


RESEARCH ARTICLE

The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures

Chiem van Straaten^{1,2}  | Kirien Whan¹ | Dim Coumou^{2,3} | Bart van den Hurk^{2,4} | Maurice Schmeits¹

¹Department of Weather and Climate Modelling, Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

²Department of Water and Climate Risk, Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

³Potsdam Institute for Climate Impact Research, Earth System Analysis, Potsdam, Germany

⁴Deltares, Delft, The Netherlands

Correspondence

C. van Straaten, Department of Weather and Climate Modelling, Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands.
Email: chiem.van.straaten@knmi.nl

Funding information

This study is part of the open research programme Aard- en Levenswetenschappen, project number ALWOP.395, which is financed by the Dutch Research Council (NWO)

Abstract

The succession of European surface weather patterns has limited predictability because disturbances quickly transfer to the large-scale flow. Some aggregated statistics, however, such as the average temperature exceeding a threshold, can have extended predictability when adequate spatial scales, temporal scales and thresholds are chosen. This study benchmarks how the forecast skill horizon of probabilistic 2-m temperature forecasts from the subseasonal forecast system of the European Centre for Medium-Range Weather Forecasts (ECMWF) evolves with varying scales and thresholds. We apply temporal aggregation by rolling-window averaging and spatial aggregation by hierarchical clustering. We verify 20 years of re-forecasts against the E-OBS dataset and find that European predictability extends at maximum into the fourth week. Simple aggregation and standard statistical post-processing extend the forecast skill horizon with two and three skilful days on average, respectively. The intuitive notion that higher levels of aggregation capture large-scale and low-frequency variability and can therefore tap into extended predictability holds in many cases. However, we show that the effect can be saturated and that there exist regional optimums beyond which extra aggregation reduces the forecast skill horizon. We expect such windows of predictability to result from specific physical mechanisms that only modulate and extend predictability locally. To optimize subseasonal forecasts for Europe, aggregation should thus be limited in certain cases.

KEYWORDS

ensemble forecasts, statistical post-processing, verification, forecast skill horizon

1 | INTRODUCTION

Extending skilful weather predictions beyond two weeks and into the subseasonal range is of great importance for humanitarian concerns such as safeguarding crop harvests

and preventing energy shortages (Coughlan de Perez *et al.*, 2015; Grams *et al.*, 2017; Guimares Nobre *et al.*, 2019). These efforts are propelled by the intuition that extreme, large-scale events can potentially be predicted in advance (Vitart and Robertson, 2018). However, producing skilful

forecasts of such large-scale events remains notoriously difficult.

The atmosphere is a dynamical system that varies on many spatio-temporal scales. Its succession of instantaneous states is deterministic but chaotic. Small disturbances can evolve to larger scales, growing in such a way that they overwhelm signals that were originally present. This means that deterministic atmospheric forecasts draw on predictability arising from initial conditions but will at some point become inaccurate (Lorenz, 1969). The forecast error will then relate to the total variance in the predicted phenomenon. The saturation of forecast error occurs most quickly at the finest scales, whereas at larger scales of motion variations are observed that have the potential for predictability at longer lead times (Hoskins, 2013; Privé and Errico, 2015; Ying and Zhang, 2017; Toth and Buizza, 2019).

These potentially predictable variations can be internal to the atmosphere, or they can form in interaction with other components of the Earth system. Internally, the mid-latitude tropospheric variability is often dominated by a few large-scale patterns that recur and evolve into each other (Vautard, 1990; Hannachi *et al.*, 2017), which are associated with predominant weather types on the ground (Grotjahn *et al.*, 2016). Variability in Europe is also steered into specific regions of phase space by slow-moving components such as Atlantic sea-surface temperatures (Czaja and Frankignoul, 2002), snow cover (Orsolini *et al.*, 2013; Henderson *et al.*, 2018), soil moisture (Prodhomme *et al.*, 2016), the stratosphere (Baldwin and Dunkerton, 2001; Tripathi *et al.*, 2015) and tropical variability such as the Madden–Julian Oscillation (MJO) (Cassou, 2008; Vitart, 2017; Yadav and Straus, 2017; Lin and Brunet, 2018). These components often interact, so in the subseasonal forecast range they represent not only the slowly evolving boundary conditions but also the part of the internal variability that provides predictability by changing the statistics of the higher-frequency weather. Thus, naturally, the seamless transition from short to extended range forecasts requires aggregations that capture the variability of the large-scale patterns in our meteorological variable of interest.

In practice, subseasonal forecasting aims to extend the time window of the predictand with increasing lead times (Nicolis, 2016; Wheeler *et al.*, 2017; Ford *et al.*, 2018; Bürger, 2020). It has been demonstrated that more aggregation indeed leads to a general predictability in upper air fields at longer lead times (Roads, 1986; Jung and Leutbecher, 2008; Buizza and Leutbecher, 2015) and in surface variables such as precipitation (Wheeler *et al.*, 2017). Studies have also tailored the aggregation to a single conditional source of predictability: rainfall events in Europe that are clustered in time due to large-scale dynamics

(Economou *et al.*, 2015; Pasquier *et al.*, 2019; Yang and Villarini, 2019), or extreme temperatures occurring simultaneously within a spatial region due to large-scale flow or sea-surface temperatures (Stefanon *et al.*, 2012; McKinnon *et al.*, 2016; Vijverberg *et al.*, 2020). The forecast skill of such derived predictands can be high, but it is conditional on the occurrence of the source mechanism, and might also lose validity for less or more extreme events (Wulff and Domeisen, 2019). To improve skill under all physical circumstances, statistical post-processing is often used to correct systematic biases and under- or over-dispersion. This aligns the model error growth with the real uncertainty growth (Wilks, 2018). In this way, studies have been able to demonstrate predictability of weekly aggregations into weeks 3 and 4 for the midlatitudes (Ferrone *et al.*, 2017; Vigaud *et al.*, 2017; Monhart *et al.*, 2018).

In conjunction with increased aggregation leading to increased predictability, based on a physical understanding one would also expect an optimum to exist. When too many different situations are aggregated, the conditional predictability in either one of them is lost, for instance by spatially aggregating hotspots of soil–atmosphere coupling with non-hotspots (Ardilouze *et al.*, 2017) or by temporally aggregating beyond the time window in which flow configuration modulates precipitation significantly (Barton *et al.*, 2016). Predictability is then only regained by aggregating even further, for example to multi-month values in order to capture the modulation of Europe's seasonal state by the El Niño–Southern Oscillation or soil moisture (Bunzel *et al.*, 2018; Lee *et al.*, 2019).

This study benchmarks how the subseasonal predictability of surface temperatures in Europe varies over the continent and changes with the amount of temporal and spatial aggregation applied. We hypothesize that the maximal extension of the forecast skill horizon (Buizza and Leutbecher, 2015) occurs under certain optimum aggregation levels and by statistical post-processing of the raw ensemble forecasts. Section 2 introduces the forecast ensemble, the scores to determine the forecast horizon and the post-processing method. Section 3 shows the resulting influences of post-processing and aggregation for events with varying exceedance thresholds. Section 4 provides a discussion and Section 5 summarizes and concludes.

2 | DATA AND METHODS

2.1 | Datasets

The forecast ensemble is the European Centre for Medium-Range Weather Forecasts' (ECMWF's) extended range forecasting system cycle 45r1, which extends their medium range ensemble twice a week to +46 days (Buizza

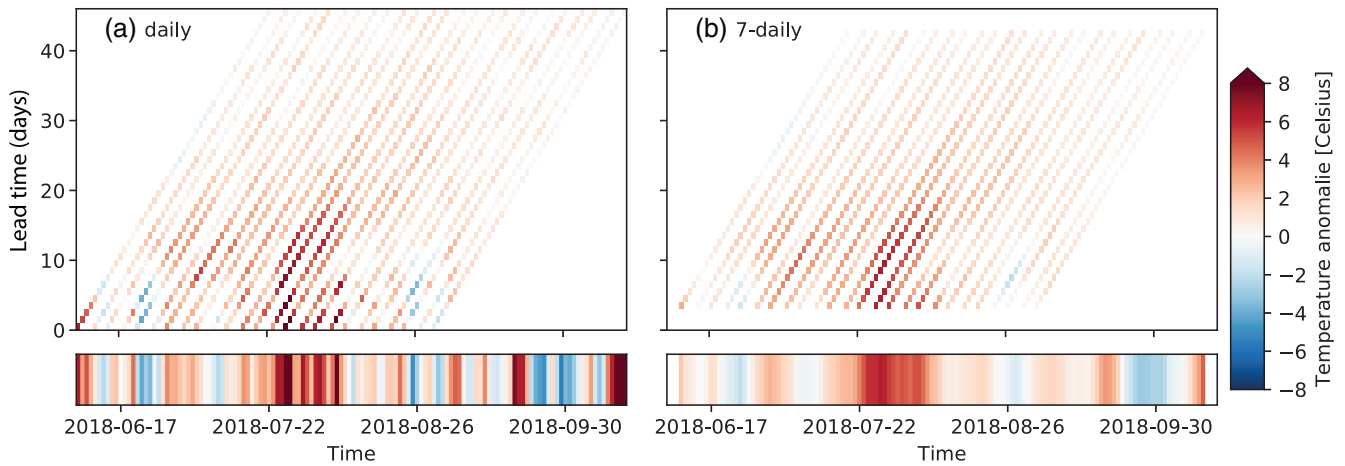


FIGURE 1 Temporal aggregation of extended range forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) and corresponding E-OBS observations, exemplified with daily mean 2-m temperature anomalies during the 2018 European heatwave. Top panels: Ensemble mean of forecasts initialized each Monday and Thursday, aggregated to (a) daily and (b) 7-day rolling means. Note that the window size remains 7 days for all lead times indicated on the vertical axis. Bars at bottom: Corresponding E-OBS observations. Data are taken from the grid cell closest to 52° latitude and 7° longitude

et al., 2018). A degradation of the resolution takes place at +16 days. We downloaded forecasts of daily mean 2-m temperatures on a regular grid of $0.38 \times 0.38^\circ$ as it equates the degraded spectral resolution in large parts of the European domain and minimizes the need for MIR interpolation on the ECMWF MARS archive. For forecasting in the extended range, a lead-time-dependent bias in the model climatology can be expected (Johnson *et al.*, 2019). All 11 members in the re-forecast period from June 1998 to May 2019 were therefore used to calculate forecast climatological means specific to the day of the year (± 5 days) and the lead time, that were subtracted from the forecast values. This results in forecast anomalies with respect to the climatology of the model re-forecast that are free from potential drifts in that climatology.

