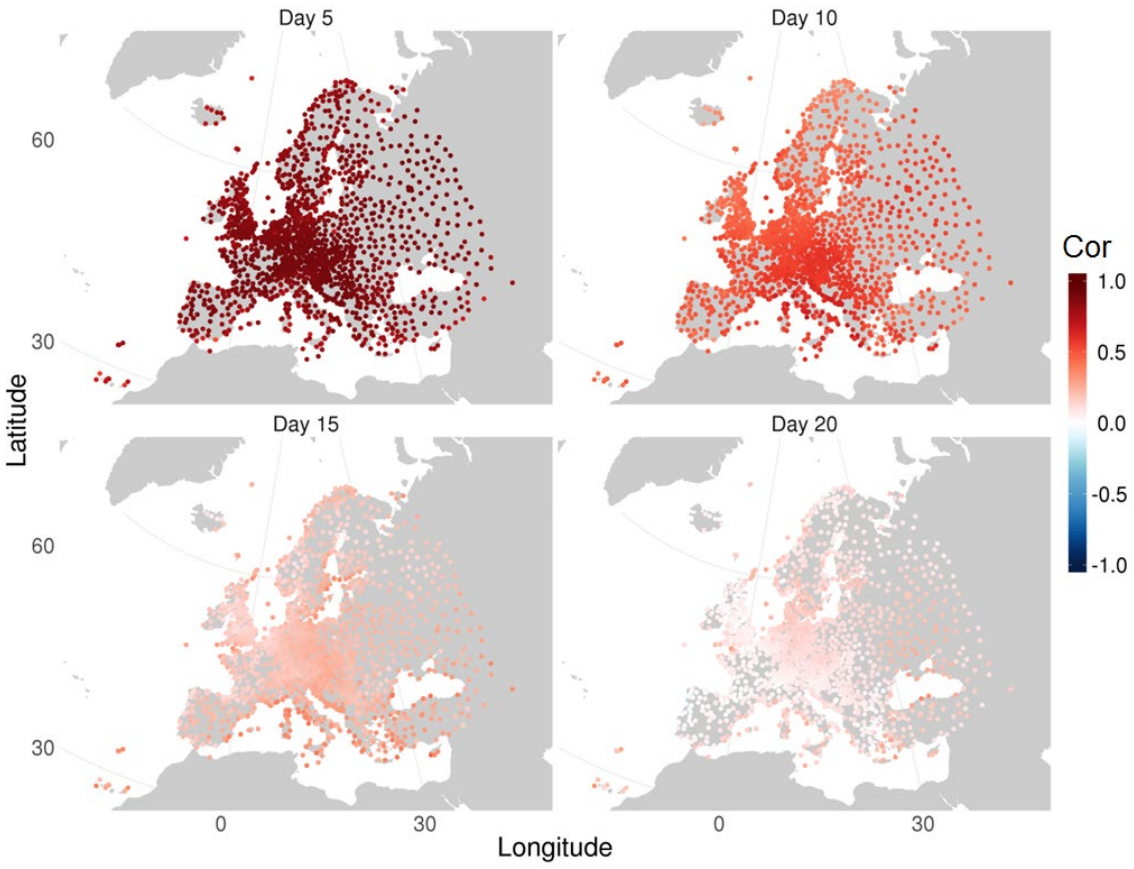

Figure 2: orrelation between the $WBGT_{shade}$ hindcasts (ensemble mean) and observations for the 1799 stations. Darker colours indicate stronger correlations.

Looking at the spatial distribution of the correlation across Europe (see Figure 2), three major groups can be identified: 1) Stations along the shores of Europe have a slightly lower than average correlation at the 5th day, but those stations exhibit higher than average correlations at the 15th and 20th day. 2) Stations in Ukraine, Western Russia and Northern Turkey have higher than average correlations at almost all lead times. 3) Over Italy, the Alps and the Balkans, the correlation is higher than the average at the 10th day.

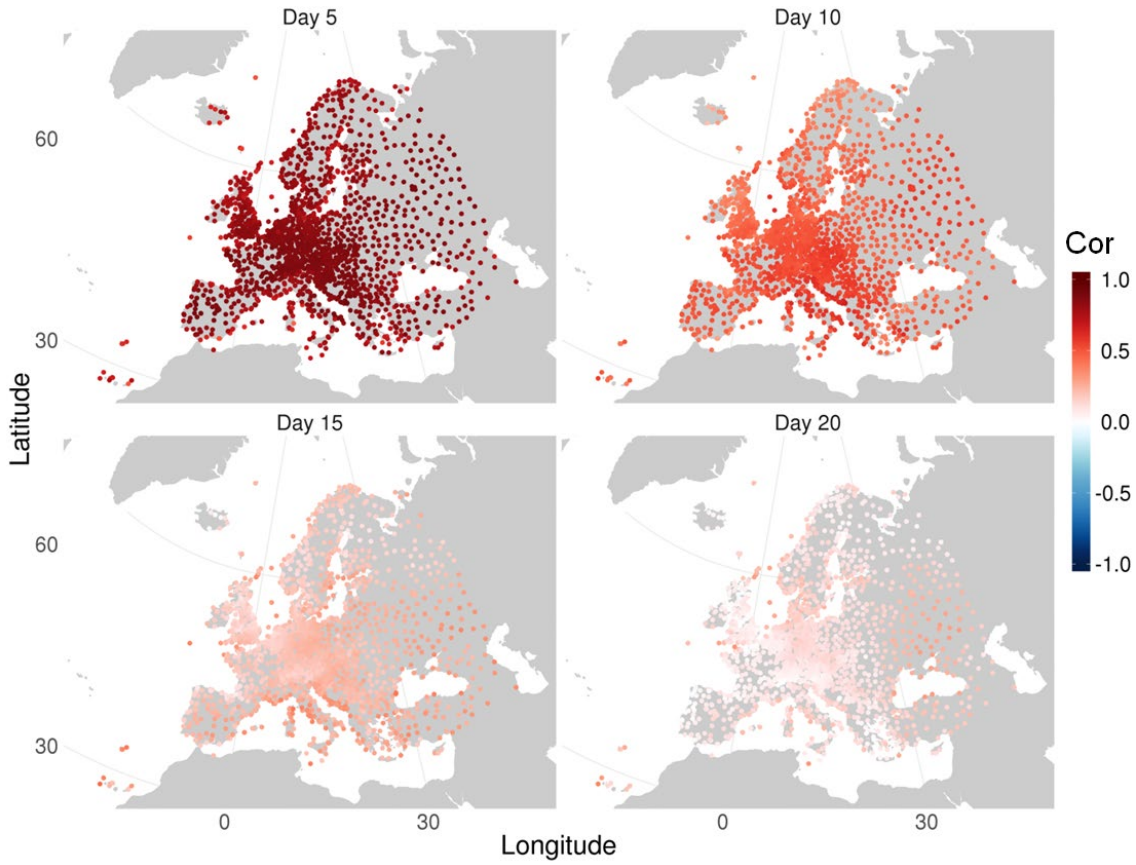

Figure 3: Correlation between the air temperature of hindcasts (ensemble mean) and observations for the 1799 stations. Darker colours indicate stronger correlations.

The correlation of the $WBGT_{shade}$ input variables show a slightly different spatial pattern, with overall smaller correlation values (see Figures 3 and 4). At the 5th day, the values are slightly lower in Scandinavian Peninsula and British Isles. Some stations in Ukraine, Russia, Northern Turkey, around the western Mediterranean Sea and North Sea show higher scores than the average. At 20th day of the forecasting, the correlation is relatively high at coastal stations and over Eastern Europe.

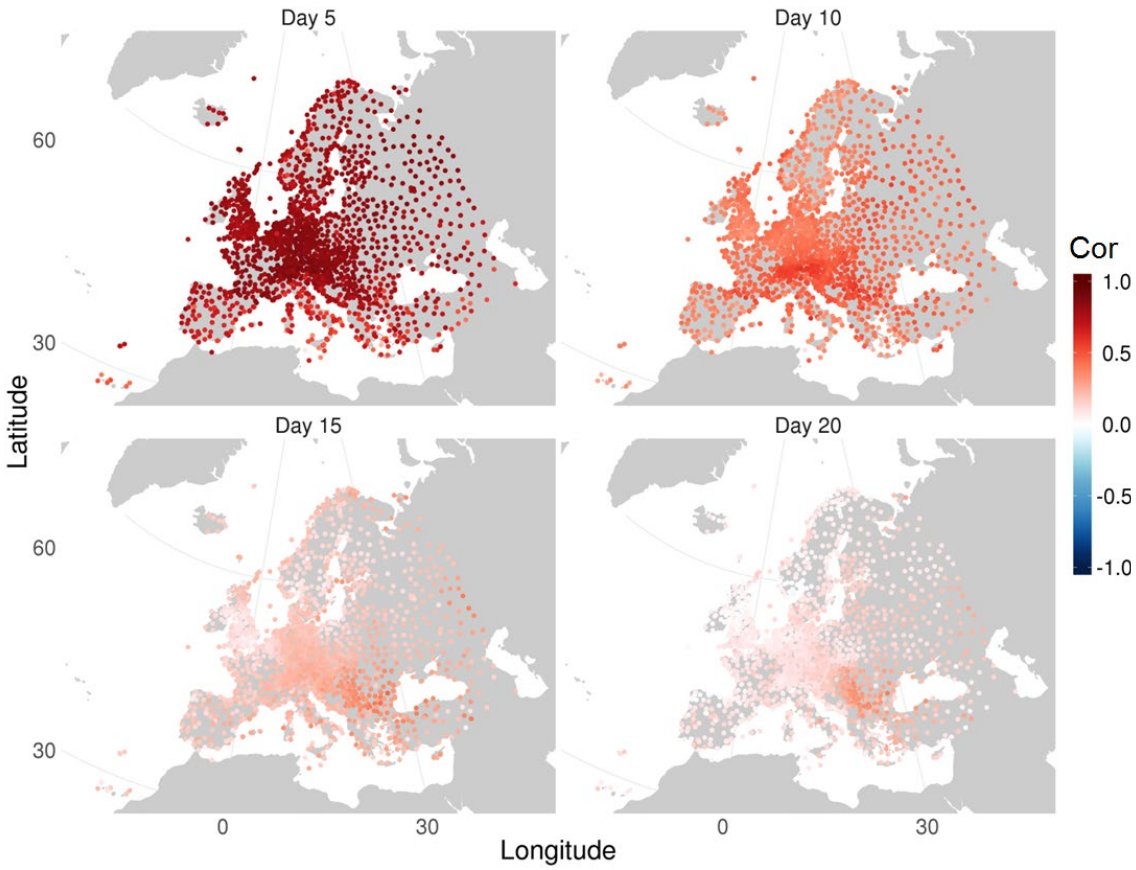

Figure 4: Correlation between the dew point temperature hindcasts (ensemble mean) and observations for the 1799 stations. Darker colours indicate stronger correlations.

The correlations of the dew point temperature hindcasts show highest values over the Alps, Pyrenees, Balkan region and Russia at the 10th day (see Figure 4). At longer lead times the Balkan region shows higher than average correlations, most pronounced at the 20th day, which agrees with results by Bedia et al. 2017 for seasonal forecasts. Many stations in the Mediterranean region have lower correlation at all lead times.

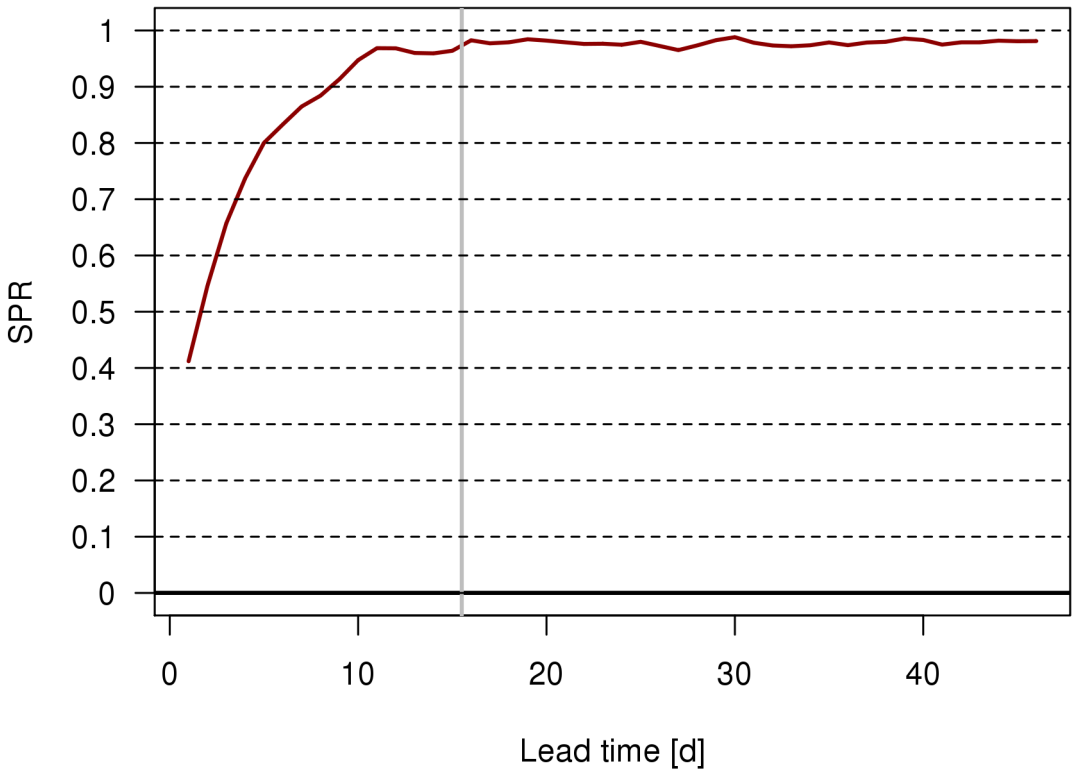

Figure 5: Fair spread to error ratio of the $WBGT_{shade}$ hindcasts averaged over the 1799 stations. The vertical grey line between the 15th and 16th indicates the change in model horizontal resolution.

The fair spread to error ratio averaged across Europe starts with 0.4 (see Figure 5) before reaching values close to one around the 10th day. It indicates that the forecasts are unreliable because of under-dispersion or overconfidence at the beginning and become reliable beyond day 10.