Observed temperature anomalies were derived from version 19.0 of the E-OBS ensemble dataset (Cornes *et al.*, 2018). Its ensemble mean forms the best guess of observed daily mean 2-m temperatures on a $0.25 \times 0.25^\circ$ grid. From the 60+ years of data in the dataset we retained those 20 years that overlapped with the re-forecasts. At each location we subtracted the observed climatological mean specific to the day of the year (± 5 days) calculated from January 1998 to December 2018.

The daily gridded anomalies from E-OBS were then paired with the 11 forecast anomalies in the nearest-neighbour forecast ensemble grid cell, representing an area that is only slightly different. The datasets span from June 1998 to December 2018 and are built separately for the winter and summer seasons, that is, December–January–February (DJF) and June–July–August (JJA). We allow days of forecasts that were

initialized before the start of the seasonal window to be included (Coelho *et al.*, 2018).

2.2 | Aggregation

The paired daily anomalies at all E-OBS grid cells in Europe were then averaged to multiple spatial and temporal levels and all combinations of those levels. The temporal levels consist of rolling 1- to 11-day window averages. Each of these windows is applied to all lead times equally, and assigns the lead time of a given forecast to the window centre, which is a compromise between the more accurate first days and the more uncertain last days in the window (Weigel *et al.*, 2008; Buizza and Leutbecher, 2015). Thus, for a window of 7 days, the first possible midpoint lead time is 4 days, which is assigned the average of the anomalies from forecast days 1–7 (see Figure 1).

The spatial levels are determined by hierarchical clustering (Hastie *et al.*, 2009). This method begins with as many clusters as there are grid cells and a dissimilarity defined between each of these, say, time series A and B :

$$d_{A,B} = 1 - \max_{\tau=-20, \dots, 20} \rho(A_{t-\tau}, B_t). \quad (1)$$

This maximum in a set of correlations ρ with lags τ ranging from -20 to $+20$ days allows cells to be similar while experiencing the same (but temporally displaced) dominant weather features (Pfleiderer and Coumou, 2018). Each level of spatial averaging is then determined by grouping all sets of grid cells below a certain dissimilarity level (e.g., the level of 0.025 requires a minimum similarity, namely

lagged correlation exceeding 0.975, between each of the cells) into single clusters, until the whole of Europe is one single cluster. We opt for an average linking rule (see Hastie *et al.*, 2009). Our progression through the dissimilarity levels from 0.025 to 1 avoids the common problem of assuming a fixed number of clusters at the beginning of a study (e.g., Yiou *et al.*, 2008). We perform the cluster extraction for winter and summer separately, using the observed daily temperatures from January 1989 to December 2018. The use of daily time series separates our spatial aggregation from the temporal aggregation and allows for a separate investigation into their effects. The supposed independence was briefly tested, and similarly shaped spatial clusters appeared at other time aggregations.

2.3 | Scoring and forecast skill horizon

Each set of anomalies, averaged to a spatial and temporal level, is evaluated by comparing the distribution of forecast anomalies to the observed one. First, we extract the forecast probability that a temperature anomaly will exceed a certain quantile in the 20-year model re-forecast climatology of averaged anomalies. The Brier score (BS) then averages the squared difference between this probabilistic prediction p_i and the binary observation o_i (whether or not the observed anomaly exceeded the equivalent quantile in the observed 20 year climatology of averaged anomalies) over the n forecast–observation pairs per lead time and per spatial cluster in each set:

$$BS = \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2. \quad (2)$$

Using two equivalent thresholds, this BS extends the mean de-biasing, performed to create anomalies, with an implicit calibration of the raw ensemble forecasts to match the observed climatological spread. Additionally, the BS of a reference forecast based on only the observed climatology is computed. It has a fixed p_i , namely 1 minus the quantile probability itself.

The full 11-member distribution is scored with the continuous ranked probability score (CRPS). The implicit calibration mentioned above has no effect on the CRPS as that score can be regarded as the BS integrated over all possible thresholds y , and accounts for reliability and sharpness (e.g., Wilks, 2011):

$$CRPS(F, y) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} (F_{for,i}(y) - F_{obs,i}(y))^2 dy. \quad (3)$$

F_{for} is the forecast cumulative distribution function (cdf) and F_{obs} is the observed single-step cdf. As the forecast

distribution is a discrete ensemble of 11 members, it receives worse CRPS scores than a version of the same underlying distribution with more members. A fair reference score is thus formed by sampling the same number of members ($M = 11$) from the empirical climatological distribution F of the observed anomalies, at intervals determined by a Weibull estimator (Wilks, 2011):

$$F^{-1} \left(\frac{m}{M+1} \right) \quad \text{for } m \text{ in } 1, \dots, M. \quad (4)$$

A persistence reference forecast might be harder to beat, but since the construction of its probability distribution is non-trivial (Smith *et al.*, 2015), we use the climatological reference to transform both scores to a skill score (SS): $BSS = 1 - BS/BS_{clim}$ and $CRPSS = 1 - CRPS/CRPS_{clim}$. For each cluster and lead time we determine a confidence interval around these skill scores by scoring random samples (with substitution) from the set of n forecast–observation pairs. Because of auto-correlation, which will differ between clusters and which will increase with larger rolling-window sizes at higher temporal aggregation levels, this bootstrapping is done with different block lengths for each. The block lengths are based on a measure of the characteristic time-scale T_0 (Figure 2; Feng *et al.*, 2011):

$$T_0 = 1 + 2 * \sum_{\tau=1}^D (1 - \tau/D) * \rho_{\tau}, \quad (5)$$

where D is a cutoff lag, which, similarly to the hierarchical clustering, we set to 20 days. ρ_{τ} is the auto-correlation between the lagged and unlagged time series of a cluster. The block bootstrapping is repeated only 200 times due to computational limitations. With these skill confidence intervals per cluster and per lead time we then deduce the local forecast skill horizons, defined as the lead time at which the lower bound of the interval (the 2.5th percentile) first crosses the zero skill line (Buizza and Leutbecher, 2015); this means the lead time at which the forecast ceases to be statistically better than the climatological reference forecast according to a one-tailed test at a 0.025 significance level.

2.4 | Statistical post-processing

Besides scoring the aggregated, but otherwise unprocessed, forecast anomalies and scoring the climatological reference, we also score a version of the ensemble that is post-processed with a non-homogeneous Gaussian regression (NGR), which is a standard post-processing method for temperatures (Wilks and Vannitsem, 2018). Its Gaussian distribution is assumed to have a location

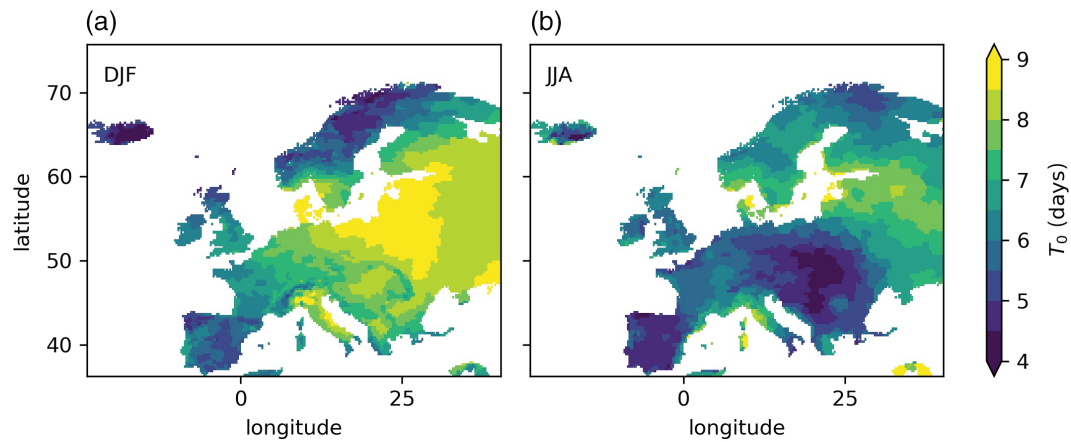


FIGURE 2 Characteristic time-scale in the daily observed temperature anomalies at the 0.025 spatial aggregation level. (a) Winter, 1,158 clusters. (b) Summer, 977 clusters

parameter μ_i and a scale parameter σ_i that vary, respectively, with the ensemble mean m_i and the ensemble standard deviation s_i :

$$\mu_i = \alpha_1 + \alpha_2 \cdot m_i, \quad (6)$$

$$\ln(\sigma_i) = \beta_1 + \beta_2 \cdot \ln(s_i). \quad (7)$$

The model is fitted using a three-fold cross-validation by minimizing the CRPS (Gneiting *et al.*, 2005; Gebetsberger *et al.*, 2018) on two thirds of the 20-year dataset and validating on the other third. The model is fitted separately for each season, aggregation level, cluster and lead time. For scoring the post-processed distribution with CRPS we extracted 11 members with the estimator in Equation 4 (a robustness test with 100 members gave similar results).

Obviously NGR is a simple method that uses simple predictors and assumes normality even when it is inappropriate. Some studies have demonstrated the usefulness of more advanced predictors and post-processing methods in the subseasonal-to-seasonal range (Rodney *et al.*, 2013; Yoo *et al.*, 2018; Hwang *et al.*, 2019; Kämäräinen *et al.*, 2019; Strazzo *et al.*, 2019). Such extensions are often specific to single sources of predictability or to a fixed time aggregation. In this study we compare the general predictability at varying aggregations, and aim to do this in a way that is simple but corrects for systematic errors.

3 | RESULTS

3.1 | The effect of post-processing

In Figure 3 the lower bound of bootstrapped BSS is plotted for the exceedance of four climatological quantiles: two for cold anomalies (0.15, 0.33) and two for warm

anomalies (0.66, 0.85). The lowest skill is seen for the stronger-coloured lines, which are the more extreme quantiles that are harder to predict than the more moderate terciles. At short lead times and the daily aggregation level (left panels in Figure 3), post-processing adds skill to the raw ensemble forecasts, even as the implicit calibration made the raw forecasts “climatologically reliable” (Van Schaeybroeck and Vannitsem, 2015). What happens is that in these first 5 days the spread of the under-dispersed raw forecasts is increased by NGR, adding ensemble reliability to the climatological reliability, leading to increased overall reliability (confirmed by a CRPS decomposition, not shown; Hersbach, 2000). After 5 days the added value becomes smaller as the raw has better dispersion properties. At the 9-day aggregation level in winter, between lead times of 5–13 days, the BSS values of the NGR and raw forecasts are even comparable (Figure 3b). Afterwards, NGR forces the ensemble spread to be similar to the observed climatological spread when uncertainty is greatest at large lead times. In this unskilful range, the zero BSS line is contained between the 2.5th and 97.5th percentiles (upper bound not shown). The upper bounds of the bootstrapped BSS distribution of the raw ensemble are close to those of NGR (not shown) while its lower bounds are lower due to its negatively skewed BSS distribution. The forecast horizon is defined by these lower bounds, meaning that NGR extends the lead time at which the skill crosses zero by about 3 days. In the following we therefore only present results from post-processed forecasts.

3.2 | The effect of aggregation

In Figure 4 we see how aggregation affects predictability in winter, as measured by the forecast skill horizon in the CRPS. For each row, increasing time aggregation tends

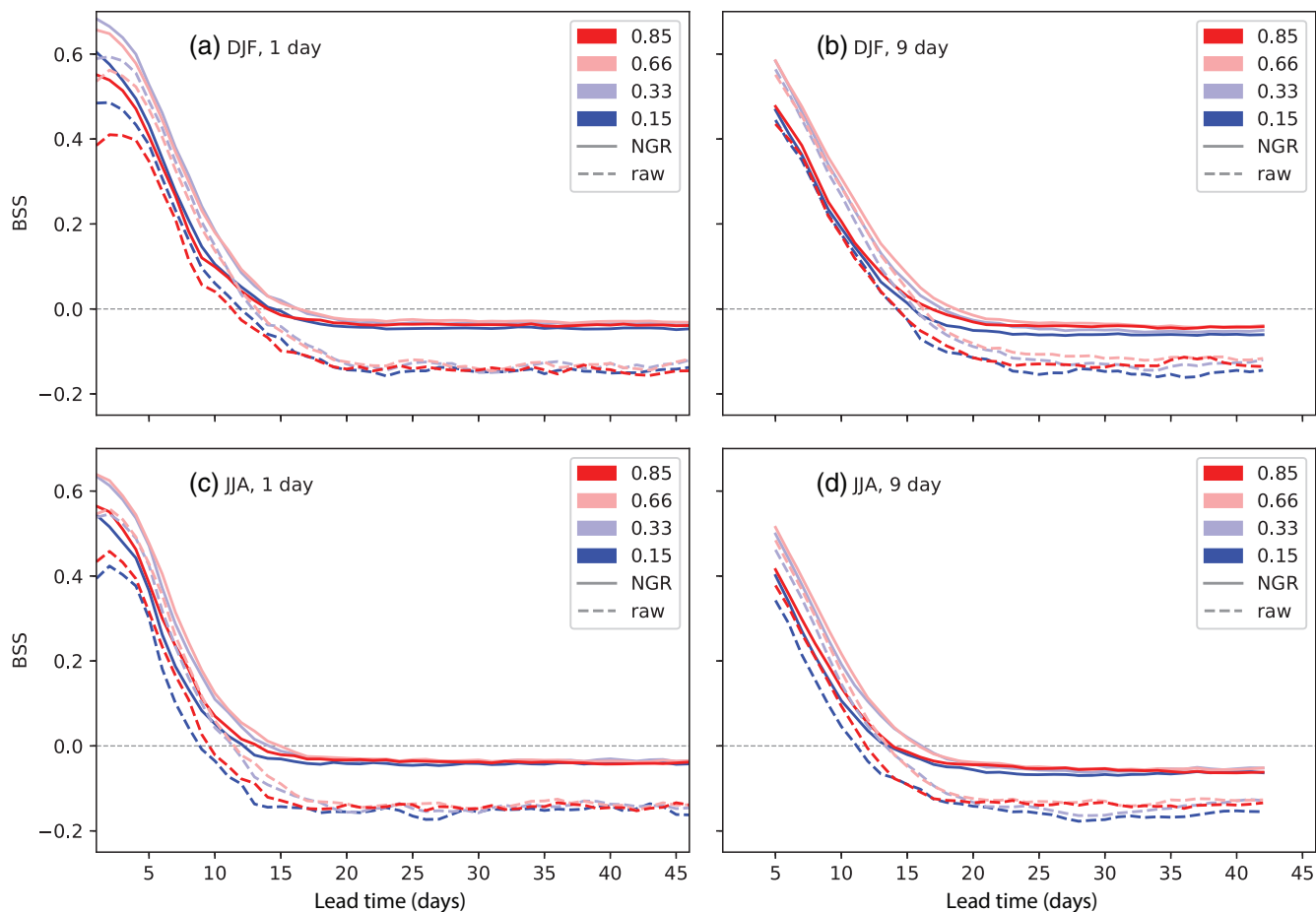


FIGURE 3 Area-weighted average of the 2.5th percentile of bootstrapped BSS over all clusters at the 0.025 spatial aggregation level, plotted for four different climatological quantiles: in winter (top panels) and summer (bottom panels), for daily (left panels) and 9 day (right panels) aggregation levels. Red lines indicate the warm mean temperature anomalies, while blue lines indicate the cool anomalies. Post-processed data are shown with a solid grey line and the raw forecasts are shown with a dashed grey line

to increase predictability. The longest forecast skill horizon is obtained for the 11-day rolling average, except for Iceland. In Iceland the temperature range within a season is quite narrow and is shifted by multiannual variability. The climatological distribution obtained by pooling the years 1998–2018 is thus wider than the range of possibilities at each point in time, leading in comparison to overly skilful post-processed anomaly forecasts (see also the discussion of Figure 8). Other regions that have, for instance, 19-day predictability when a 9-day aggregation is applied to all lead times (lightest, indicating a day 15–23 mean) and 20-day predictability when an 11-day aggregation is applied to all lead times (lightest, indicating a day 15–25 mean) imply that the forecast skill horizon can be extended by using the 11-day window. The longest forecast horizons are obtained for hardly any spatial aggregation (top row, 1,158 clusters) or full aggregation to the European scale (bottom row, 1 cluster). At an intermediate level of spatial aggregation (from 483 to 20 clusters) some local regions have reducing and later increasing forecast

skill horizons, indicating that space aggregation can work both as a benefit and as a disadvantage.

Similar results for summer are shown in Figure 5. Skilful forecasts do not extend as far as they do for winter, indicating the lower general predictability of summer temperature anomalies. Time aggregation has the largest influence at the lowest spatial aggregation. At larger spatial aggregations, the forecast skill horizon of regions sometimes hardly changes, meaning that the averaging works equivalently to a smoother.

Whether aggregation changes the forecast skill horizon merely due to smoothing of the skill of underlying regions/days or due to the extraction of a signal with a truly different predictability is illustrated in Figures 6 and 7. In Figure 6 the lower bound of the bootstrapped CRPSS in each cluster is plotted against the lead time. At the 0.025 aggregation level the clusters form a spatial distribution, which at the European level is only one value per lead time bin. Both in winter and in summer at the daily time aggregation (Figure 6a,c) the European