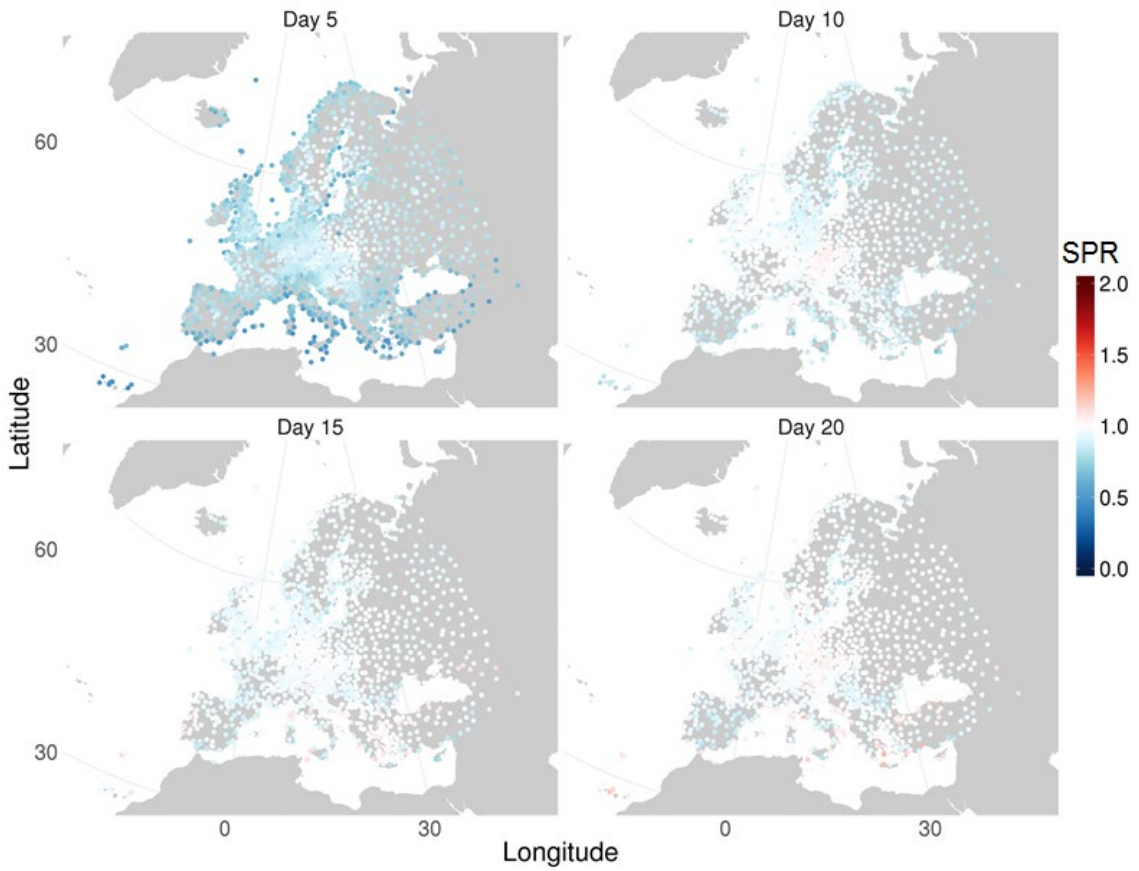

Figure 6: Fair spread to error ratio of $WBGT_{shade}$ hindcasts. Bluish colours indicate an underdispersion and reddish colours imply overdispersive forecasts.

The spatial analysis of SPR also shows the significant overconfidence at day 5 and that the average reliability beyond day 10 is composed of both areas with slight overconfidence and overdispersion, as for example in Mid to Eastern Europe (see Figure 6).

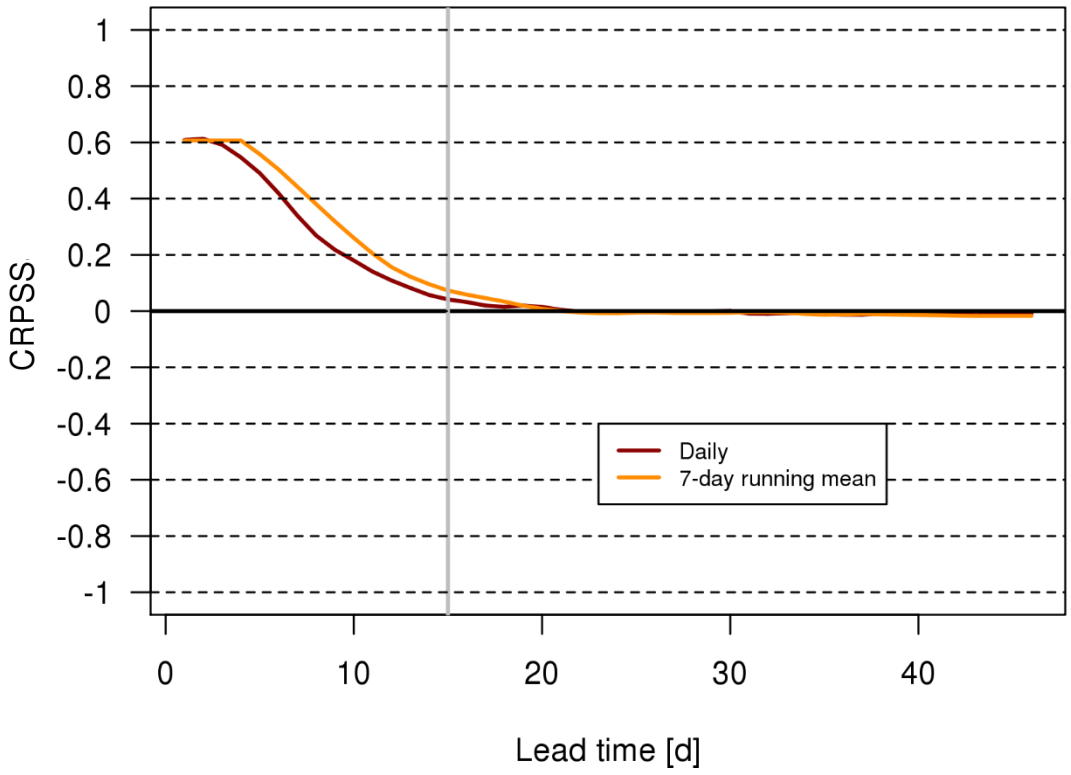

Figure 7: Fair continuous ranked probability skill score of the daily and 7-day running mean summer $WBGT_{shade}$ hindcasts averaged over the 1799 stations. The vertical grey line between the 15th and 16 day indicates the change in model horizontal resolution.

Correlation as shown above represents a potential measure of forecast performance as it does not consider absolute differences between forecasts and observations but only its (relative) association. In contrast, continuous ranked probability score (CRPS) represents an integral measure of probabilistic forecast performance considering also accuracy in absolute terms. The averaged fair CRPSS for stations across Europe starts with 0.6 and falls below 0.2 around the 10th forecast day (see Figure 7). As already seen in terms of correlation, time aggregation results in increased predictability, the skill for weekly values is higher, a gain of about 3 days in terms of forecast horizon. After the 20th day, the forecasts provide no added value as compared to the climatological reference forecast.

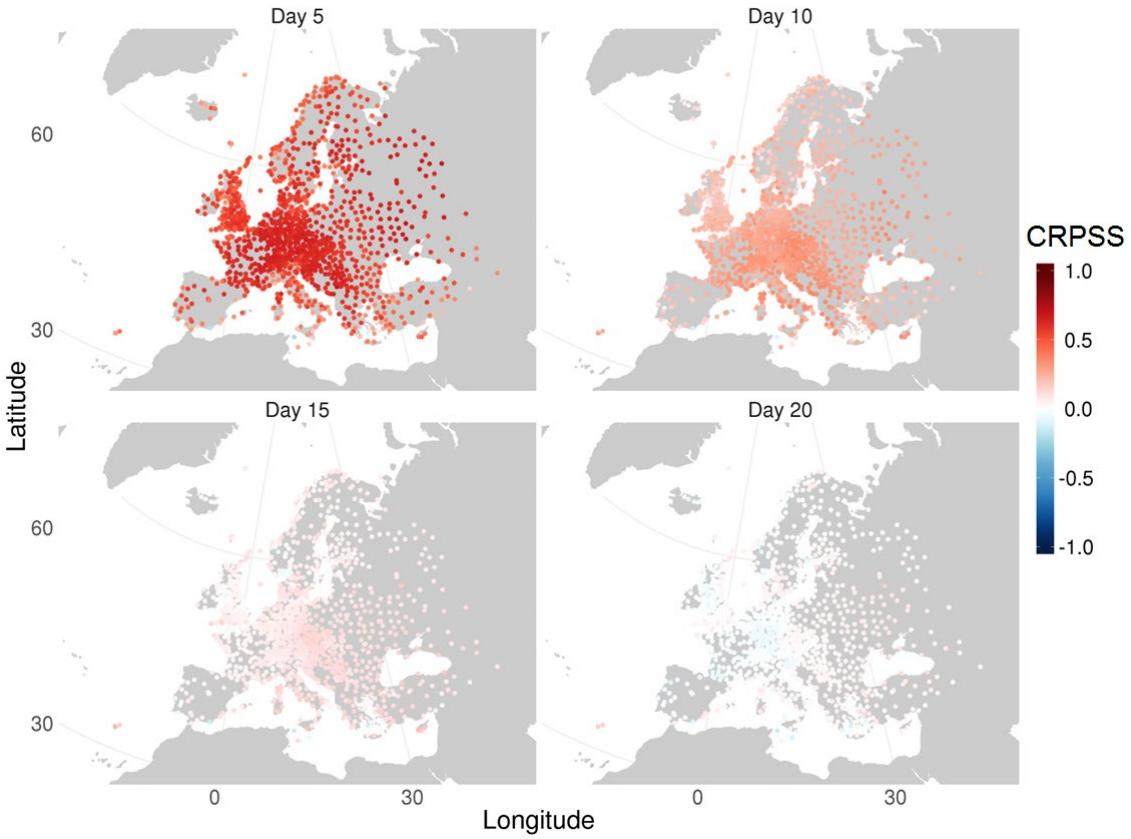

Figure 8: Fair continuous ranked probability skill score of the $WBGT_{shade}$ hindcasts. Dark red colours means the forecast is better in predicting the tercile than the climatology. White and bluish colours indicate that the forecast has no skill.

Considering the spatial pattern of $WBGT_{shade}$ skill in terms of CRPSS at day 5, below average values are found at coastal regions (see Figure 8), highest skill values are found in the Balkans, Germany, France, and Russia. At longer lead times, the skill scores in coastal regions improve relatively to those of the mainland, and highest skill scores are reached in Central, Eastern and South-eastern Europe.

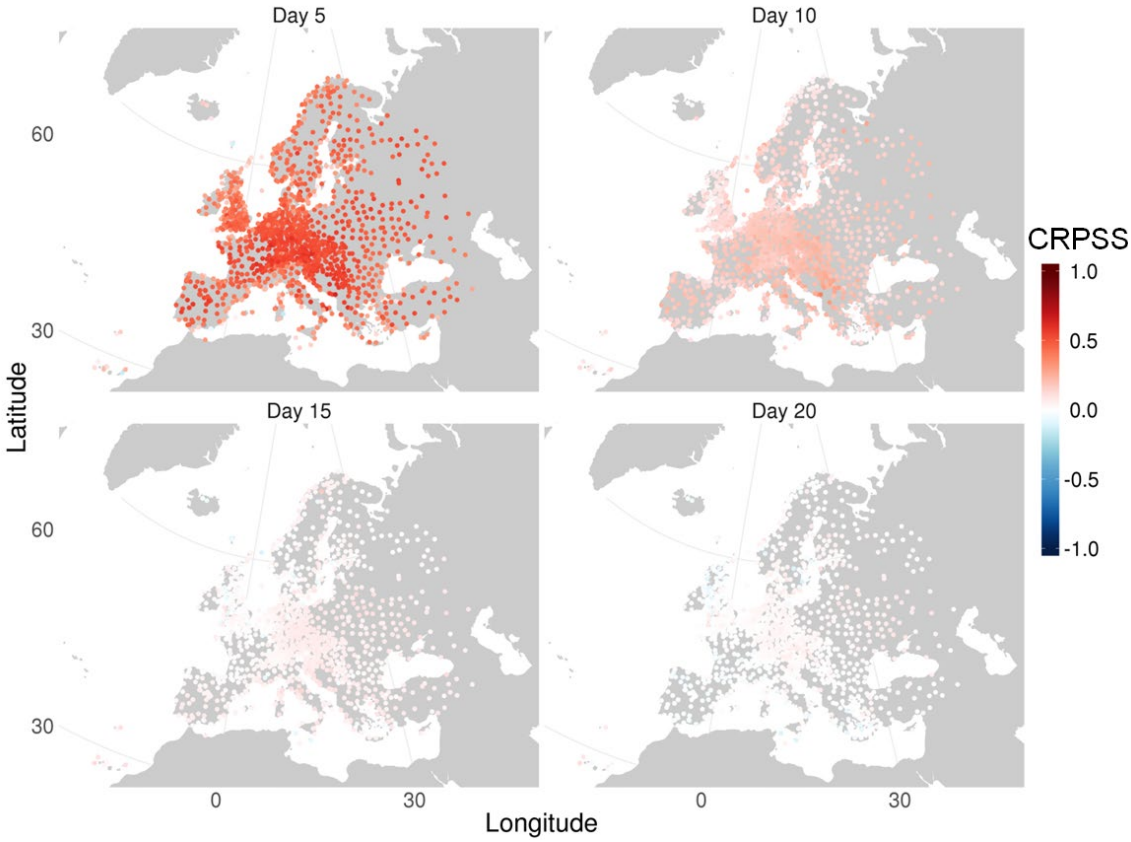

Figure 9: Fair continuous ranked probability skill score of the air temperature hindcasts. Dark red colours means the forecast is better in predicting the tercile than the climatology. White and bluish colours indicate that the forecast has no skill.

It is also interesting to compare forecast performance of WBGT with that of its underlying variables air and dew point temperatures. Figures 9 and 10 show that skill of the bias-corrected variables decreases in similar fashion with lead time as that of $WBGT_{shade}$ (see Figure 8), but skill of $WBGT_{shade}$ is higher than that of the underlying variables, as previously seen for correlation. The spatial skill pattern of the underlying variables are also slightly different. At day 5, the CRPSS of air temperature hindcasts is high across Europe, apart from a few stations at the coasts. 5 days later, CRPSS has decreased significantly, with lower values in Scandinavian Peninsula and British Isles and higher values over Central Europe, the Balkan region and Russia. At longer lead times, skill is marginal, whereas some stations in Ukraine, Russia, Northern Turkey, around the western Mediterranean Sea and North Sea show higher skill scores than on average.

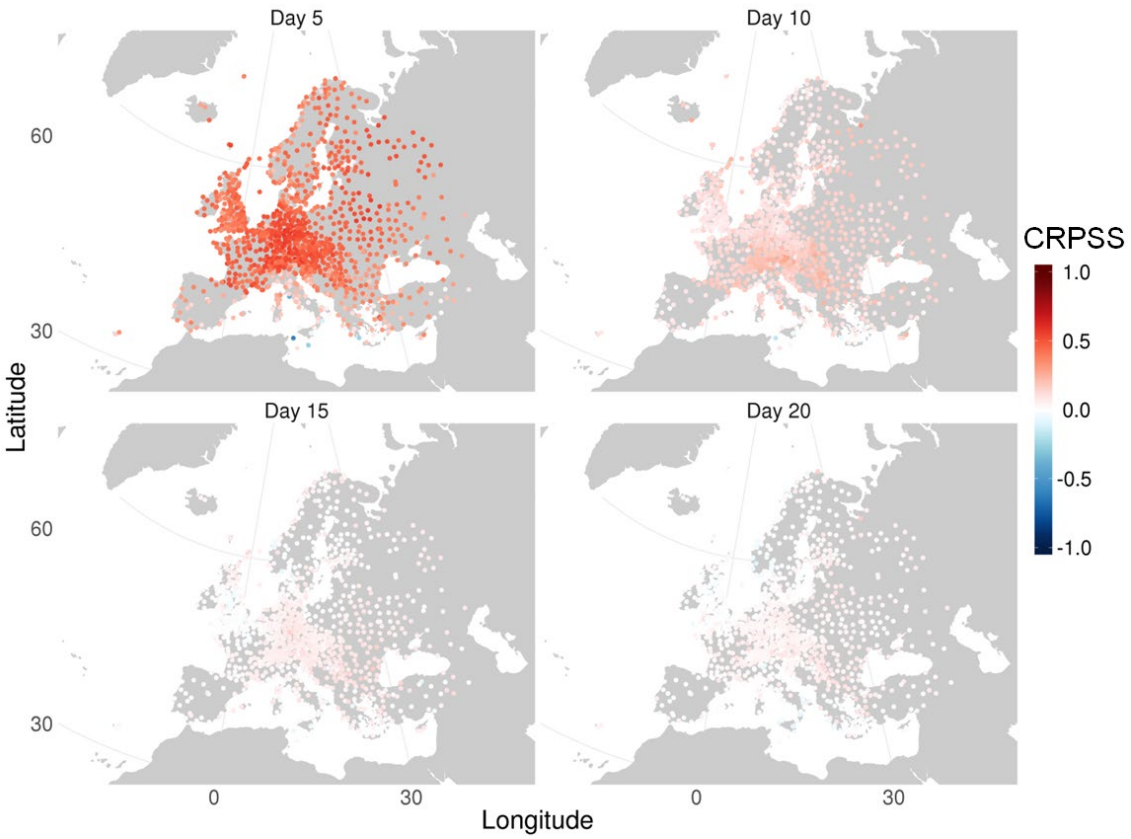

Figure 10: Fair continuous ranked probability skill score of the dew point temperature hindcasts. Red colours mean the forecast is better than climatology, bluish colours indicate that forecasts are worse than climatology.