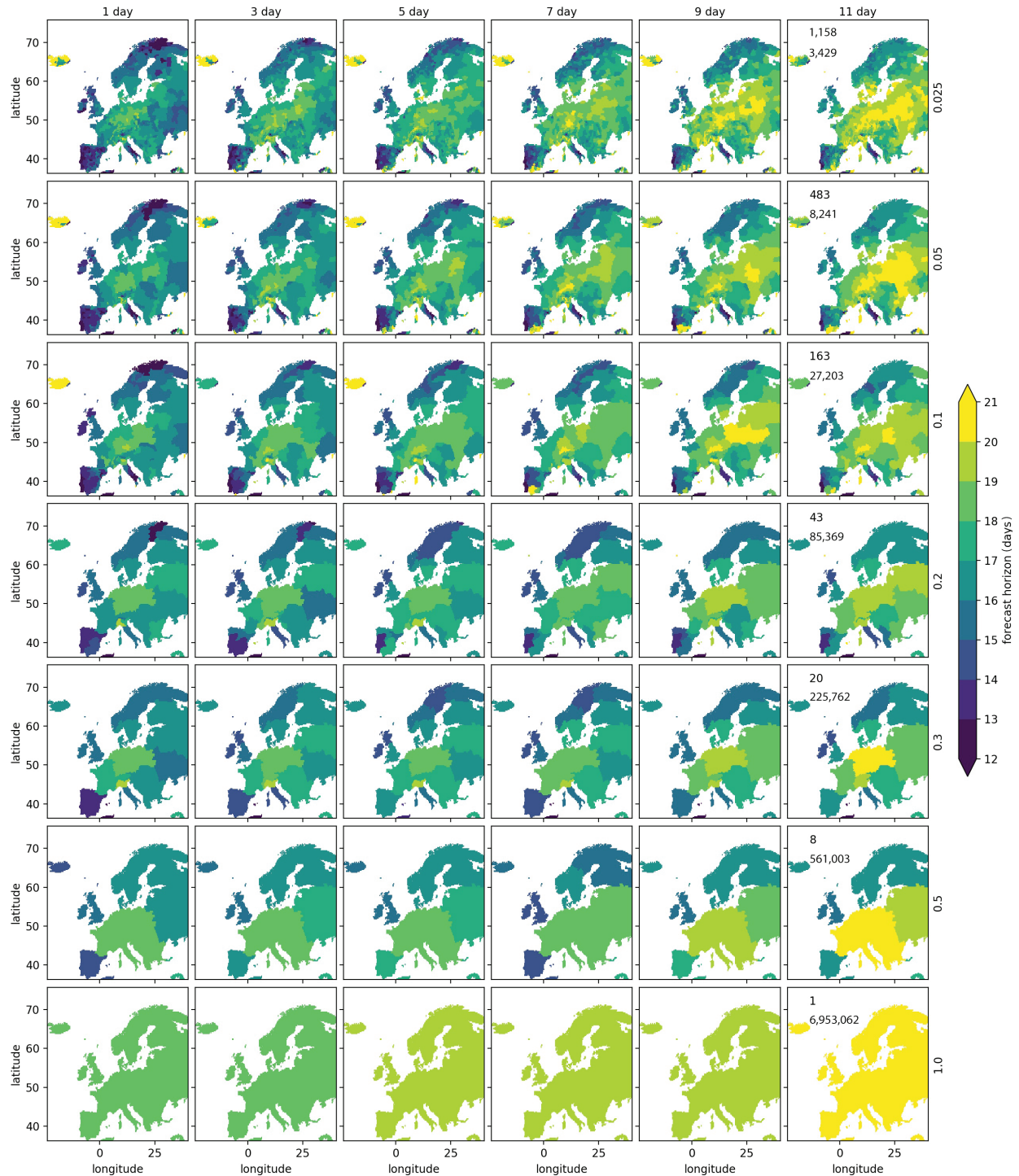


FIGURE 4 Forecast skill horizon (days) for post-processed winter temperature anomalies for different levels of spatial aggregation from small-scale to Europe-wide averages (rows), and for temporal aggregation from daily to 11-day rolling averages (columns). In the right column annotations indicate the dissimilarity threshold, the number of clusters and their median size (km^2)

CRPSS is clearly higher than the average of the underlying clusters. Particularly at lead times shorter than 11 days, the European aggregate has more predictability than the ensemble of regions. Near the forecast horizon (where most clusters cross the zero line) it tends to the interquartile range, which implies that spatial aggregation acts as a

smoother. Some individual regions have more skill and a more extended forecast horizon when the degree of spatial aggregation is limited. The dots above the zero line at very long lead times are locations with variable scores, not with interminable forecast horizons; their lower bound will equally have jumped below zero at earlier lead times.

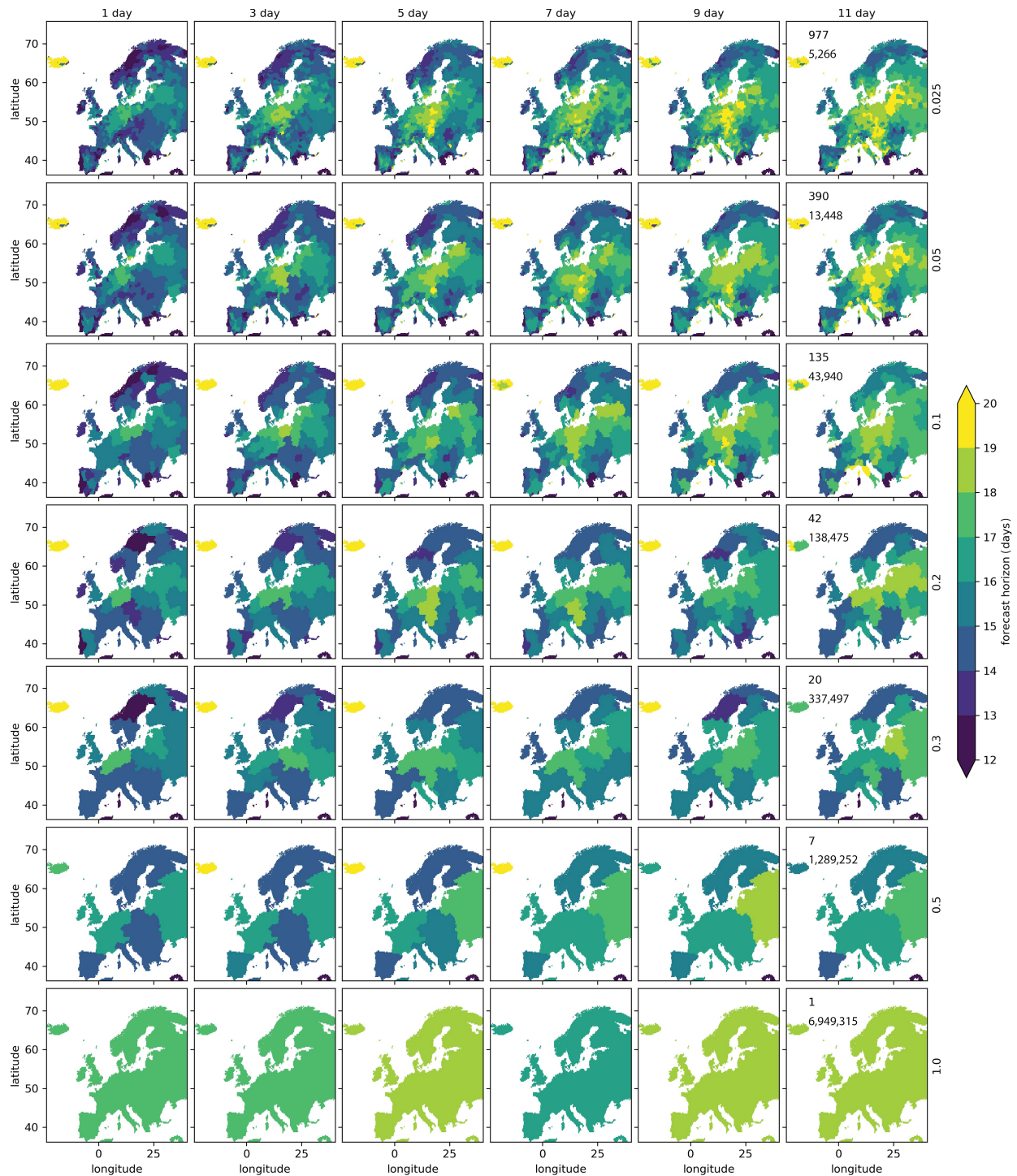


FIGURE 5 As Figure 4 but for summer. Note that a different spatial clustering and a different colour scale has been applied than for the winter season

At longer time aggregations (Figure 6b,d), the mentioned effects of space aggregation are less pronounced but still present.

In Figure 7 the spatial CRPS distributions belonging to two time aggregations are compared (outliers are not shown for clarity). The interquartile ranges of the daily and the 9-day scores only become distinguishable at lead times

exceeding 6 days, indicating that beyond this lead time a predictable multi-day variability was well initialized and was captured by the simple 9-day average. This extends the median forecast horizon by about 2 days. Some of the differentiation between the time aggregation levels can also be related to the convex shape of the curve between lead times of 9 and 15 days. There the temporal window

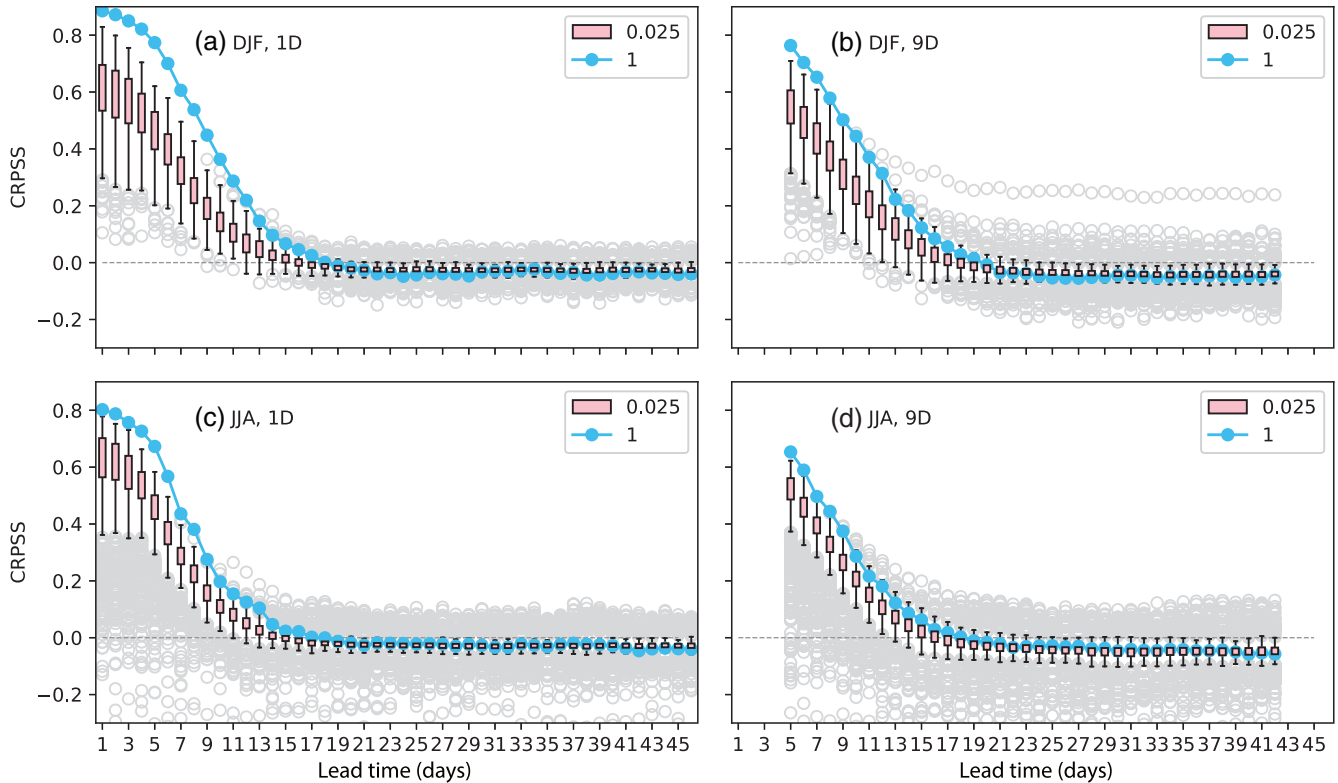


FIGURE 6 The influence of spatial aggregation on the 2.5th percentile of bootstrapped CRPSS at 1-day (left) and 9-day (right) time aggregations. The spatial distribution at the 0.025 aggregation level (1,158 clusters in winter [top] and 977 clusters in summer [bottom]) is shown with a box for the inter-quartile range, with whiskers extending to 1.5 times that range and with open dots for the outliers. The European averaged aggregation (1 cluster) is shown with solid connected dots

is a favourable mixture of initial days that are much more predictable than its centre day (to which its lead time is assigned) and final days that are only slightly less predictable. However, higher CRPSS values also appear along the straight section between lead times of 5 and 9 days, and Buizza and Leutbecher (2015) demonstrated that the skill of time-averaged variables is higher than the skill of time-averaged scores. Therefore, we are confident that the increased forecast horizon can be attributed to the temporal aggregation applied.

The extensions and regional optimums that we find have to be related to sources of predictability. We can expect such sources to be related to particular types of events, and for their conditional predictability to emerge at a certain level of intensity. Becker *et al.* (2013) found for instance increased signal-to-noise ratios for extreme events, despite an equally increasing error in predicting them. In Figure 8 we show the BSS forecast skill horizon for predicting the exceedance of varying climatological quantiles. Note that the forecast skill horizon for Iceland is now strongly reduced compared to Figures 4 and 5. This difference is not surprising because the CRPS is an integration of the BS over all possible thresholds per point in time, while the BSS in Figure 8 is created by first summing

over time. The inflated CRPSS for Iceland followed from a reference that was too wide for the varying set of possibilities at each point in time. In the case of the BSS, the varying exceedance probability is over-estimated in some years and under-estimated in others, but aggregated over all forecast occasions the reference is by definition exactly right and the skill is not inflated. The source of these multiannual changes can be sought in the sea-surface temperatures (Frajka-Williams *et al.*, 2017). These are able to dominate because E-OBS includes only coastal stations in Iceland (Cornes *et al.*, 2018).

Particularly for the largest time aggregation (Figure 8, right column), the forecast horizon for the different quantiles shows an asymmetry. The upper tercile is more predictable than the lower tercile and this predictability is primarily located in the east of the domain. In the raw forecasts this spatial structure is also present but less pronounced (Figure 9), indicating that the asymmetry is partially caused by NGR. Closer investigation reveals that the climatological distribution in regions with long forecast skill horizons is negatively skewed. Initially NGR corrects the under-dispersion of the raw forecast and performs well, but when uncertainty increases with long lead times and dispersion approaches climatology the thicker lower

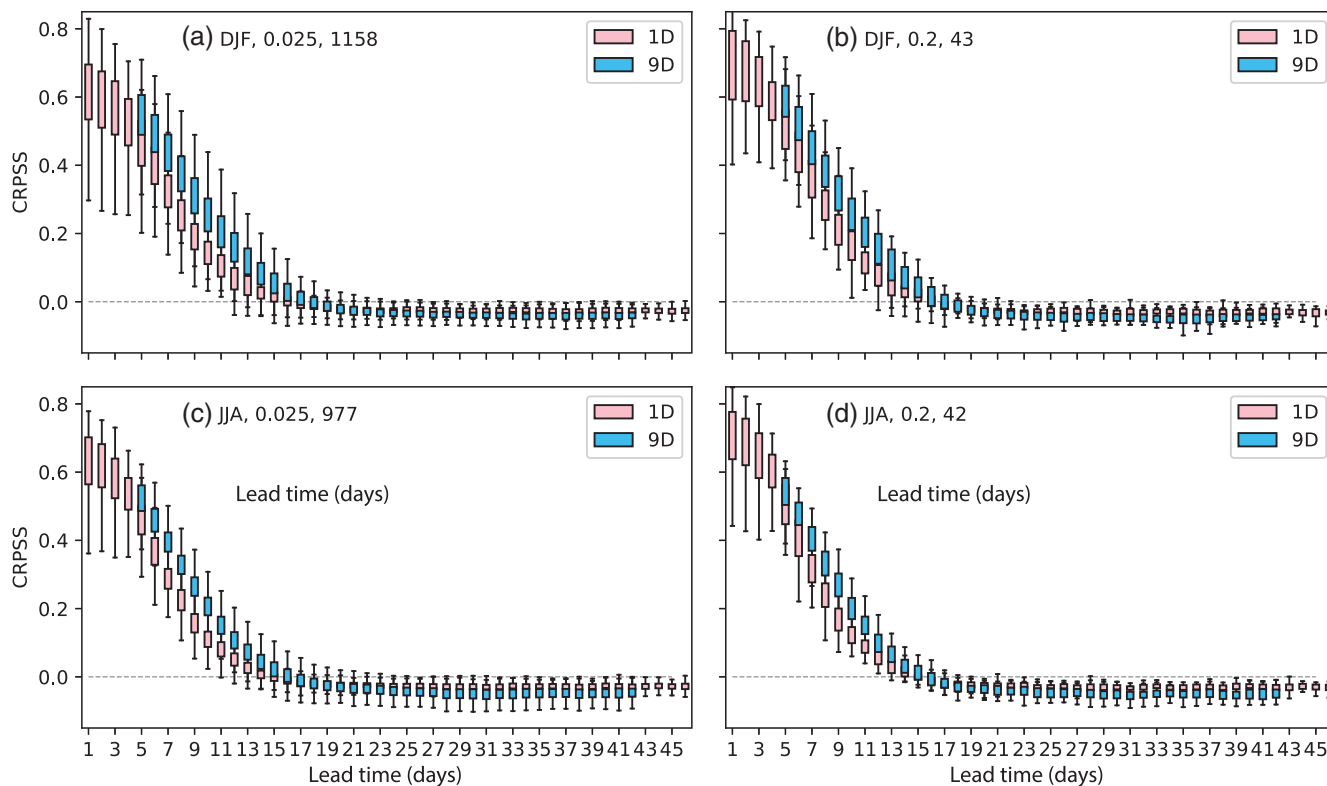


FIGURE 7 The influence of time aggregation on the 2.5th percentile of bootstrapped CRPSS. The spatial distribution of daily time series is shown in pink and that of the 9-day rolling averages in light blue box-and-whisker plots. Boxes represent the interquartile range, whiskers extend to 1.5 times that range and outliers are not shown. Annotations indicate the season, spatial aggregation level and number of clusters

tail is badly represented by the post-processed Gaussian shape and we see the performance for cold quantiles drop relative to the warm quantiles.

For summer (Figure 10), time aggregation does not clearly reveal an asymmetry in the forecast skill horizons. At quantiles 0.1 and 0.9 some regions show consistently short forecast horizons, despite time aggregation. Generally the moderate events in the bulk can be better predicted than events in the tails. This ordering is in contrast with the study of Wulff and Domeisen (2019), who found that European warm extremes in summer, exceeding the 90th percentile and at a 5 day temporal aggregation, are more predictable than moderate events between the 25th and 75th percentiles. They found this for the warm tail only, so they hypothesized that the emergent source of conditional predictability related to land–atmosphere feedbacks and large-scale circulation. Here we find no indication of an emergent source.

4 | DISCUSSION

The forecast skill horizons presented above are in agreement with other estimates of European unconditional predictability in bias-corrected forecasts (Ferrone *et al.*, 2017;

Monhart *et al.*, 2018). We find the forecast skill horizon for the full forecast distribution to be highest in winter, where midpoint lead times extend to slightly above 21 days, meaning that the windows of predictability can be extended up to weeks 3 and 4. In this study we have varied the level of aggregation to test its impact on the predictability horizon. Our findings show that no distinct aggregation captures the one and only predictable subseasonal signal in Europe. Results suggest that the predictable mode of variability varies over the domain and that aggregation can increase predictability (but does not always do so).

For areas where subseasonal predictability exists, time aggregation increases skill, predominantly beyond a given lead time (Figure 7). This confirms that it is optimal to apply aggregation only when uncertainty has increased with lead time and when the predictable low-frequency signal remains (Ford *et al.*, 2018; Bürger, 2020). In other areas, however, especially for more extreme quantiles, time aggregation had almost no effect on forecast skill horizon. It just smoothed the skill (or the absence thereof) over time and no predictable signal captured by simple averaging appeared.