The CRPSS of the dew point temperature is lower than that of temperature (compare Figures 10 and 11) at all lead times. The spatial pattern (see Figure 10) at the 5th day shows relatively low values at many coasts and highest skill score values in Central Europe and Russia. At the 10th day, the highest dew point CRPSS values are found in the Balkan region and over the Alps. Skill at longer lead times is again marginal.

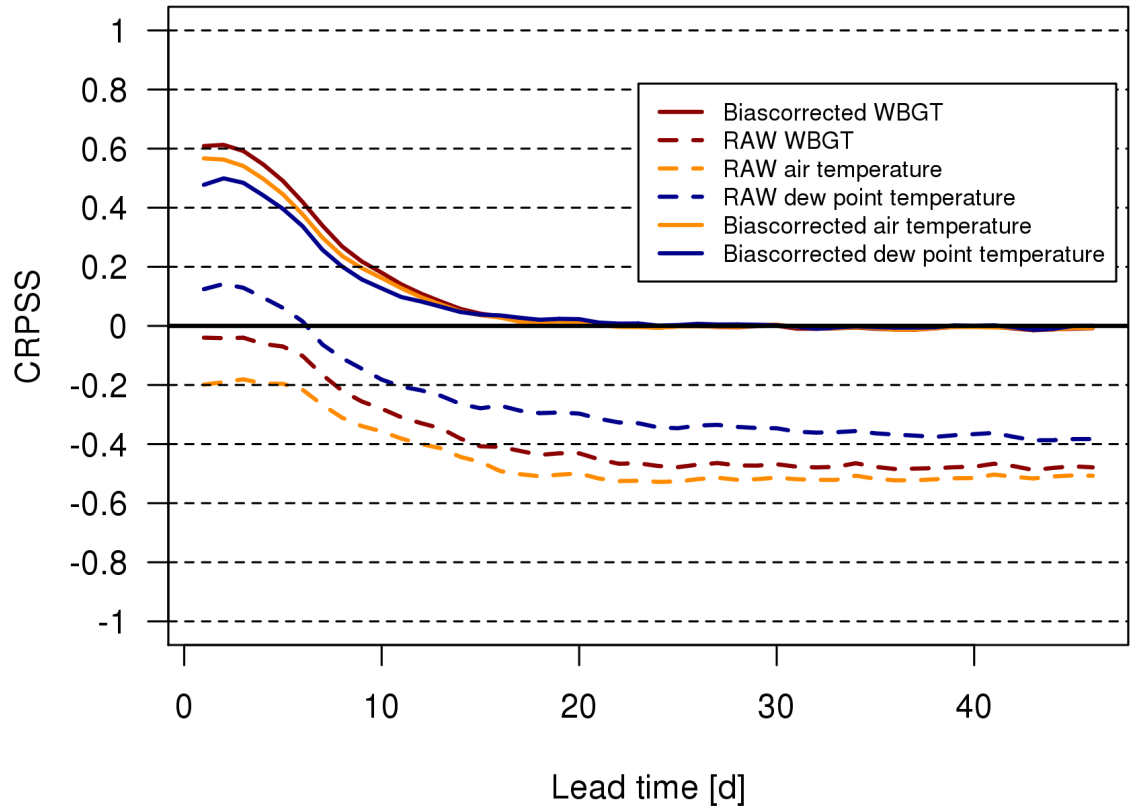

Figure 11: Fair continuous ranked probability skill scores of the summer $WBGT_{shade}$ hindcasts and its underlying variables.

The bias correction improves the skill of the forecasts substantially (see Figure 11) throughout the lead times. The bias-corrected joint product gets even higher skill than the underlying variables, whereas the raw joint product has an CRPSS which lies between the underlying variables. The higher skill of WBGT forecasts as compared to those of its underlying variables is striking. Considering the spatial picture, WBGT forecasts seem to benefit from relative skill in both underlying variables (see Figure 12), as the skill pattern of WBGT forecasts is almost combining the strengths in both underlying variables.

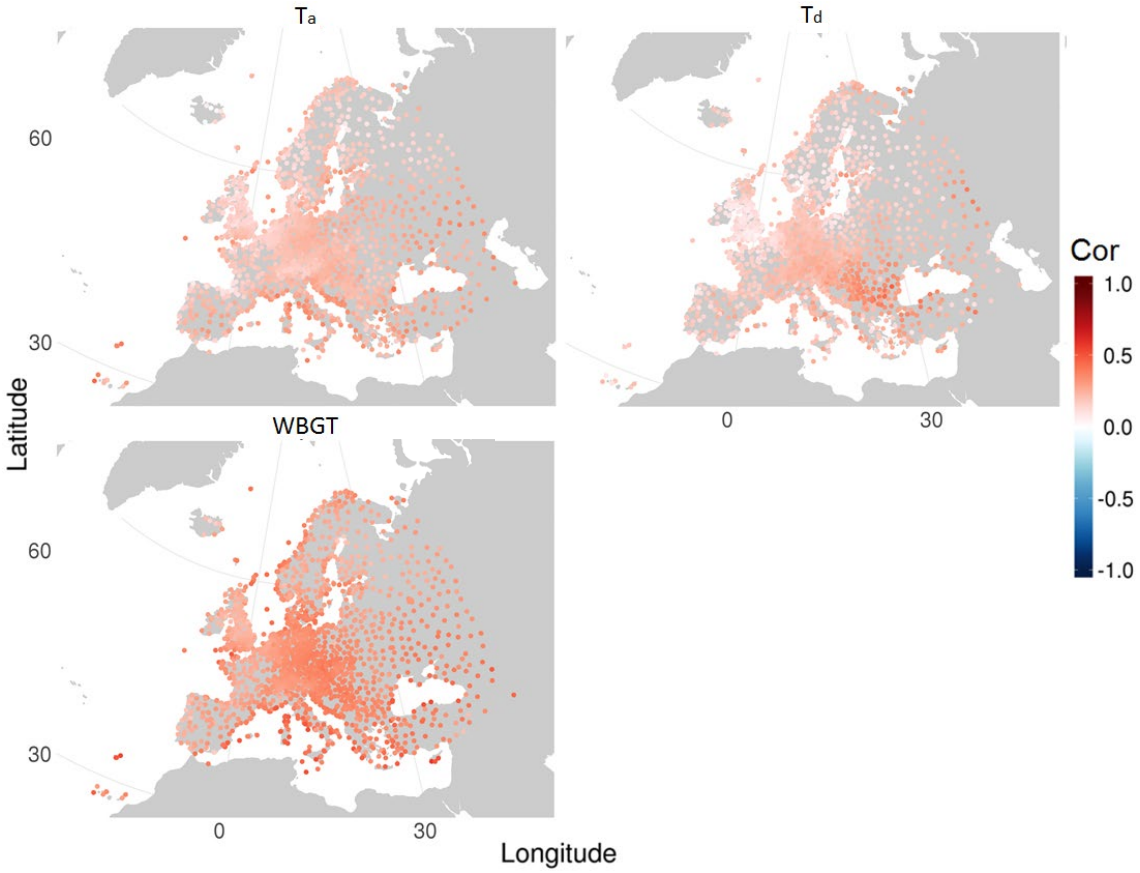

Figure 12: Correlation between the air temperature, dew point temperature and $WBGT_{shade}$ hindcasts with observations for the 1799 stations on the 10th day of the forecast. Darker colours indicate stronger correlations.

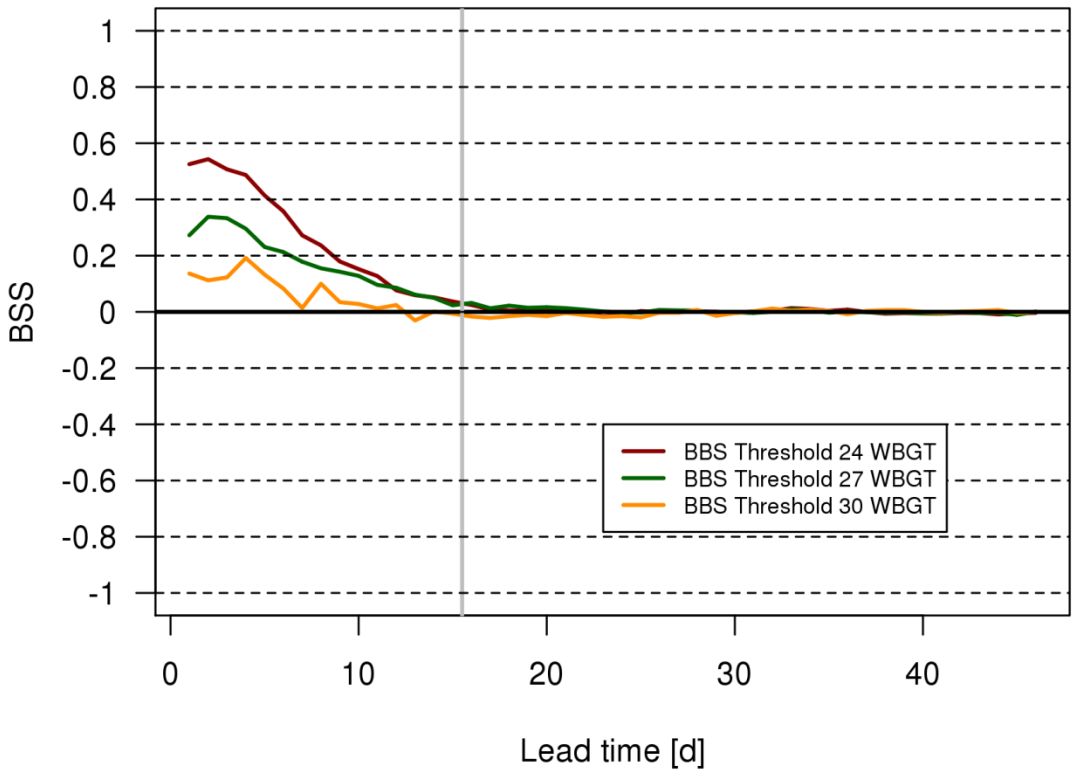

Figure 13: Fair Brier skill scores for three $WBGT_{shade}$ thresholds of the summer $WBGT_{shade}$ hindcasts averaged for the 1799 stations. The vertical grey line between the 15th and 16th day indicates the change in the model horizontal resolution.

The ability of discriminating a $WBGT_{shade}$ threshold of 24 °C is presented in the following results. The averaged fair Brier skill score (BSS) for the $WBGT_{shade}$ threshold 24 °C across Europe begins with values larger than 0.5 in the first days (see Figure 13). There were significant amount of stations with skill scores of negative infinity, “Not-a-Number” or “Not-Available”. BSS can only be computed if full pairs of observations and forecasts are available, so averaging was across stations with available BSS. Around the 8th day, the fair BSS falls below 0.2 and reaches 0 around the 20th day. The fair BSS for the $WBGT_{shade}$ threshold of 27 °C starts with a value larger than 0.2 and gets larger after the 2nd day. Around the 6th day, the skill score falls below 0.2. After the 10th day, the fair BSS for both $WBGT_{shade}$ thresholds of 24 °C and 27 °C are very similar. The skill score for the $WBGT_{shade}$ threshold of 30 °C shows noisy behaviour already at the beginning. The fair BSS varies around values larger than 0.1 at the beginning. The score starts to oscillate around 0 already at the 10th forecast day. The larger the $WBGT_{shade}$ threshold is set, the fewer stations with fair BSS remain.

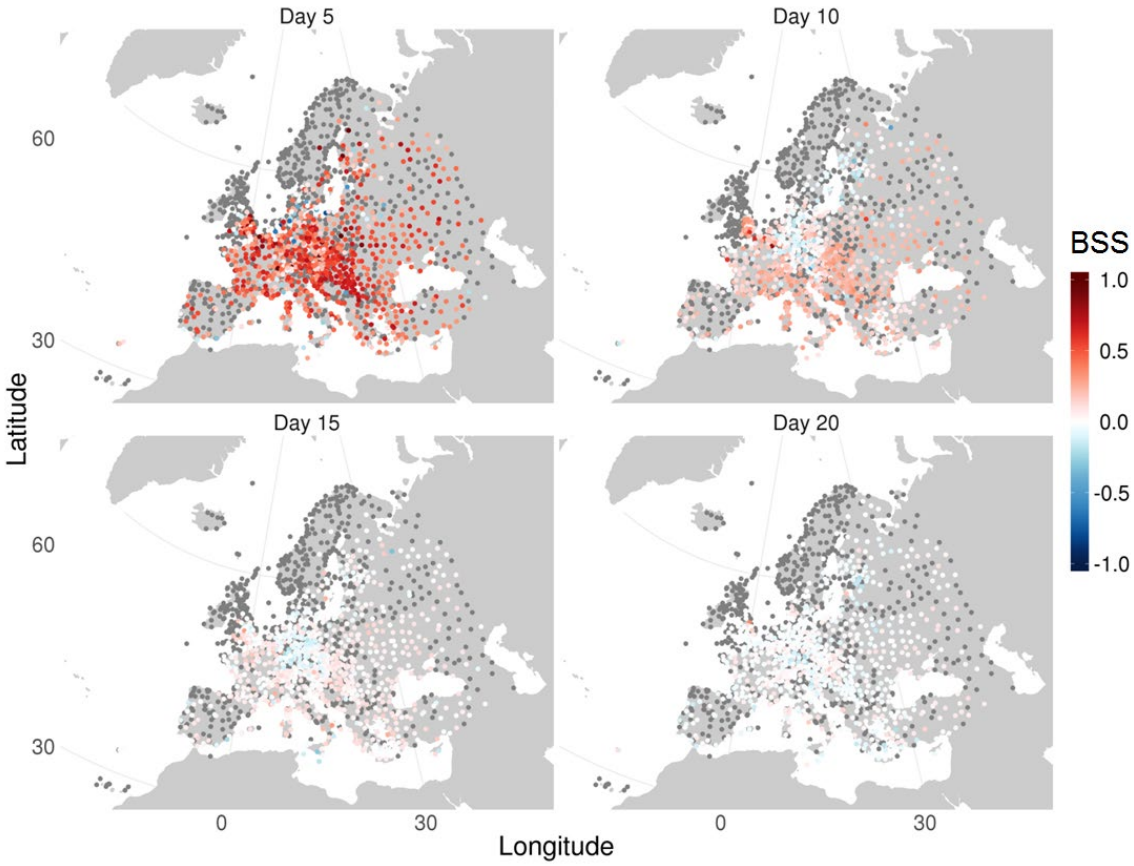

Figure 14: Fair Brier skill score of the summer hindcasts for the $WBGT_{shade}$ threshold of 24°C. Dark red colours mean that the forecast is better in predicting the category than the climatology. White and bluish colours indicate that the forecast has no skill. The grey colour represents stations, where the brier skill score could not have been computed because of insufficient observations above the threshold.

In climatological terms, $WBGT_{shade}$ gets never higher than 24 °C in Iceland, the North-western Great Britain and Scandinavia. Therefore, the fair BSS for the WBGT threshold 24 °C could not be computed for those region in North-western Europe (see Figure 14). At the 5th forecast day, negative fair BSS are randomly scattered apart from a region in North-western Germany. South-eastern England, Balkan region, Ukraine and Central Europe are areas with higher skill scores. At the 10th day, negative fair BSS are found at many stations in Germany, Southern Sweden and Southern Finland. Regions with higher skill scores than the average are found in South-eastern England, Balkans, Italy, Ukraine and Southern France. At the 15th forecast day, the major region of negative fair BSS lies in Germany.

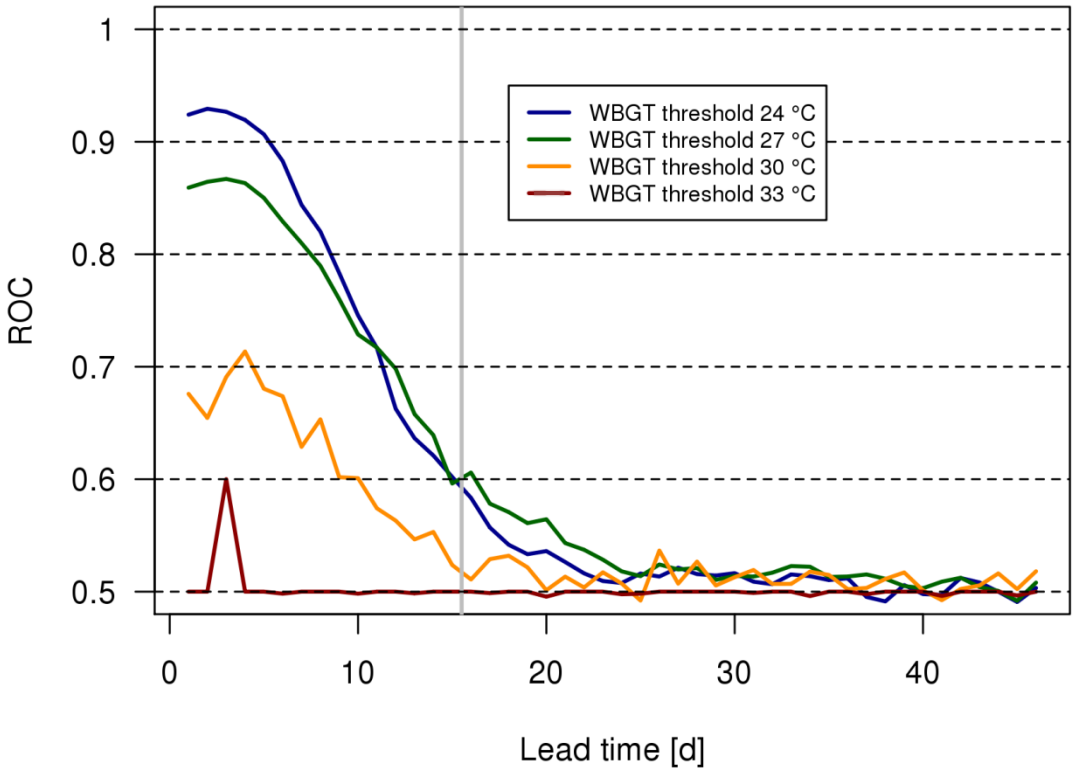

Figure 15: ROC area score of the summer $WBGT_{shade}$ hindcasts averaged for the 1799 stations. The vertical grey line between the 15th and 16th day indicates the chande in model horizontal resolution.

The analysis of the receiver operating characteristics (ROC) area score for different thresholds of $WBGT_{shade}$ clearly demonstrates the effects of the sample size of the underlying observations. Whereas the threshold of 24°C is reached throughout the whole investigation area, such of 30°C and higher are only observed in certain parts of Europe, resulting in a significantly smaller sample for this threshold, resulting in much noisier averaged ROC values in Figure 15. The ROC for a threshold of 24°C starts with values above 0.9 and remains at skilful values (i.e. >0.5) up to more than 20 days. The ROC area score for the $WBGT_{shade}$ threshold of 27 °C score behaves similarly as that of 24 °C, although it has lower values at short lead times. The ROC area score for the $WBGT_{shade}$ threshold of 30 °C is clearly lower and shows no more skill beyond the 15th day. The ROC area score for the $WBGT_{shade}$ threshold of 33 °C demonstrates that forecasts of such WBGT thresholds provide no added value as compared to naïve climatological forecasts.

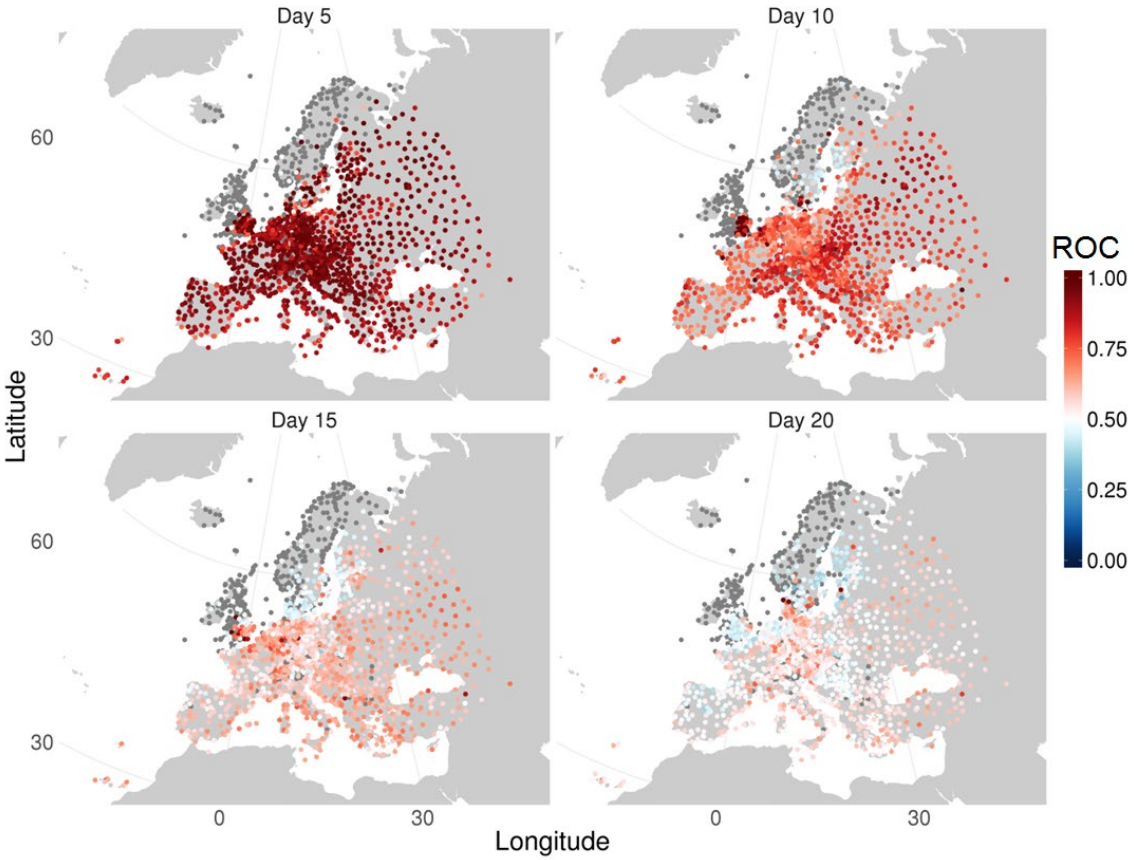

Figure 16: ROC area score of the summer hindcasts for the $WBGT_{shade}$ threshold of 24°C. Dark red colours means the forecast outperforms the climatology. White and bluish colours indicate that the forecast has no skill. The grey colour represents stations, where ROC area score could not be computed because of insufficient observations above the threshold.

Considering the spatial pattern of $WBGT_{shade}$, the highest ROC area scores for the threshold of 24°C are found in England, Central Europe, Southern Sweden and Czechia (see Figure 16). 5 days later, the highest scores are in England, the Alps, Central Europe, Slovakia and Russia. A cluster of negative ROC area scores lies in Southern Scandinavian Peninsula. At the 15th forecast day, good skill scores are reached by stations in England, Russia, in the Mediterranean and around the Benelux countries. There are still many stations in Southern Scandinavian Peninsula with slightly negative ROC area scores.

4 Setup of operational WBGT forecast proto-type

ECMWF initializes the real-time extended range forecasts every Monday and Thursday at 00 UTC and provides them at about 23 UTC on the same day. The hindcasts for the past 20 years initialized on the same date are provided about 4 days earlier. Forecasts and hindcasts of all variables required to compute WBGT are retrieved from ECMWF as soon as they are available in the form of ensembles at daily granularity. The bias correction for a given forecast is performed on the basis of hindcasts and corresponding observations. The hindcast sample for correcting a given forecast consists of the hindcasts with the same initialization date as the forecast, plus those the previous and following date, resulting in a sample size of 660, i.e. 3 dates x 20 years x 11 members. Figure 17 illustrates the concept of forecasts and corresponding hindcasts as used for bias-correcting the operational forecasts.

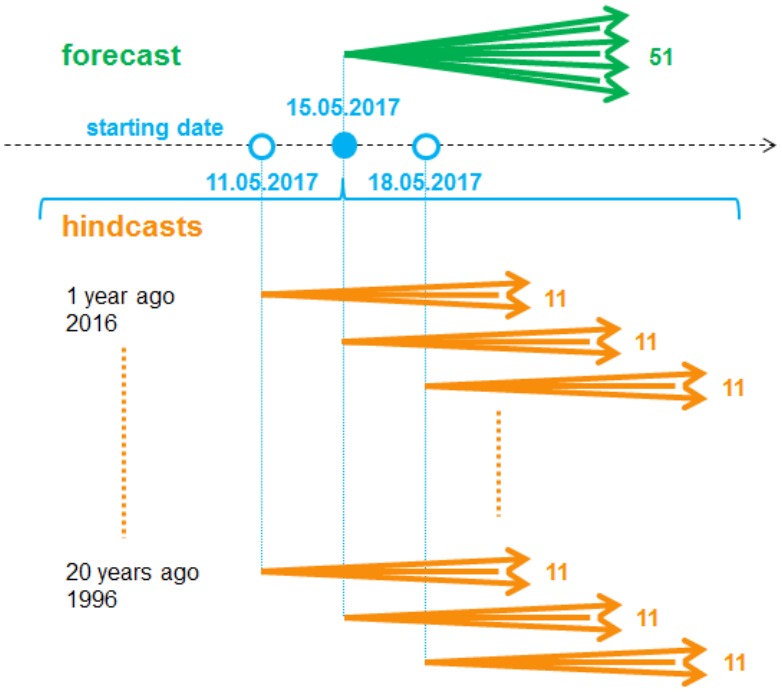

Figure 17: Concept of the operational real-time forecast and corresponding hindcasts. The 51 members ensemble forecasts (green) of any initialization date are accompanied by 11 members ensemble re-forecasts (hindcasts) of the past 20 years (orange). For bias-correcting a given forecast, hindcasts of three initialization dates (blue) are used.

The production of the WBGT forecast consists of two main computational steps, the bias correction of base variables and the WBGT computation. Figure 18 shows the production chain for the case of $WBGT_{shade}$ with its base variables air temperature and dew point temperature. As described in Chapter 2.3 the raw forecast is bias corrected by applying a quantile mapping to the distributions of daily values in the hindcast and observations. The quantile mapping correction was done in a lead time dependent fashion, by considering values in a 7 day moving window along the lead time for estimating hindcast and observation distributions. The quantile mapping is produced with the R-package "biascorrection". The WBGT computation is done with the R-package "HeatStress". In the second step, ensembles of daily WBGT forecasts are computed from the bias-corrected forecasts of the base variables. This primary WBGT forecast product serves as the starting point to derive any further user-specific product, such as probabilities of exceeding certain WBGT thresholds. The radiation and wind parameters have to be loaded and bias corrected additionally, in order to compute the $WBGT_{sun}$.

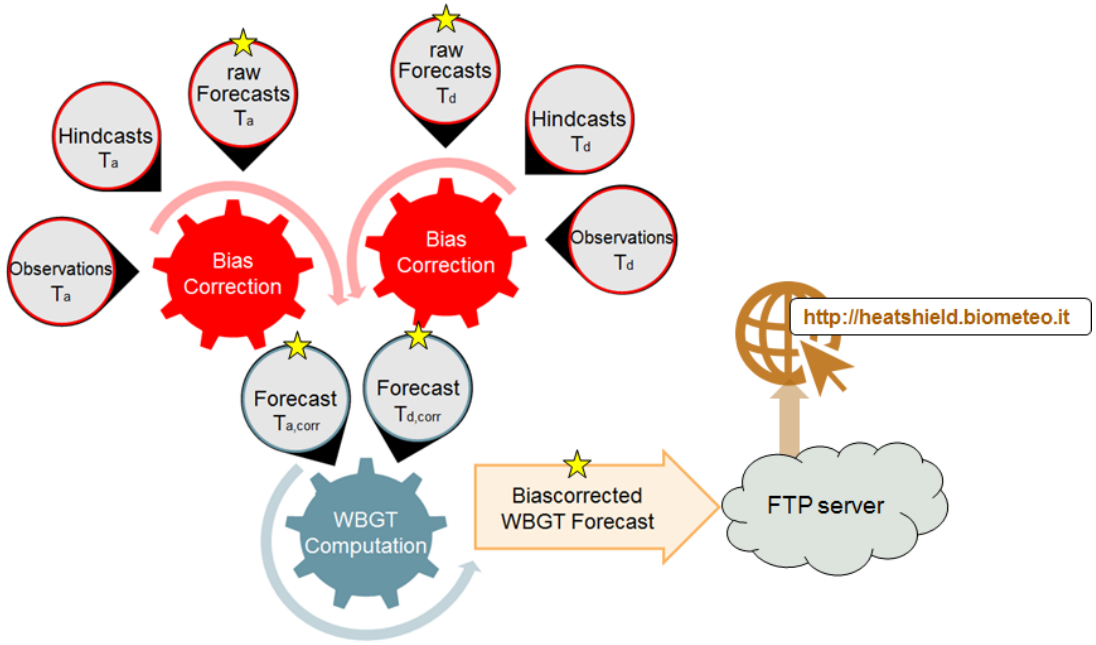

Figure 18: Illustration of the forecast processing for producing $WBGT_{shade}$ forecasts.