In contrast, space aggregation changed the signal considerably at short lead times (Figure 6). For the first 11 days, the European average was easier to predict than

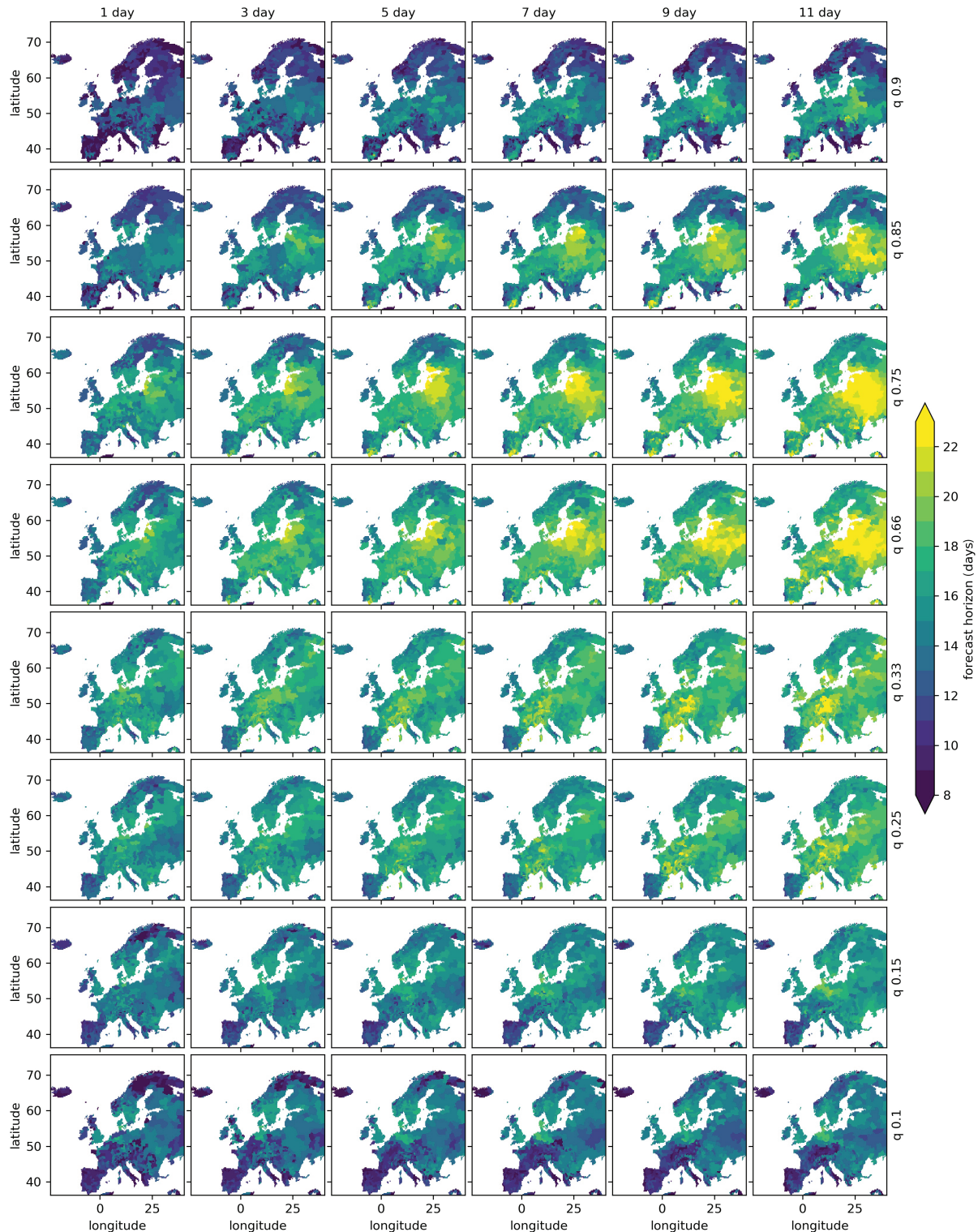


FIGURE 8 Forecast skill horizon (days) of post-processed forecasts in winter for different climatological quantiles (rows) for varying levels of time aggregation (columns) and at the 0.025 spatial aggregation level (1,158 clusters). Bottom: Cold tail. Top: Warm tail

the ensemble of regions. At this largest spatial aggregation, winter results (Figure 4) showed that it is best to also aggregate in time. This confirms that the dominant features are best captured by changing both the space and time filters as both scales are related (the North Atlantic Oscillation

varies slowly and influences temperatures over the entire European continent) (World Meteorological Organization, 2015). However, this study also found conflicting evidence, namely that for certain regions the forecast horizon is not maximized by increasing spatial aggregation. We think

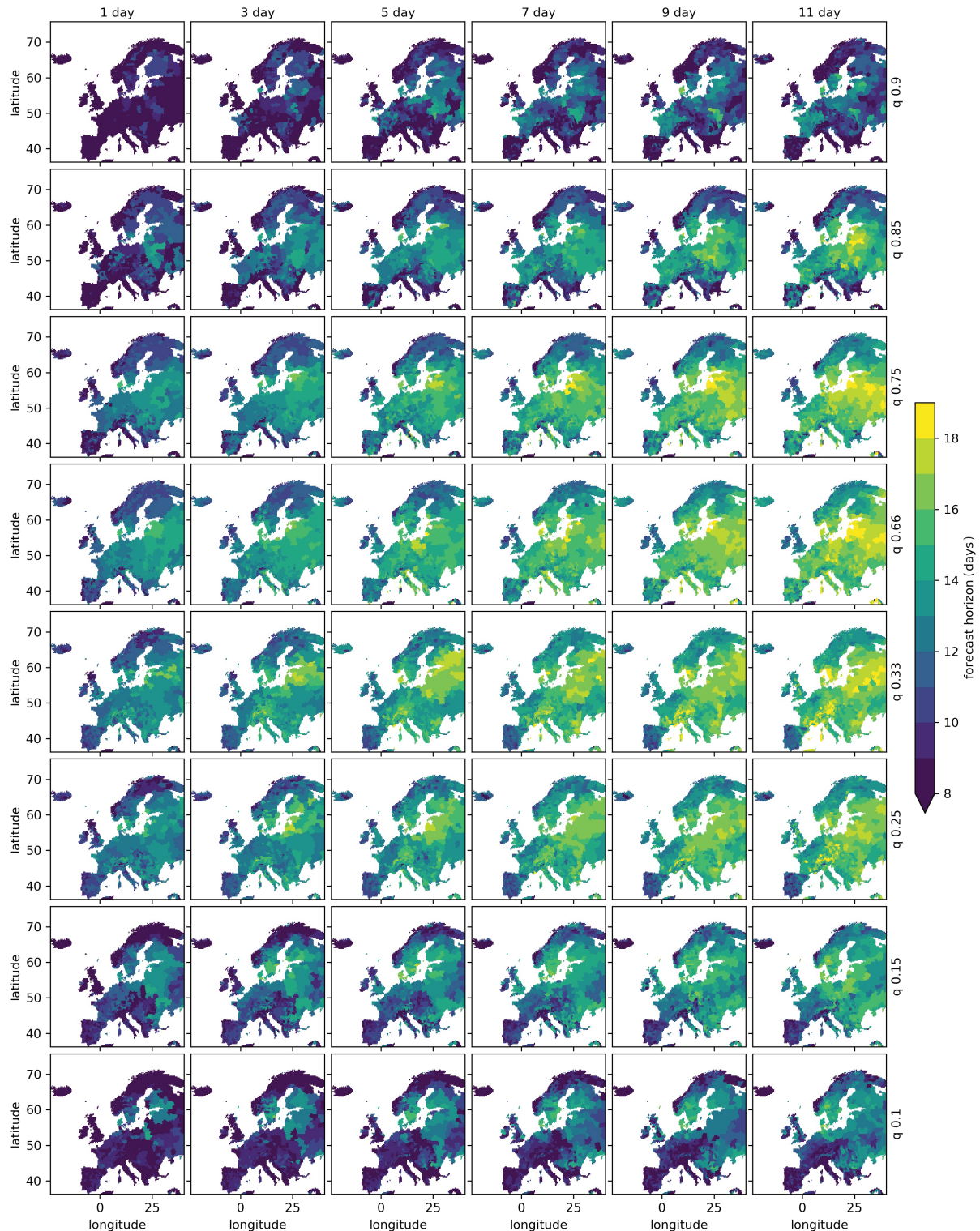


FIGURE 9 As Figure 8 but for raw DJF forecasts. Note the different colour scale

that predictability in either of the underlying regions is then lost by mixing the physical mechanisms that modulate locally.