The heat stress forecast systems runs every Tuesday and Friday morning. The operational forecast production is accomplished using the R computing software (R Core Team 2016), and a specific R package "mfcdaily" has been built including functions for retrieving, bias-correcting and displaying forecasts as presented in this report. The R-package was designed to allow parallel running of the underlying calculations on several cores. The resulting runtime of one operational forecast for 1800 locations is about 13 minutes when using 12 cores. Functions from the R-packages "biascorrection" and "HeatStress" were packed into new function that enable them to wrap through datasets and to run in parallel mode. The package is released as a CAT (Climate Analysis Tool) on the MeteoSwiss servers.

The verification of the forecast system needs to process the observations and re-forecasts. Since the additional datasets are larger, the whole process chain gets more memory intense. The requested

memory goes beyond the installed memory of a single node on the server Kesch. The operational forecast production starts with single mother session, which process the whole datasets in parallel without any problems. For enabling the verification processing, the datasets were split into several mother sessions, which were parallelised with maximum 4 cores. The splitting reduces the memory of the mother and daughter sessions and speeds up the processing.

4.1 Verification

The heat stress forecasting system runs every Tuesday and Friday morning. The output of the forecast contains arrays of the bias-corrected $WBGT_{sun}$ and $WBGT_{shade}$ forecasts, and the probabilities of exceeding $WBGT_{shade}$ of 24 °C, $WBGT_{sun}$ of 27 °C and 30 °C. Those probabilities were computed as the running mean of members exceeding the thresholds with window size of 5 days. The forecast products are discussed in the following.

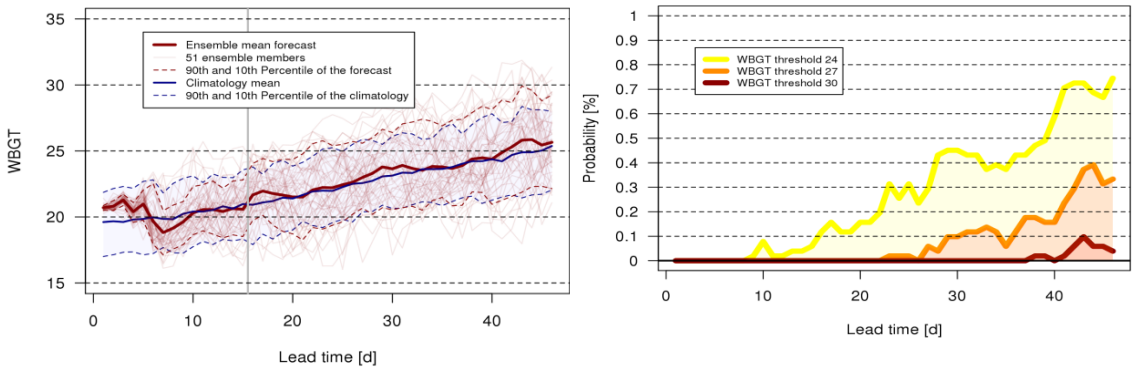

Figure 19: Left) $WBGT_{sun}$ ensemble forecast for the site Reggio Calabria (Italy) from 15.5.17, and right) corresponding probabilities of exceeding different $WBGT_{sun}$ thresholds.

The primary forecast output are daily ensembles of WBGT (both in the sun and in the shade) for 1799 locations in Europe (see an example of such a WBGT forecast in Figure 19, left). From these WBGT ensembles, probabilities of exceeding any threshold can be computed, thus allowing to address specific user needs on particular WBGT thresholds. The plot on the right of Figure 19 shows the daily probabilities of exceeding different $WBGT_{sun}$ thresholds for the same exemplary forecast. The better skill for weekly as compared to daily predictions can be accounted for in creating forecast products displaying the predictions aggregated over several days rather than daily information.

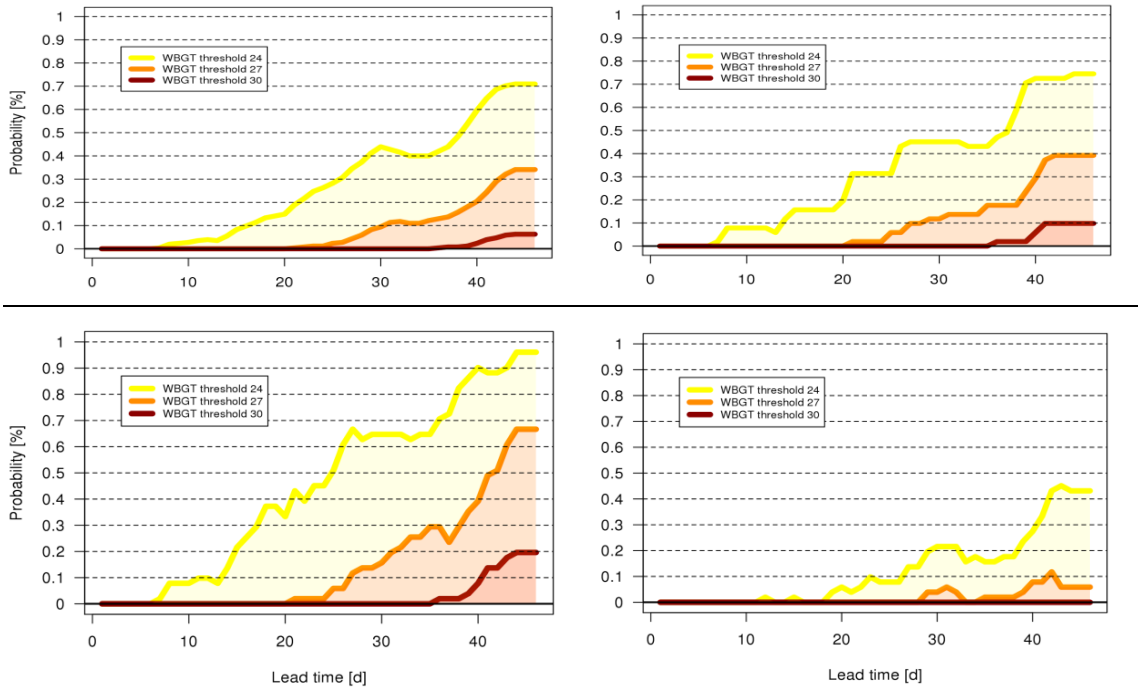

Figure 20: Variants of same forecast displaying probabilities for moving windows of 5 days, top left: average probabilities, top right: maximum probabilities, bottom left: probabilities of one or more threshold exceedances, bottom right: probabilities of two consecutive threshold exceedances within moving window.

Figure 20 shows different variants of the same forecast as in Figure 19, all presenting forecast information for moving windows of 5 days. Obviously it will depend on users' preferences which of the presentation formats and what WBGT thresholds should be chosen for operational forecasts. For the initial forecast prototype, we choose probabilities of exceeding $WBGT_{sun}$ thresholds of 27° and 30° as a basis and present weekly summaries in the form of average probabilities of exceeding these thresholds (top left in Figure 20).

The prototype of an early-warning system specifically for workers has been made operational as part of a web platform within the HEAT-SHIELD project (http://heatshield.biometeo.it/), along with short range heat risk forecasts (5-days forecasts) for the Tuscany region (see Figure 21).

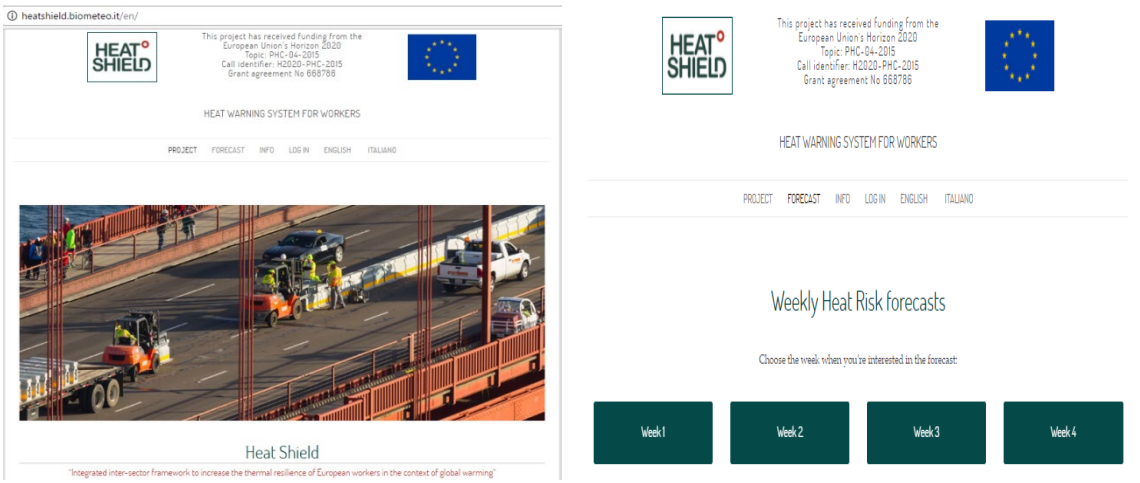

Figure 21: The web platform with occupational heat risk forecasts (left: welcome page) and its subpage on weekly heat risk forecasts for the upcoming month (right).

For each week, two European maps are available showing the averaged probabilities of $WBGT_{sun}$ threshold exceedances. In particular, the probabilities of exceeding $WBGT_{sun}$ thresholds of 27°C (hot days) and 30°C (very hot days) are shown for several European localities (see Figure 22). These maps are also shown in the web site of the HEAT-SHIELD project (https://www.heat-shield.eu/maps-forecast).

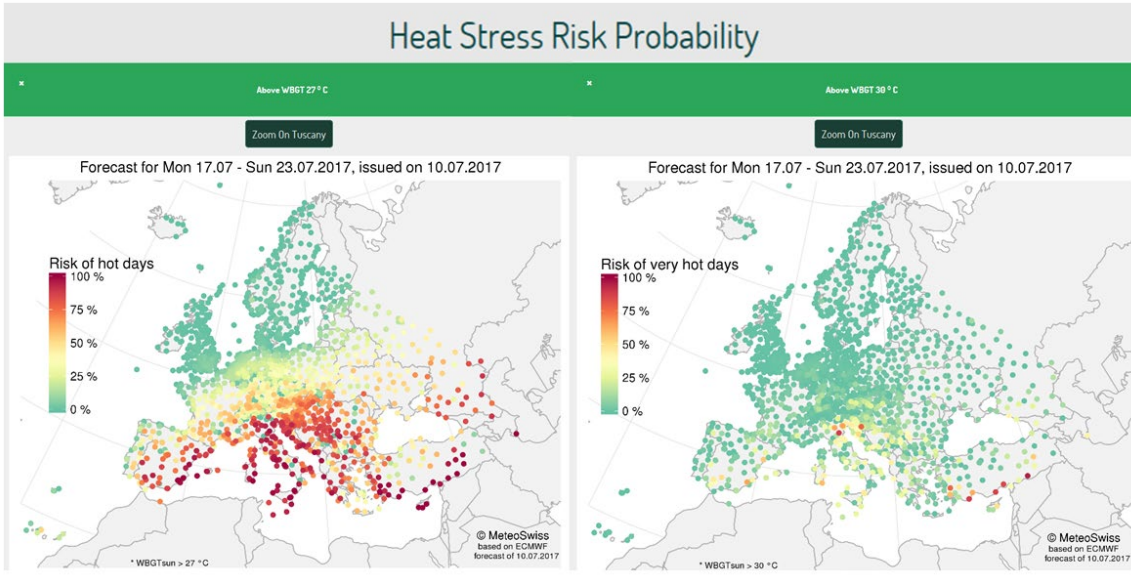

Figure 22: Weekly heat stress risk probabilities for hot (left, $WBGT_{sun}$> 27 °C) and very hot (right, $WBGT_{sun}$ > 30 °C) days for different European localities.

In the final version of the occupational early heat warning system, it will be possible to choose a specific location from these summary weekly maps to visualize more detailed information. This infor-

mation will only be accessible through a login where users need to provide some information, including the location (all over Europe), the physical activity of involved workers of interest and whether they are acclimatized or not, and also one or more email addresses to which warnings will be sent. In this way, a notification email will be sent twice per week (whenever the long-term probabilistic forecast is updated). In the email body, there will also be a link that allows the user to view the daily details of the probability of exceeding the $WBGT_{sun}$ heat threshold specific to the given working activity over a long period (up to 46 days), by means of a coloured calendar-based on an equal range probability scale. At the moment, the prototype does not include yet the login mask and the detailed daily information in form of the calendar is only available for the Tuscany region (see Figure 23).

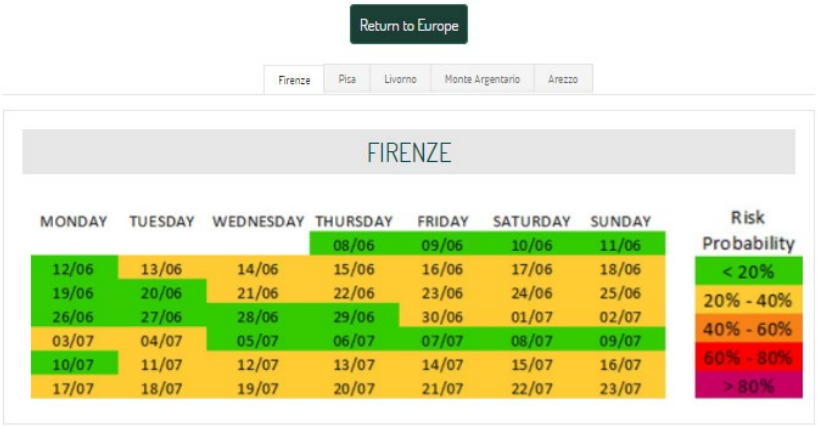

Figure 23: The example of the calendar with the probability forecast of exceeding the WBGT heat threshold for a specific working activity (in this case for an unacclimatized worker engaged in a moderate activity in the sun).

In the final version of the early heat warning system, this detailed probabilistic forecast will be available for all the localities included on the European maps. The web platform http://heatshield.biometeo.it/ demonstrates how this long term information can be combined with more detailed forecasts on the short term. The short term heat risk forecasts are represented by hourly and daily information for the current day up to the 5th following days. That product have been developed by the research group of the University of Florence involved in the HEAT-SHIELD project, based on a probabilistic model simulation. The hourly heat risk forecast for each day is provided for different working intensities and also includes recommendations for fluid intake and rest/break suggestions. The actual forecast is represented by 4 daily time bands (night: from 1 a.m. to 6 a.m; morning: from 7 a.m to 12 a.m.; afternoon: from 13 p.m. to 18 p.m; evening: from 19 p.m. to 24 p.m.) and allows to obtain forecasts for the WBGT heat stress risk for different metabolic rates (low, moderate, high and very high), accounting for acclimatized and unacclimatized workers exposed in the sun and in the shadow. Based on the available heat risk categories adapted from several international (especially American) organizations, five heat risk categories corresponding to different WBGT thresholds were adopted: No risk; low, medium, high and very high risks. This tailored heat risk information is provided in the form of maps (see Figure 24).

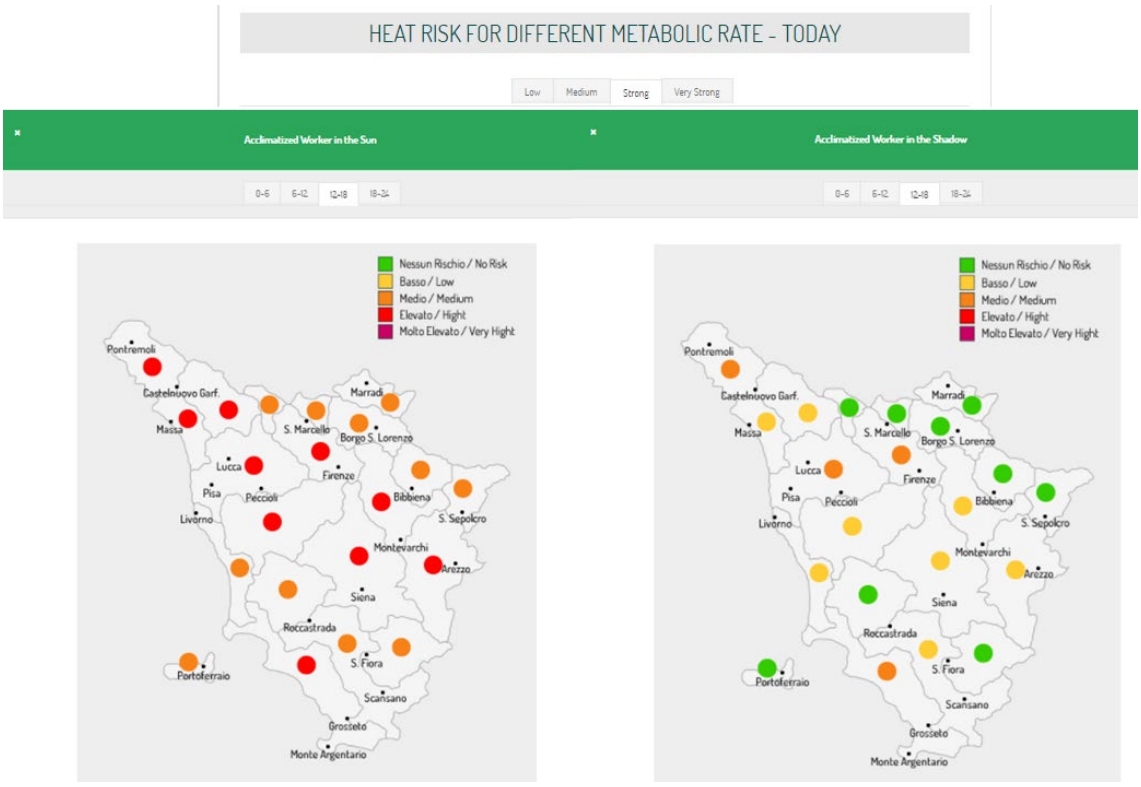

Figure 24: Heat risk categories of WBGT for the time band 13-18 for an acclimatized worker in the sun (left) and shadow (right), engaged in hard work (high metabolic rate) in the range band 12-18.

In addition to the subdaily products, the daily heat risk forecast provides daily maps of alert levels (see Figure 25) reporting a critical heat stress level when high level WBGT thresholds for acclimatized and unacclimatized workers engaged in moderate activity in the sun are reached. In this case, when the first day of a series or an isolated day has these features the warning is "caution", two consecutive days "alarm", from the third consecutive day onwards the warning is “emergency” (see Figure 25).

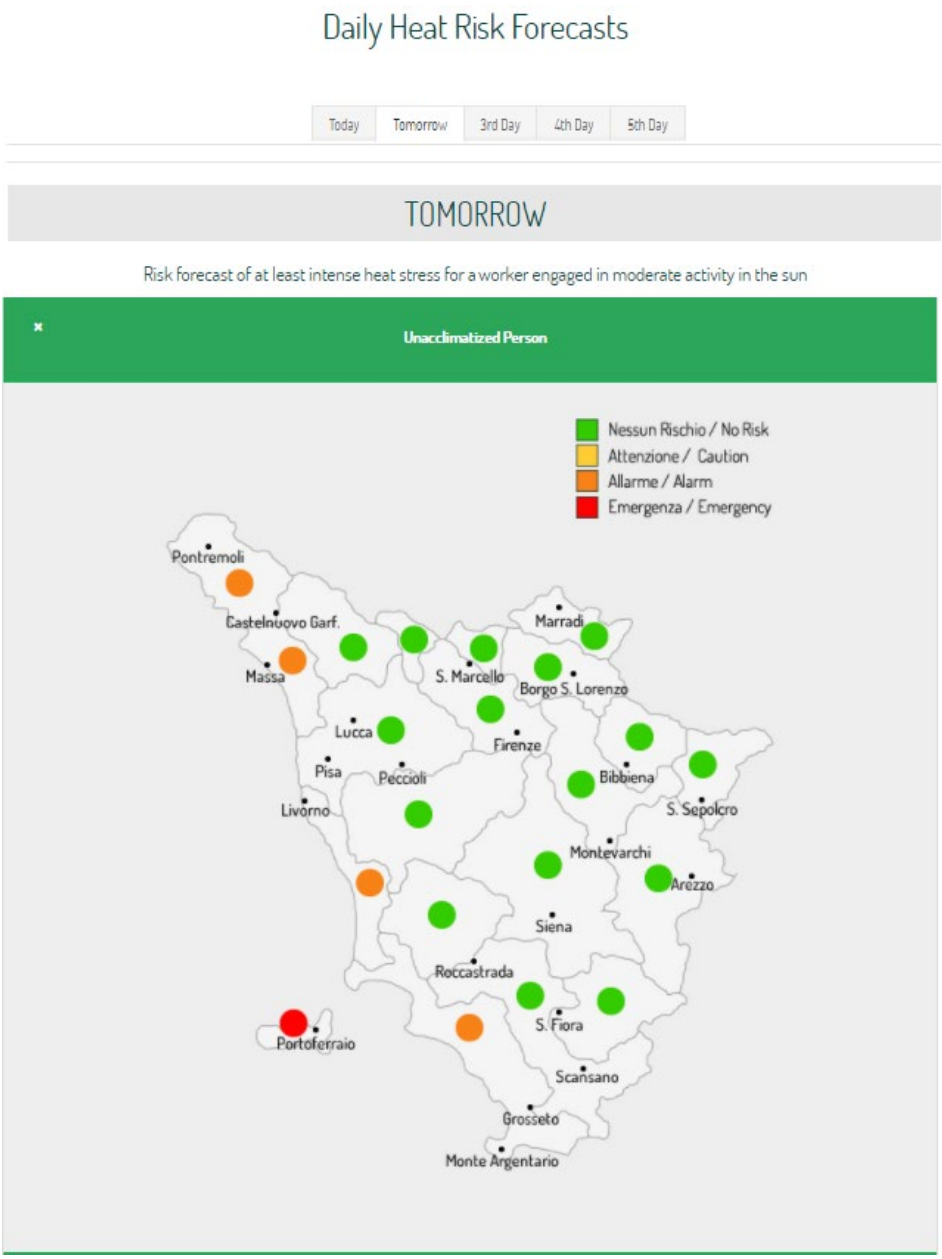

Figure 25: Daily heat risk forecast for unacclimatized workers engaged in moderate activity in the sun in different areas of Tuscany.

5 Discussion

The verification of $WBGT_{shade}$ forecasting contains metrics for assessing the association, reliability and accuracy. The association of the forecasts was validated with the correlation and we obtained that the hindcasts are associated up to 20 days to the observations. The accuracy was assessed by computing the fair CRPSS and RPSS and our results showed that the accuracy of the hindcasts remains positive until the14^{th} of the forecasts. The reliability was evaluated on the basis of the SPR. In the first 10 days of the forecasting period, the heat stress forecast system is underdispersive (i.e. overconfident), which makes the forecasts unreliable. In contrast, the correlation, RPSS and CRPSS imply a skilful forecast for the first 14 to 20 days. The time window remains short for getting good assessments of all verification metrics.

The correlation analysis shows that the 7-day running mean re-forecasts reach better scores than the daily ones. Similarly, the CRPSS of the 7-day running mean products tops the daily hindcasts. The scores of the averaged forecasts remain high for longer lead times. Therefore, there is a gain in the skill of the forecast, although the moving window reduces information (e.g. daily fluctuations). There are some stations with low association in the WBGT and air temperature at the very early stage of the model run. The correlation of the dew point temperature is low at stations in the Mediterranean region. Furthermore, they share high score region through the entire run, as at the 10^{th}, 15^{th} and 20^{th} day of forecast run. The high score regions in the correlation analysis of the dew point temperature differ from the other two parameters. Interestingly, the correlation of the dew point temperature is high in the Balkan region, away from the coast. It can be assumed that the coupled soil model of the ECMWF maintains skill even up to higher lead times. Low correlation scores at the shores and coasts might be due to coarse resolution of the global model. Therefore, not all coasts and islands are well represented in the grid cells.

Orth and Seneviratne (2014) came to similar results regarding the decrease of CRPSS along the lead time of a sub-seasonal forecast system of ECMWF. In their analysis, the CRPSS of the summer temperatures decay from a skill score of nearly 1 in the first week, to 0.6 in the second week, to 0.4 in the third week and to 0 in the fourth week. Those values are higher than the CRPSS of the air temperature computed in our study, probably due to the fact that we targeted maximum temperature. The spatial pattern of the CRPSS matches each other very well, with the highest skill scores in Central Europe throughout the skilful period. The values are low in Britain, Scandinavia and in the Mediterranean region. Buizza et al. (2008) computed the CRPSS of 2m temperature for the European domain of the ECMWF during the summer season. The resulting curve of CRPSS along the lead time starts around 0.55 and falls below 0 around9 days lead time. Considering the developing status of the models in those studies, the skill scores in the older ECMWF forecast system have to be lower throughout the lead time. Interestingly, all the CRPSS from those studies get slightly negative with

higher lead times as observed in our study. Hagedorn et al. (2012) verified winter ECMWF hindcasts of air temperature over Europe at 850 hPa, obtaining a decay curve of the CRPSS quite similar to the one shown in the current study. Both curves start with values more than 0.6 and gets close to 0 around 14 days lead time.

Fully reliable forecasts cannot be guaranteed with QM, because the correlation between raw forecasts and observations is not considered (Zhao et al., 2017). Therefore, skill scores of the forecast can be negative, if the raw forecast does not correlate significantly with the observations. Zhao et al. (2017) postulated that even if the forecasts and observations correlate significantly, QM cannot guarantee reliability and coherence in the post-processing of the forecasts.

Comparing the CRPSS of the WBGT, air and dew point temperatures, the skill scores of both input variables are significantly lower than the heat stress index during the first days. The increased skill score may be caused by the combination of both parameters, since some of the spatial pattern of the CRPSS of the WBGT can be derived out of the spatial patterns of the skill scores of the input variables. At 5 days lead time the CRPSS of the WBGT is low at many stations near the sea, as for the CRPSS of air temperature. There are a few stations in the subtropics, which show clearly negative CRPSS for the WBGT and for the dew point temperature, but this feature is not observed in the CRPSS of the air temperature. The CRPSS of the dew point is low in wide areas of the Mediterranean region. Nevertheless, the CRPSS of the WBGT is higher in Italy and Spain, which might be due to higher CRPSS of the air temperature in those areas.

At 10 days lead time of the forecast, the CRPSS of WBGT scores higher values over Central, Eastern Europe, Italy and Balkan region, similarly to air temperature. The higher CRPSS of dew point temperature are clustered in the Alps and Balkan region and far the sea, which are also retained in the CRPSS of WBGT. The computation of the WBGT is not linearly and only a third is directly linearly contributed by the air temperature. The other two thirds are made by more complex formulas, which might explain the resulting complex pattern. Compared to the correlation analysis, the CRPSS of the WBGT and air temperature do not imply a close connection as in the correlation. Differences in the cumulative distribution of both parameters may be responsible for the disagreements.

Fischer and Knutti (2012) validated projections of extreme health indicators and its contributing variables, obtaining that the contributing variables have larger uncertainties than the joint products. Furthermore, their investigations showed that such relationships exist even under present-day conditions. Model biases can be largely cancelled by joining variables, as shown by Casanueva et al. 2017 and the current study. A combination of the contributing variables may compensate failures of one of the variables if the other variable is better forecasted. Physical processes might lead to the mechanism which improves the forecast skill of the joint product as it was suggest by Fischer and Knutti (2012).

The ROC area score and the BSS were computed to evaluate the discrimination skill of the forecast system. The ROC area score could only be computed for areas where the hindcasts and observations have enough events. The higher the WBGT thresholds gets, the more south the limit of ROC area computation moves. Therefore, it is harder to apply higher thresholds in the forecasts even in subtropical Europe. The forecast system can make a discrimination until the 15th day for $WBGT_{shade}$ 30°C and until the 25th days for $WBGT_{shade}$ 24°C. ROC area scores lower than 0.5 are found near the limit of of $WBGT_{shade}$ threshold of 24°C. Surprisingly, high ROC area scores up to 15 days lead

time are reached by stations in the southeast of England, next to the spatial limit of the threshold. In the first days, there is a cluster of stations with ROC area scores lower than 0.5 in the Baltic Sea.

The BSS behaves similar as the ROC area score: The higher the thresholds is set, the lower the BSS and the noisier the lead-time dependent curves of the BSS becomes. The BSS with the threshold of $WBGT_{shade}$ 30°C reaches value of 0 around the 7th day whereas the BSS with the thresholds of $WBGT_{shade}$ 24°C and 27°C gets close to 0 around the 17th day. Similarly to the ROC area score, there is a cluster of high BSS in the south-east of England, which is predominant beyond the 10th day. In Germany, there many stations with negative BSS already at the 5th day, and then many more at the 10th day. The BSS get earlier to 0 than the ROC area score for all the thresholds, by a lag of about 5 days, thus the range which contains good discrimination along the lead time is limited. The skilful range is short for higher thresholds and widens to the half forecast horizon for moderate high thresholds.

6 Conclusions

The prototype of forecast system detecting heat stress risk has been setup and runs operationally. The index for identifying heat stress is the Wet Bulb Globe Temperature (WBGT). WBGT can be calculated for sunny conditions ($WBGT_{sun}$) and for shaded/sheltered conditions ($WBGT_{shade}$). The operational forecast is launched every Monday and Friday morning on the CSCS sever “Kesch” (see Chapter 4). A R-package (MeteoSwiss CAT) named as “mfcdaily” has been created in order to run R-codes operationally. The computational chain includes data loading, downscaling and bias correction, WBGT computation, assessing probabilities of exceeding thresholds, plotting diagrams and saving the outputs. So far, the whole setup of the “mfcdaily” is designed to generate the heat stress forecast with multiple cores. Nevertheless, this R-package can easily be extended for other extended and long range forecasts. The functions for the data loading, downscaling and bias correction can already be used to load major parameters such as air temperature, dew point temperature, radiation and wind speed. A restriction for new implementations can be the requested data structure of the input variables. In sum, the “mfcdaily” package generates the heat stress forecast by multiple cores. Furthermore, the package can already produce local sub-seasonal and seasonal forecasts for a few parameters. There is even a high potential to extend the forecast system with little effort.

The heat stress forecast system produces predictions of $WBGT_{sun}$ and $WBGT_{shade}$. The performance of predicting $WBGT_{shade}$ has been assessed according to verification metrics. The verification metrics consist of measures of association, reliability, accuracy and discrimination. The association was validated with the correlation, the reliability by the SPread to error Ratio (SPR), the accuracy by the Continuous Ranked Probability Skill Score (CRPSS), and the discrimination of certain thresholds was assessed with the Receiver Operating Characteristics area score (ROC area score) and the Brier Skill Score (BSS).