We hypothesized that specific sources of predictability could be identified from anomalies exceeding specific climatological quantiles. One school of thought is that

extreme events are related to predictable large-scale drivers and can therefore be better predicted themselves (Sillmann *et al.*, 2017). The other is that extreme events are actually harder to predict because they require a rare synchronization of processes at all relevant scales. We did not investigate extremes beyond the 10th and 90th percentiles,

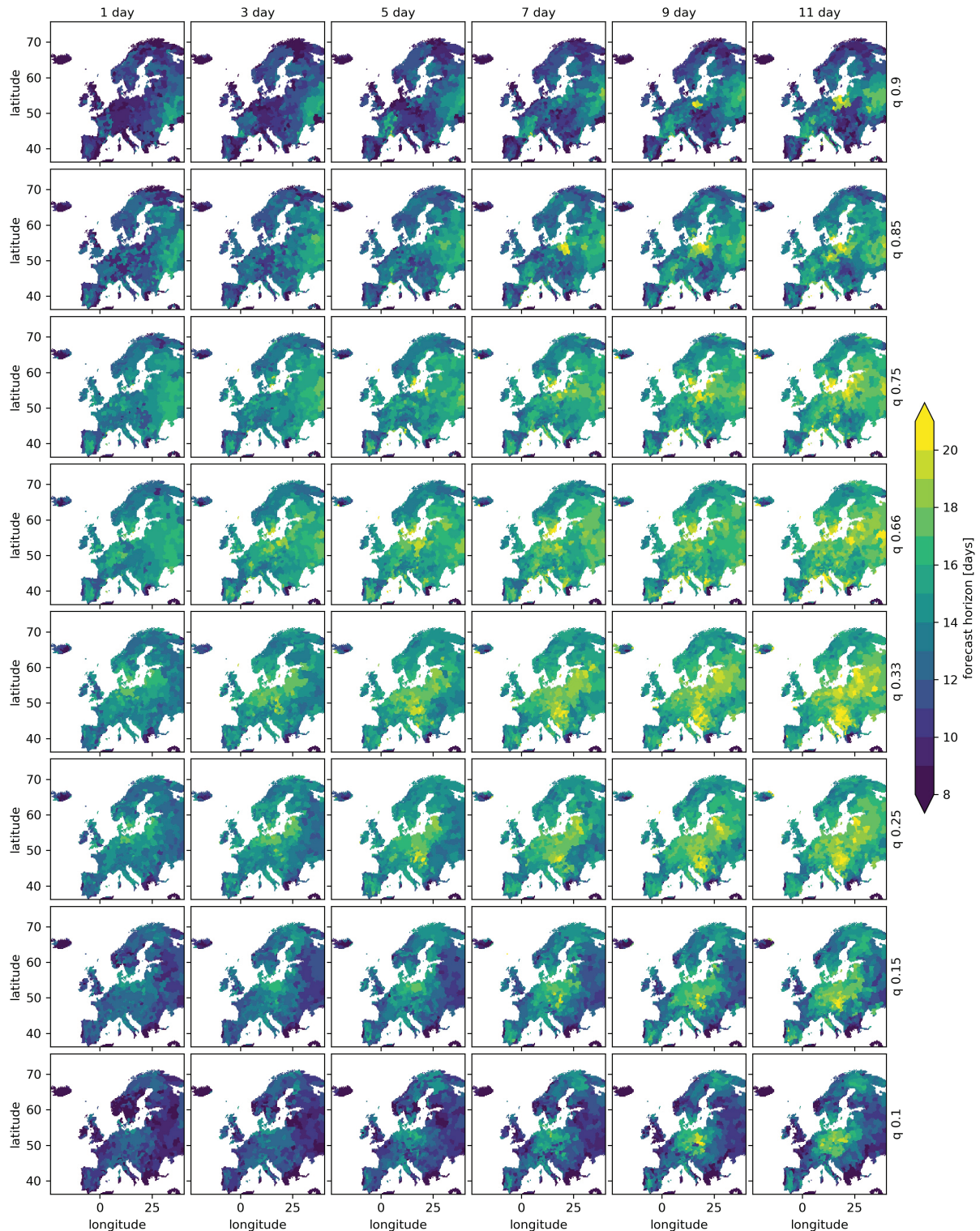


FIGURE 10 As Figure 8 but for summer. Note the different colour scale

but our results support both schools of thought. The BSS curves (Figure 3) showed that tail events are harder to predict, and we also found no indication of increased predictability of summer warm extremes (Figure 10; Wulff and Domeisen, 2019). On the other hand, the winter BSS (Figure 8) displayed a regional signal in the upper

quantiles (visible only at larger time aggregations) which we might relate to an emergent predictable phenomenon.

A candidate mechanism associated with above-normal temperatures in winter for a ± 10 -day time aggregation could be the early disappearance of the snow pack, as this increases the absorption of short-wave radiation and takes

some time to rebuild itself again. Certainly, the regionally extended forecast skill horizons are at least also partly due to the persistence of weather. The region around the Baltic Sea and Denmark is persistent in winter (Figure 2) and strongly imprinted by the first principal component of the large-scale Euro-Atlantic atmospheric variability (Ferranti *et al.*, 2018). This results in a relatively skilful region in our analysis (as in Monhart *et al.* (2018)).

Clusters that displayed weak predictability can be interpreted as being devoid of sources of real extended predictability, but might also just indicate biases in the model. With NGR we attempted to remove biases and under- and over-dispersion for each lead time, season and cluster. This led to noticeable increases of skill, but also to a bias for winter cold anomalies at long lead times in regions with a skewed climatological distribution (Figure 8). A correction approach that can better handle such distributions and, for example, the multiannual variability change in Iceland, might be realized with other simple (Ferrone *et al.*, 2017; Vigaud *et al.*, 2017) or more advanced (Yoo *et al.*, 2018; Hwang *et al.*, 2019; Kämäräinen *et al.*, 2019; Strazzo *et al.*, 2019) post-processing methods.

5 | CONCLUSION

This study has demonstrated that the forecast skill horizon for average temperatures varies over the European domain and can be extended to weeks 3 and 4 without preconditioning. A standard non-homogeneous Gaussian regression post-processing step added three skilful forecast days on average. The influence of space and time aggregation was explored by a protocol that allowed a clean comparison of different aggregation levels. We found that simple averaging captures predictable large-scale patterns in high-frequency weather and that this aggregation becomes especially effective beyond lead times of a few days, adding two skilful days on average. For some regions, however, time aggregation simply smoothed skill over time, showing that it is not everywhere that a signal is extracted by aggregation. Also, space aggregation, when applied at an intermediate level, was found to lead to smoothing, therefore discarding the local extended forecast horizons present in some regions. To optimize subseasonal predictability in Europe, aggregation should thus be limited in certain cases, especially when it is important to trace back the signals to the associated sources of predictability. This tracing is further eased when, in addition to particular spatio-temporal scales, the types and intensity levels of the events are also known. We have demonstrated that quantiles can be used for such a stratification, but that a source of extended predictability does not always emerge for the more extreme cases. A recommended extension of

this study is to explore other statistics than the average (e.g., DelSole and Tippett, 2009). The predictable modes of variability might be better detected with meteorological index variables, such as the clustering of warm days or rainfall events, than with temperature or rainfall averages.

ACKNOWLEDGMENTS

The software used to download and analyse the data can be accessed at <https://github.com/chiemvs/SubSeas>. This study is part of the open research programme Aard- en Levenswetenschappen, project number ALWOP.395, which is financed by the Dutch Research Council (NWO).

ORCID

Chiem van Straaten  <https://orcid.org/0000-0002-7291-6777>

REFERENCES

- Ardilouze, C., Batté, L., Bunzel, F., Decremer, D., Déqué, M., Doblas-Reyes, F.J., Douville, H., Fereday, D., Guemas, V., MacLachlan, C., Müller, W. and Prodhomme, C. (2017) Multi-model assessment of the impact of soil moisture initialization on mid-latitude summer predictability. *Climate Dynamics*, 49(11–12), 3959–3974. <https://doi.org/10.1007/s00382-017-3555-7>
- Baldwin, M.P. and Dunkerton, T.J. (2001) Stratospheric harbingers of anomalous weather regimes. *Science*, 294(5542), 581–584. <https://doi.org/10.1126/science.1063315>
- Barton, Y., Giannakaki, P., Von Waldow, H., Chevalier, C., Pfahl, S. and Martius, O. (2016) Clustering of regional-scale extreme precipitation events in southern Switzerland. *Monthly Weather Review*, 144(1), 347–369. <https://doi.org/10.1175/MWR-D-15-0205.1>
- Becker, E.J., Van Den Dool, H. and Peña, M. (2013) Short-term climate extremes: prediction skill and predictability. *Journal of Climate*, 26(2), 512–531. <https://doi.org/10.1175/JCLI-D-12-00177.1>
- Buizza, R., Balsamo, G. and Haide, T. (2018) IFS upgrade brings more seamless coupled forecasts. *ECMWF Newsletter*, 156, 18–22. <https://doi.org/10.21957/729r3bdsx6>
- Buizza, R. and Leutbecher, M. (2015) The forecast skill horizon. *Quarterly Journal of the Royal Meteorological Society*, 141(693), 3366–3382. <https://doi.org/10.1002/qj.2619>
- Bunzel, F., Müller, W.A., Dobrynin, M., Fröhlich, K., Hagemann, S., Pohlmann, H., Stacke, T. and Baehr, J. (2018) Improved seasonal prediction of European summer temperatures with new five-layer soil-hydrology scheme. *Geophysical Research Letters*, 45(1), 346–353. <https://doi.org/10.1002/2017GL076204>
- Bürger, G. (2020) A seamless filter for daily to seasonal forecasts, with applications to Iran and Brazil. *Quarterly Journal of the Royal Meteorological Society*, 146(726), 240–253. <https://doi.org/10.1002/qj.3670>
- Cassou, C. (2008) Intraseasonal interaction between the Madden-Julian Oscillation and the North Atlantic Oscillation. *Nature*, 455(7212), 523. <https://doi.org/10.1038/nature07286>
- Coelho, C.A.S., Firpo, M.A.F. and de Andrade, F.M. (2018) A verification framework for South American sub-seasonal precipitation