The forecast skills for validating the association, accuracy and discrimination are high in the first week and get lower in the following 1.5 weeks (see Chapter 3). However, the forecast is unreliable and overconfident during the first week. After the first week, the reliability improves and reaches an optimal level. The forecast system produces predictions for $WBGT_{shade}$ and $WBGT_{sun}$, in particular focusing on the exceedance of specific thresholds (24°C for $WBGT_{shade}$ and 27°C and 30°C for $WBGT_{sun}$). The discrimination of certain $WBGT_{shade}$ thresholds succeeds for the first 1-1.5 weeks for 30°C threshold and for 2-3 weeks for 24°C and 30°C thresholds. Since the $WBGT_{sun}$ is larger than $WBGT_{shade}$, the discrimination of $WBGT_{sun}$ can surely be guaranteed up to the third week. Furthermore, the results have been compared to findings of other studies (see Chapter 5). The performed verification of the air temperature hindcasts agrees with published literature on the predictability and forecast performance. In sum, the skill analysis imply that the forecast system for $WBGT_{shade}$ remain skilful for the first 2.5 weeks. There are regional differences in the skill, best forecast performance for

the summer months are reached by stations in the Balkan region. Furthermore, the forecast $WBGT_{shade}$ outperforms the forecasts of the individual input parameters.

Abbreviations

a.m.	ante meridiem.
AUC	Area Under the Curve
BS	Brier Skill
BSS	Brier Skill Score
CAT	Climate Analysis Tool
Cor	Correlation
CRPS	Continous Ranked Probability Skill
CRPSS	Continous Ranked Probability Skill Score.
ECDF	Empirical Cumulative Distribution function
ECMWF	European Centre for Medium-range Weather Forecasts
e.g.	exempli gratia
h	hour
hPa	hectopascal
IFS	Integrated Forecasting System
p.m.	post meridiem
QM	Quantile Mapping
ROC	Receiver Operating Characteristics
RPS	Ranked Probability Skill
RPSS	Ranked Probability Skill Score
SPR	Spread to error ratio
WBGT	Wet Bulb Globe Temperature
WP5	Workpackage 5
UTC	Coordinated Universal Time

List of figures

References

Baran S., and S. Lerch, 2015: Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society, 141* (691), 2289-2299.

Bedia J., N. Golding, A. Casanueva, M. Iturbide, C. Buontempo, and J. M. Gutiérrez, 2017: Seasonal predictions of Fire Weather Index: Paving the way of their operational applicability in Mediterranean Europe. *Climate Services*.

Bernard T. E., and M. Pourmoghani, 1999: Prediction of workplace wet bulb global temperature. *Applied Occupational and Environmental Hygiene, 14*, 126-134.

Bhend J., I. Mahlstein, and M. A. Liniger, 2017: Predictive skill of climate indices compared to mean quantities in seasonal forecasting. *Quarterly Journal of the Royal Meteorological Society, 143* (702), 184-194.

Bird R. B., W. E. Stewart, and E. N. Lightfoot, 2007: *Transport phenomena*. John Wiley & Sons.

Broecker J., 2012: Probability forecasts, in Forecast Verification: A Practitioner's Guide in Atmospheric Science (2nd ed., chap. 8), edited by I. T. Jolliffe and D. B. Stephenson, 119-139. Oxford, West Sussex and Hoboken: *John Wiley & Sons*, *Ltd*.

Buizza R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society, 125* (560), 2887-2908.

Buizza R., P. L. Houtekamer, Z. Toth, G. Pellerin, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCP global ensemble prediction systems. *Monthly Weather Review, 133*, 1076-1097.

Buizza R., R. Hagedorn, and F. Vitart, 2008: *Recent results in ensemble prediction*.

Budd G. M., 2008: Wet-bulb globe temperature (WBGT) – its history and it limitations. *Journal of Sicence and Medicine in Sport, 11* (1), 20-32.

Casanueva A., S. Kotlarski, M. Liniger, and A. Fischer, 2017: Heat stress indices over Europe for current climate conditions and for future conditions based on climate model simulations. *HEAT-SHIELD Deliverable 1.2*.

Crochemore L., M.-H. Ramos, and F. Pappenberger, 2016: Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences, 20*, 3601-3618.

De Freitas C. R., E. A. Grigorieva, 2015: A comprehensive catalogue and classification of human thermal climate indices. *International Journal of Biometeorology, 59*, 109-120.

Dee, D. P., S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society, 137*, 553–597.

Déqué M., 2012: Deterministic forecasts of continuous variables, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd ed., chap. 8), edited by I. T. Jolliffe and D. B. Stephenson, 77-94. Oxford, West Sussex and Hoboken: John Wiley & Sons, Ltd.

Fischer E. M. and R. Knutti, 2012: Robust projections of combined humidity and temperature extremes. *Nature Climate Change, 3*, 126-130.

Ferranti L., T. N. Palmer, F. Molteni, and E. Klinker,1990: Tropical-Extratropical Interaction Associated with the 30-60 Day Oscillation and Its Impact on Medium and Extended Range Prediction. *Journal of the Atmospheric Sciences, 47* (18), 2177-2199.

Ferro C. A., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and ranked probability scores. *Meteorological Applications, 15*, 19-28.

Ferro C. A., 2014: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society, 140*, 1917-1923.

Gneiting T., 2014: *Calibration of medium-range weather forecasts*. European Centre of Medium-Range Weather Forecasts.

Hagedorn R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society, 138* (668), 1814-1827.

Haiden T., M. Janousek, P. Bauer, J. Bidlot, M. Dahoui, L. Ferranti, F. Prates, D. S. Richardson, and F. Vitart., 2015: Evaluation of ECMWF forecasts, including 2013–2014 upgrades. *ECMWF technical memorandum, 765*.

Haiden T., M. Janousek, J. Bidlot, L. Ferranti, F. Prates, F. Vitart, P. Bauer, and D. S. Richardson, 2016: Evaluation of ECMWF forecasts, including the 2016 upgrade. *ECMWF technical memorandum, 792*.

Ho C. K., E. Hawkins, L. Shaffrey, J. Brocker, L. Hermanson, J. M. Murphy, D. M. Smith, and R. Eade, 2013: Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. *Geophysical Research Letters, 40*, 5770-5775.

ISO, 1989: Hot environments - Estimation of the heat stress on working man, based on the WBGT-index (wet bulb globe temperature). ISO Standard 7243. Geneva: *International Standards Organization*; 1989a.

Jolliffe I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atomspheric Science* (2nd ed.). Oxford, West Sussex and Hoboken: John Wiley & Sons, Ltd.

Khajehei S., and H. Moradkhani, 2017: Towards an improved ensemble precipitation forecast: A probabilistic post-processing approach. *Journal of Hydrology, 546*, 476-489.

Lemke B., and T. Kjellstrom, 2012: Calculating workplace WBGT from meteorological data. *Industrial Health, 50*, 267-278.

Lerch S., and T. L. Thorarinsdottir, 2013: Comparison of non-homogenous regression models for probabilistic wind speed forecasting. *Tellus Series a Dynamic Meteorology and Oceanography, 65*.

Liljegren J. C., R. A. Carhart, P. Lawday, S. Tschopp, and R. Sharp, 2008: Modeling the wet bulb globe temperature using standard meteorological measurements. *Journal of occupational and environmental hygiene, 5* (10), 645-655.

Lundgren K., K. Kuklane, G. A. O. Chuansi, and I. Holmer, 2013: Effects of heat stress on working populations when facing climate change. *Industrial Health, 51* (1), 3-15.

Kingsolver J. G., S. E. Diamond, and L. B. Buckley, 2013: Heat stress and the fitness consequences of climate change for terrestrial ecotherms. *Functional Ecology, 27.6*, 1415-1423.

Klein Tank A. M. G, J. B. Wijingaard, G. P. Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessermoulin, G. Müller-Westermeier, M. Tsanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, M. Maugeri, A. Leitass, A. Bikantis, R. Aberfeld, A. F. V. van Engelen, E. Forland, M. Mietus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J. Antonio López, B. Dahlström, A. Moberg, W. Kirchhofer, A. Ceylan, O. Packaliuk, L. V. Alexander, and P. Petrovic, 2002: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology, 22* (12), 1441-1453.

McPherson M. J., 2008: *Surface ventilation and environmental engineering*, 2nd Ed. Chapter 17. Physiological reactions to climatic conditions. Mine Ventilation Services Inc., Clovis. http://www.mvsengineering.com/index.php?cPath=25.

Müller V., C. Gray, and K. Kosec, 2014: Heat stress increases long-term human migration in rural Pakistan. *Nature climate change, 4.3*, 182-185.

Müller W. A., C. Appenzeller, F. J. Doblas-Reyes, and M. A. Liniger, 2005: A Debiased Ranked Probability Skill Score to Evaluate Probabilistic Ensemble Forecasts with Small Ensemble Size. *Journal of Climate, 18*, 1513-1523.

Murphy A. H., 1993: What is a Good Forecast? An Essay on the Goodness in Weather Forecasting. *Weather Forecasting, 8*, 281-293.

Orth R., and S. I. Seneviratne, 2014: Using soil moisture forecasts for sub-seasonal summer temperature predictions in Europe. *Climate Dynamics, 43* (12), 3403-3418.

Posselt R., R. W. Mueller, R. Stöckli, J. Trentmann, 2012: Remote sensing of solar surface radiation for climate monitoring - the CM-SAF retrieval in international comparison. *Remote Sensing of Environment, 118*, 186–198.

Posselt R., R. Mueller, J. Trentmann, R. Stockli, M. A. Liniger, 2014: A surface radiation climatology across two Meteosat satellite generations. *Remote Sensing of Environment, 142*, 103-110.

R Core Team, 2016: *R: A language and environment for statistical computing*. R foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rajczak J., S. Kotlarski, and C. Schär; 2016a: Does Quantile Mapping of Simulated Precipitation Correct for Biases in Transition Probabilities and Spell Lengths? *Journal of Climate, 29*, 1605-1615.

Rajczak J., S. Kotlarski, N. Salzmann, and C. Schär, 2016b: Robust climate scenarios for sites with sparse observations: a two-step bias correction approach. *International Journal of Climatology, 36* (3), 1226-1243.

Schepen A., and Q. J. Wang, 2014: Ensemble forecasts of monthly catchment rainfall out to long lead times by postprocessing coupled general circulation model output. *Journal of Hydrology, 519*, 2920-2931.

Schwierz C., C. Appenzeller, H. C. Davies, M. A. Liniger, W. Müller, T. F. Stocker, and M. Yoshimore, 2006: Challenges posed by and approaches to the study of seasonal-to-decadal climate variability. *Climate Change, 79*, 31-63.

Smith L. A., H. Du, E. B. Suckling, and F. Niehörster, 2015: Probabilistic skill in ensemble seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society, 141* (689), 1085-1100.

Teixeira E. I., G Fischer, H. van Velthuizen, C. Walter, and F. Ewert,2013: Global hot-spots of heat stress on agricultural crops due to climate change. *Agricultural and Forest Meteorology, 170*, 206-215.

Themessl M. J., A. Gobiet, and A. Leuprecht, 2011: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology, 31* (10), 1530-1544.

Verkade J. S., J. D. Brown, P. Reggiani, and A. H. Weerts, 2013: Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology, 501*, 73-91.

Vitart F., 2014: Evolution of ECMWF sub-seasonal forecast skill scores. *Quarterly Journal of the Royal Meteorological Society, 140* (683), 1889-1899.

Vitart F., G. Balsamo, R. Buizza, L. Ferranti, S. Keeley, L. Magnusson, F. Molteni, A. Weisheimer, 2014: Sub-seasonal predictions. *ECMWF Research Department Technical Memorandum, 734* (47).

Vitart F., C. Ardilouze , A. Bonet, A. Brookshaw, M. Chen, C. Codorean, ... & H. Hendon, 2016: The sub-seasonal to seasonal prediction (S2S) project database. *Bulletin of the American Meteorological Society*, (2016).

Weigel A. P., F. K. Chow, and M. W. Rotach, 2007a: The effect of mountainous topography on moisture exchange between the ‘surface’ and the free atmosphere. *Boundary Layer Meteorology, 125*, 227-244.

Weigel A. P., M. A. Liniger, and C. Appenzeller, 2007b: The discrete brier and ranked probability skill scores. *Monthly Weather Review, 135*, 118-124.

Weigel A. P., M. A. Liniger, and C. Appenzeller, 2007c: Generalization of the discrete Brier and ranked probability skill scores for weighted multimodel ensemble forecasts. *Monthly Weather Review, 135*, 2778-2785.

Weigel A. P., M. A. Liniger, and C. Appenzeller, 2008a: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society, 134*, 241-260.

Weigel A. P., M. A. Liniger, and C. Appenzeller, 2008b: Seasonal Ensemble Forecasts: Are Recalibrated Single Models better than Multimodels? *Monthly Weather Review, 137*, (4), 1460-1479.

Weigel A. P, D. Baggenstos, and M. A. Liniger, 2008c: Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review, 136*, 5156-5182.

Weigel A. P., 2012: Ensemble forecasts, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd ed., chap. 8), edited by I. T. Jolliffe and D. B. Stephenson, 141-166. Oxford, West Sussex and Hoboken: John Wiley & Sons, Ltd.

Wilks D. S., and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts, *Monthly Weather Reviews, 135* (6), 2379-2390.

Wilks D. S., 2011: *Statistical Methods in the Atmospheric Science* (3rd ed.). London: International Geophysics Series, Vol. 100, Academic Press Inc.

Williams R. M., C. A. T. Ferro, and F. Kwasniok, 2014: A comparison of ensemble post-processing methods for extreme events, *Quarterly Journal of the Royal Meteorological Society, 140*, 1112-1120.

Yaglou C. P., and D. Minard, 1956: *Prevention of heat causalities at marine corps training centres*. Armed Services Technical Information Agency Document Service Centre AD099920.

Zhao T., J. Bennett, Q. J. Wang, A. Schepen, A. Wood, D. Robertson, and M. H. Ramos, 2017: How suitable is quantile mapping for post-processing GCM precipitation forecasts? *Journal of Climate, 30*, 3185-3196.

Acknowledgement

This project was part of an internship under the supervision of Christoph Spirig and Mark Liniger. Wewould like to express our thanks to MeteoSwiss and HEAT-SHIELD for supporting this work. HEAT-SHIELD is funded by EU Horizon 2020 research and innovation programme under grant agreement No 668786.

We thank all our colleagues from MeteoSwiss, especially Ana Casanueva, Sven Kotlarski and Jonas Bhend, and HEAT-SHIELD who were involved in the realisation of this project to setup a prototype of a forecasting system for heat stress.

Also we would like to thank Marco Stoll and Matteo Buzzi who reviewed this report.

A Appendix 1: WBGT computation

Heat stress is obtained by the popular Wet Bulb Globe Temperature (WBGT). This heat measure was introduced in the 50's during a campaign to investigate heat illness on soldiers in training camps of the United States Army and Marine Corps (Budd 2008). A measuring device was constructed to scale the heat stress more appropriately. Since the WBGT cannot be directly extracted from climate models, the WBGT has to be estimated with meteorological parameters which are available from the model data and measured by meteorological stations. The approaches used in this study to determine the WBGT are presented hereafter.

Lemke and Kjellstrom (2012) compared published methods of computing indoor and outdoor WBGT from standard weather and climate data. They recommended to use the implementation of Bernard and Pourmoghani (1999) for the calculation of the indoor WBGT. For computing the outdoor WBGT, Lemke and Kjellstrom (2012) advised to take the formula derived by Liljegren et al. (2008). The indoor or in the shade WBGT ($\mathrm{WBGT_{shade}}$) is defined by Bernard and Pourmoghani (1999) as:

$$\mathrm{WBGT_{shade}} = 0.67 * \mathrm{T_{pwb}} + 0.33 * \mathrm{T_a} - 0.048 * \log_{10} v * \left(T_a - T_{pwb}\right) \quad (4)$$

, where the psychrometric wet bulb temperature is denoted as $\mathrm{T_{pwb}}$, v stands for the horizontal wind speed and $\mathrm{T_a}$ is the air temperature. This formula is valid for wind speed in the range 0.3-3ms^{-1} and if the wind speed is approximated to 1 ms^{-1} (slow walk), then the equation reduces to:

$$\mathrm{WBGT_{shade}} = 0.67 * \mathrm{T_{pwb}} + 0.33 * \mathrm{T_a} \quad (5)$$

, which is the implementation followed in the forecast system for the $\mathrm{WBGT_{shade}}$.

McPherson (2008) formulated the equation for calculating the $\mathrm{T_{pwb}}$ by iteration as:

$$1556 * \mathrm{e_d} - 1.484 * \mathrm{e_d} * \mathrm{T_{pwb}} - 1556 * \mathrm{e_w} + 1.484 * \mathrm{e_w} * \mathrm{T_{pwb}} + 1010\left(\mathrm{T_a} - \mathrm{T_{pwb}}\right) = 0 \quad (6)$$

, where the air temperature is denoted as $\mathrm{T_a}$. The vapour pressure at the dew point $\mathrm{e_d}$ and the saturation water pressure $\mathrm{e_w}$ are defined as:

$$\mathrm{e_d} = 6.106 * \mathrm{e}^{\left[\frac{17.27*\mathrm{T_d}}{(237.3+\mathrm{T_d})}\right]} \quad (7)$$

$$\mathrm{e_w} = 6.106 * \mathrm{e}^{\left[\frac{17.27*\mathrm{T_{pwd}}}{(237.3+\mathrm{T_{pwd}})}\right]} \quad (8)$$

Thus, $\mathrm{T_{pwb}}$can be calculated by iteration with the Equations 6, 7 and 8.

Lemke and Kjellstrom (2012) suggest to use the implementation of Liljegren et al. (2008) for the calculation of the WBGT outdoor or in the sun ($\mathrm{WBGT_{sun}}$). Furthermore, the WBGT computation as

stated by Liljegren et al. (2008) underestimates the targeted values in sheltered conditions, whereas the implementation of Bernard and Pourmoghani (1999) does a better job. The basic formula of the outdoor WBGT is referred from Yaglou and Minard (1956), which introduced it as:

$$\mathrm{WBGT_{sun}} = 0.7 * \mathrm{T_{nwb}} + 0.2 * \mathrm{T_g} + 0.1 * \mathrm{T_a} \tag{9}$$

, where $\mathrm{T_{nwb}}$ is defined as the natural wet bulb temperature; $\mathrm{T_g}$ stands for the globe temperature. That formula is suitable for cloudy and sunny conditions, because $\mathrm{T_g}$ contains the direct and diffuse components of the sunlight radiation. The method of Liljegren et al. (2008) calculates the $\mathrm{T_g}$ as:

$$\mathrm{T_g^4} = \frac{1}{2}(1+\varepsilon_a) * \mathrm{T_a^4} - \frac{\mathrm{h}}{\varepsilon_g * \sigma}(\mathrm{T_g} - \mathrm{T_a}) + \frac{\mathrm{S}}{2 * \varepsilon_g * \sigma}(1-\alpha_g)\left[1 + \left(\frac{1}{2 * \cos\theta} - 1\right) * \mathrm{f_{dir}} + \alpha_{sfc}\right] \tag{10}$$

, where the globe emissivity ε_g is set to $\varepsilon_g = 0.95$; ε_a is the emissivity of the atmosphere calculated from air temperature and relative humidity; θ stands for the solar zenith angle and σ for Stefan-Boltzmann constant; $\mathrm{f_{dir}}$ represents the fraction of the total horizontal solar irradiance S due to the direct beam of the sun which we set to 0.8; the surface albedo α_{sfc} was set to $\alpha_{sfc} = 0.4$ and the albedo of the globe α_g to $\alpha_g = 0.05$; h stands for the convective heat transfer for the flow around a sphere calculated as $\mathrm{h} = \mathrm{Nu} * k/D$, where *D* is the diameter of the globe (50mm), *k* is the thermal conductivity of the air and Nu is the Nusselt Number.

The natural bulb temperature $\mathrm{T_{nwb}}$ is approximated as:

$$\mathrm{T_{nwb}} = \mathrm{T_a} - \frac{\Delta\mathrm{H}}{\mathrm{c_p}}\frac{\mathrm{M_{H20}}}{\mathrm{M_a}}\left(\frac{\mathrm{Pr}}{\mathrm{Sc}}\right)^{0.56}\left(\frac{\mathrm{e_{nwb}} - \mathrm{e_a}}{\mathrm{P} - \mathrm{e_{nwb}}}\right) + \frac{\Delta\mathrm{F_{net}}}{\mathrm{A} * \mathrm{h}} \tag{11}$$

, where $\Delta\mathrm{H}$ stands for the heat of evaporation, $\mathrm{c_p}$is the specific heat capacity of dry air at constant pressure; $\mathrm{M_{H20}}$ is the molecular weight of water vapour; $\mathrm{M_a}$ ist he molecular weight of dry air; the Prandtl Number Pr can be calculated as $Pr = \mathrm{c_p}\mu/\mathrm{k}$ (μ for the viscosity and k for the thermal conductivity); the Schmidt number Sc is denoted as $\mathrm{Sc} = \mu/(\rho\mathcal{D})$; $\mathcal{D}$ stands for the diffusivity and ρ for the density; $\mathrm{e_a}$ and $\mathrm{e_{nwb}}$ are the water vapour pressure in the atmosphere and in the surface of the wick, respectively; h is, in this case, the heat transfer coefficient for a long cylinder in cross flow; the parameter $\Delta\mathrm{F_{net}}$ is the net radiant heat flux from the environment to the wick and A is the area of the wick (diameter of 7mm and longitude of 25.4mm are assumed).

The detail code for the computation of the WBGT in the shade and in the sun is uploaded on https://github.com/anacv/HeatStress.

B Appendix 2: Skill scores

Correlation

A widely in use measure in weather and climate forecasting is the correlation coefficient, which has invariance properties (Déqué 2012). The correlation has dimensionless and positively oriented characteristics. Shifts and multiplication with constant values do not affect the correlation, because of invariance to changed scales. Therefore, bias correction and linear post-processing conserve the correlation.

In this study the ensemble mean correlation (Pearson) was computed with the “EnsCorr”-function from the R-project package “easyVerification”. The function uses the ensemble mean and the observations to compute the correlation. A score value of 0 implies a weak connection, whereas -1 and 1 means a perfectly anti-correlated respectively perfectly correlated association. The skill score is likely affected by trends and memory effects. The observation and hindcast datasets contain 20 years. The recent climate change can be found in both datasets. Furthermore, the lead time of 46 days is long enough that the correlation in the annual cycle has an impact. The effects of climate change and annual cycles on the correlation was not investigated in this study, but these effects were considered in the interpretation of the correlation scores and as well for other verification metrics.

Fair RPSS and BSS

For binary outcomes of discrete probability forecasts, the Brier Score (BS) is an adequate verification measure (Weigel, 2012). The BS is defined as:

$$\mathrm{BS} = \frac{1}{n}\sum_{t=1}^{n}(\hat{p}_t - y_t) \tag{12}$$

, where $\hat{p}_t$ is the probability of the event for the tth forecast, y_t denotes tth observation of the event. A skill score can be derived of the BS by using the climatology as reference:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{Cl}}} \quad \text{with} \quad \mathrm{BS_{Cl}} = \frac{1}{n}\sum_{t=1}^{n}(c - y_t) \tag{13}$$

The Ranked Probability Score (RPS) as a multicategorical generalisation of the BS is defined as (Weigel, 2012):

$$\mathrm{RPS} = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\left(\widehat{P}_{t,k} - Y_{t,k}\right)^2 \tag{14}$$

$$\text{with } \widehat{P}_{t,k} = \sum_{l=1}^{k}\hat{p}_{t,l}\,,\ \widehat{Y}_{t,k} = \sum_{l=1}^{k}\hat{y}_{t,l} \text{ and } C_k = \sum_{l=1}^{k} c_l \tag{15}$$

, where K is the number of categories, c_k denotes the climatological probability in the category k; $\hat{p}_{t,k}$ denotes the probability of the tth forecasts for the kth category; $\widehat{P}_{t,k}$ and $\widehat{Y}_{t,k}$ are vectors, which contain the tth cumulative forecasts and observations for the kth category; the kth category of the cumulative climate distribution is denoted as C_k. The RPSS is calculated by using the climatology as reference:

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS_{Cl}}} \quad \text{with} \quad \mathrm{RPS_{Cl}} = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=1}^{K}\left(C_k - Y_{t,k}\right)^2 \tag{16}$$

Weigel et al. (2007b) reformulated the RPSS on the basis of the new reference strategy of Müller et al. (2005). The conventional climatological reference score is corrected by the term D:

$$\mathrm{RPSS_D} = 1 - \frac{\mathrm{RPS_m}}{\mathrm{RPS_{Cl}} + D} \tag{17}$$

, where m indicate that the forecasts uses ensemble members. The term D is definite as:

$$D = \frac{1}{m}\sum_{k=l}^{K}\sum_{l=1}^{k}\left[c_l\left(1 - c_l - 2\sum_{i=l+1}^{k} c_i\right)\right] \tag{18}$$

, where c denotes the climatological probability of an event and K is the number of forecast categories. If K is equiprobable, then D can be simplified to:

$$D = \frac{K^2 - 1}{6Km} \tag{19}$$

The BBS can be similarly reformulated:

$$\mathrm{BSS_D} = 1 - \frac{\mathrm{BS_m}}{\mathrm{BS_{Cl}} + \frac{1}{m}c(1-c)} \tag{20}$$

The FRPSS is only available if the ensemble size is not equal in every verification sample (Weigel et al., 2008c). Furthermore, the output of several ensemble prediction systems has to be combined to a weighted multi-model ensemble (Weigel et al., 2007c). Ferro et al. (2008) introduced an unbiased estimator to conserve the RPS from the ensemble size effect as:

$$\mathrm{E(RPS_M)} \cong \mathrm{RPS_M} - \frac{M - m}{M(m-1)n} \times \sum_{t=1}^{n}\sum_{k=1}^{K}\widehat{P}_{t,k}\left(1 - \widehat{P}_{t,k}\right) \tag{21}$$

, where $\mathrm{E(RPS_M)}$ stands for RPS of a forecast with an ensemble size of M instead of m; $\cong$ denotes that is estimated without bias from. For the binary event/no-event case of the BS, the formula gets simpler:

$$E(BS_M) \cong BS_M - \frac{M-m}{M(m-1)n} \times \sum_{t=1}^{n}\sum_{k=1}^{K} \hat{p}_t(1-\hat{p}_t) \tag{22}$$

The CRPS was similarly treated by Ferro et al. (2008). The RPSS for infinite ensemble size can be derived by taking $M \to \infty$ in the equations above:

$$E(RPSS_\infty) = 1 - \frac{E(RPS_\infty)}{RPS_{Cl}} \cong 1 - \frac{RPS_m}{RPS_{Cl}} + \frac{\sum_{t=1}^{n}\sum_{k=1}^{K} \hat{P}_{t,k}\left(1-\hat{P}_{t,k}\right)}{RPS_{Cl}(m-1)n} \tag{23}$$

Fair CRPSS

The continuous ranked probability score (CRPS) is an Integral of the squared differences between the cumulative forecast and observation distribution (Weigel, 2012). The CRPS is defined as (Ferro et al., 2008):

$$CRPS_m = \frac{1}{n}\sum_{t=1}^{n} \int_{-\infty}^{\infty} \left[\hat{Q}_T(u) - I_T(u)\right]^2 du \tag{24}$$

, where $\hat{Q}_T(u)$ is the proportion of m ensemble members not exceeding the threshold u at the time step T; $I_T(u)$ denotes if the observation exceeded the threshold u at the time step T ($I_T(u) = 0$) or if not (($I_T(u) = 1$)). Similar to the BS and RPS, a fair score can be derived for the CRPS:

$$CRPS_M = \frac{m(M+1)}{M(m+1)} CRPS_m \tag{25}$$

The infinite ensemble sized CRPS can be computed by taking $M \to \infty$. The Fair Continuous Ranked probability skill score (CRPSS) takes the CRPSS of the climatology ($CRPS_{Cl}$) as reference:

$$E(CRPSS_\infty) = 1 - \frac{E(CRPS_\infty)}{CRPS_{Cl}} \cong 1 - \frac{m}{(m+1)CRPS_{Cl}}\left(\frac{1}{n}\sum_{t=1}^{n} \int_{-\infty}^{\infty} \left[\hat{Q}_T(u) - I_T(u)\right]^2 du\right) \tag{26}$$

Fair SPR

The validation of the forecast reliability is often performed with the Spread to Error Ratio. The reliability is the agreement of the forecast probabilities to the observed frequencies. The SPR is defined as the ensemble size corrected ratio of the time mean standard deviation to the root mean squared error (Weigel, 2012; Ho et al., 2013):

$$SPR = \sqrt{\frac{m+1}{m}} \frac{\sigma_e}{RMSE} \tag{27}$$

$$\text{with} \quad \sigma_e^2(\tau) = \frac{m}{(m+1)} MSE(\tau) \tag{28}$$

The forecast is considered overconfident if the ratio is smaller than 1. In this case, the ensemble spread is too narrow to estimate the observed uncertainties. Whereas a ratio larger than 1 means that the forecast produces an overdispersion.

ROC area score

The Receiver Operating Characteristic curve (ROC) is used as a discrimination metrics for dichotomous forecasts. A common output statistics for the ROC is the Area Under the Curve (AUC hereafter). A perfect association is reached if the score is 1. A score of already 0.5 means a zero association.

$$\mathrm{AUC} = -\sum_{n=1}^{n} \mathrm{H(n)[F(n) - F(n-1)]} = \frac{1}{N_0 N_1}\left[\sum_{n=1}^{N} \mathrm{ny(n)} - \frac{N_1(N_1+1)}{2}\right] \quad (29)$$

MeteoSchweiz
Operation Center 1
CH-8044 Zürich-Flughafen
T +41 58 460 99 99
www.meteoschweiz.ch

MeteoSvizzera
Via ai Monti 146
CH-6605 Locarno Monti
T +41 58 460 97 77
www.meteosvizzera.ch

MétéoSuisse
7bis, av. de la Paix
CH-1211 Genève 2
T +41 58 460 98 88
www.meteosuisse.ch

MétéoSuisse
Chemin de l'Aérologie
CH-1530 Payerne
T +41 58 460 94 44
www.meteosuisse.ch