- predictions. *Meteorologische Zeitschrift*, 27(6), 503–520. <https://doi.org/10.1127/metz/2018/0898>
- Cornes, R.C., van der Schrier, G., van den Besselaar, E.J.M. and Jones, P.D. (2018) An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*. <https://doi.org/10.1029/2017JD028200>
- Coughlan de Perez, E., van den Hurk, B.J.J.M., Van Aalst, M.K., Jongman, B., Klose, T. and Suarez, P. (2015) Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts. *Natural Hazards and Earth System Sciences*, 15(4), 895–904. <https://doi.org/10.5194/nhess-15-895-2015>
- Czaja, A. and Frankignoul, C. (2002) Observed impact of Atlantic SST anomalies on the North Atlantic Oscillation. *Journal of Climate*, 15(6), 606–623. [https://doi.org/10.1175/1520-0442\(2002\)015<0606:OIOASA>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<0606:OIOASA>2.0.CO;2)
- DelSole, T. and Tippett, M.K. (2009) Average predictability time. Part II: seamless diagnoses of predictability on multiple time scales. *Journal of the Atmospheric Sciences*, 66(5), 1188–1204. <https://doi.org/10.1175/2008JAS2869.1>
- Economou, T., Stephenson, D.B., Pinto, J.G., Shaffrey, L.C. and Zappa, G. (2015) Serial clustering of extratropical cyclones in a multi-model ensemble of historical and future simulations. *Quarterly Journal of the Royal Meteorological Society*, 141(693), 3076–3087. <https://doi.org/10.1002/qj.2591>
- Feng, X., DelSole, T. and Houser, P. (2011) Bootstrap estimated seasonal potential predictability of global temperature and precipitation. *Geophysical Research Letters*, 38(7), L07702. <https://doi.org/10.1029/2010GL046511>
- Ferranti, L., Magnusson, L., Vitart, F. and Richardson, D.S. (2018) How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe?. *Quarterly Journal of the Royal Meteorological Society*, 144, 1788–1802. <https://doi.org/10.1002/qj.3341>
- Ferrone, A., Mastrangelo, D. and Malguzzi, P. (2017) Multimodel probabilistic prediction of 2 m-temperature anomalies on the monthly timescale. *Advances in Science and Research*, 14, 123–129. <https://doi.org/10.5194/asr-14-123-2017>
- Ford, T.W., Dirmeyer, P.A. and Benson, D.O. (2018) Evaluation of heat wave forecasts seamlessly across subseasonal timescales. *npj Climate and Atmospheric Science*, 1(20), 2397–3722. <https://doi.org/10.1038/s41612-018-0027-7>
- Frajka-Williams, E., Beaulieu, C. and Ducheze, A. (2017) Emerging negative Atlantic Multidecadal Oscillation index in spite of warm subtropics. *Scientific Reports*, 7(1), 11224. <https://doi.org/10.1038/s41598-017-11046-x>
- Gebetsberger, M., Messner, J.W., Mayr, G.J. and Zeileis, A. (2018) Estimation methods for nonhomogeneous regression models: minimum continuous ranked probability score versus maximum likelihood. *Monthly Weather Review*, 146(12), 4323–4338. <https://doi.org/10.1175/MWR-D-17-0364.1>
- Gneiting, T., Raftery, A.E., Westveld, A.H. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5), 1098–1118. <https://doi.org/10.1175/MWR2904.1>
- Grams, C.M., Beerli, R., Pfenninger, S., Staffell, I. and Wernli, H. (2017) Balancing Europe's wind-power output through spatial deployment informed by weather regimes. *Nature Climate Change*, 7(8), 557. <https://doi.org/10.1038/nclimate3338>
- Grotjahn, R., Black, R., Leung, R., Wehner, M.F., Barlow, M., Bosilovich, M., Gershunov, A., Gutowski, W.J., Gyakum, J.R., Katz, R.W., Lee, Y.-Y., Lim, Y.-K. and Prabhat (2016) North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. *Climate Dynamics*, 46(3–4), 1151–1184. <https://doi.org/10.1007/s00382-015-2638-6>
- Guimares Nobre, G., Hunink, J.E., Baruth, B., Aerts, J.C.J.H. and Ward, P. (2019) Translating large-scale climate variability into crop production forecast in Europe. *Scientific Reports*, 9(1), 2045–2322. <https://doi.org/10.1038/s41598-018-38091-4>
- Hannachi, A., Straus, D.M., Franzke, C.L.E., Corti, S. and Woollings, T. (2017) Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55(1), 199–234. <https://doi.org/10.1002/2015RG000509>
- Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer. Springer Series in Statistics.
- Henderson, G.R., Peings, Y., Furtado, J.C. and Kushner, P.J. (2018) Snow–atmosphere coupling in the Northern Hemisphere. *Nature Climate Change*, 8(11), 954–963. <https://doi.org/10.1038/s41558-018-0295-6>
- Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570. [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2)
- Hoskins, B. (2013) The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, 139(672), 573–584. <https://doi.org/10.1002/qj.1991>
- Hwang, J., Orenstein, P., Cohen, J., Pfeiffer, K. and Mackey, L. (2019). Improving subseasonal forecasting in the western U.S. with machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, Anchorage, AK; pp. 2325–2335. New York, NY: Association for Computing Machinery.
- Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., Tietsche, S., Decremmer, D., Weisheimer, A., Balsamo, G., Keeley, S.P.E., Mogensen, K., Zuo, H. and Monge-Sanz, B.M. (2019) SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12(3), 1087–1117. <https://doi.org/10.5194/gmd-12-1087-2019>
- Jung, T. and Leutbecher, M. (2008) Scale-dependent verification of ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 134(633), 973–984. <https://doi.org/10.1002/qj.255>
- Kämäräinen, M., Uotila, P., Karpechko, A.Y., Hyvärinen, O., Lehtonen, I. and Räisänen, J. (2019) Statistical learning methods as a basis for skillful seasonal temperature forecasts in Europe. *Journal of Climate*, 32(17), 5363–5379. <https://doi.org/10.1175/JCLI-D-18-0765.1>
- Lee, R.W., Woolnough, S.J., Charlton-Perez, A.J. and Vitart, F. (2019) ENSO modulation of MJO teleconnections to the North Atlantic and Europe. *Geophysical Research Letters*. <https://doi.org/10.1029/2019GL084683>
- Lin, H. and Brunet, G. (2018) Extratropical response to the MJO: nonlinearity and sensitivity to the initial state. *Journal of the Atmospheric Sciences*, 75(1), 219–234. <https://doi.org/10.1175/JAS-D-17-0189.1>

- Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3), 289–307. <https://doi.org/10.3402/tellusa.v21i3.10086>
- McKinnon, K.A., Rhines, A., Tingley, M.P. and Huybers, P. (2016) Long-lead predictions of eastern United States hot days from Pacific sea surface temperatures. *Nature Geoscience*, 9(5), 389. <https://doi.org/10.1038/ngeo2687>
- Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C. and Liniger, M.A. (2018) Skill of sub-seasonal forecasts in Europe: effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres*, 123(15), 7999–8016. <https://doi.org/10.1029/2017JD027923>
- Nicolis, C. (2016) Error dynamics in extended-range forecasts. *Quarterly Journal of the Royal Meteorological Society*, 142(696), 1222–1231. <https://doi.org/10.1002/qj.2724>
- Orsolini, Y.J., Senan, R., Balsamo, G., Doblas-Reyes, F.J., Vitart, F., Weisheimer, A., Carrasco, A. and Benestad, R.E. (2013) Impact of snow initialization on sub-seasonal forecasts. *Climate Dynamics*, 41(7–8), 1969–1982. <https://doi.org/10.1007/s00382-013-1782-0>
- Pasquier, J.T., Pfahl, S. and Grams, C.M. (2019) Modulation of atmospheric river occurrence and associated precipitation extremes in the North Atlantic region by European weather regimes. *Geophysical Research Letters*, 46(2), 1014–1023. <https://doi.org/10.1029/2018GL081194>
- Pfleiderer, P. and Coumou, D. (2018) Quantification of temperature persistence over the Northern Hemisphere land-area. *Climate Dynamics*, 51(1–2), 627–637. <https://doi.org/10.1007/s00382-017-3945-x>
- Privé, N.C. and Errico, R.M. (2015) Spectral analysis of forecast error investigated with an observing system simulation experiment. *Tellus A*, 67(1), 25977. <https://doi.org/10.3402/tellusa.v67.25977>
- Prodhomme, C., Doblas-Reyes, F., Bellprat, O. and Dutra, E. (2016) Impact of land-surface initialization on sub-seasonal to seasonal forecasts over Europe. *Climate Dynamics*, 47(3–4), 919–935. <https://doi.org/10.1007/s00382-015-2879-4>
- Roads, J.O. (1986) Forecasts of time averages with a numerical weather prediction model. *Journal of the Atmospheric Sciences*, 43(9), 871–893. [https://doi.org/10.1175/1520-0469\(1986\)043<0871:FOTAWA>2.0.CO;2](https://doi.org/10.1175/1520-0469(1986)043<0871:FOTAWA>2.0.CO;2)
- Rodney, M., Lin, H. and Derome, J. (2013) Subseasonal prediction of wintertime North American surface air temperature during strong MJO events. *Monthly Weather Review*, 141(8), 2897–2909. <https://doi.org/10.1175/MWR-D-12-00221.1>
- Sillmann, J., Thorarindottir, T., Keenlyside, N., Schaller, N., Alexander, L.V., Hegerl, G., Seneviratne, S.I., Vautard, R., Zhang, X. and Zwiers, F.W. (2017) Understanding, modeling and predicting weather and climate extremes: challenges and opportunities. *Weather and Climate Extremes*, 18, 65–74. <https://doi.org/10.1016/j.wace.2017.10.003>
- Smith, L.A., Du, H., Suckling, E.B. and Niehörster, F. (2015) Probabilistic skill in ensemble seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141(689), 1085–1100. <https://doi.org/10.1002/qj.2403>
- Stefanon, M., D'Andrea, F. and Drobinski, P. (2012) Heatwave classification over Europe and the Mediterranean region. *Environmental Research Letters*, 7(1), 014023. <https://doi.org/10.1088/1748-9326/7/1/014023>
- Strazzo, S., Collins, D.C., Schepen, A., Wang, Q.J., Becker, E. and Jia, L. (2019) Application of a hybrid statistical–dynamical system to seasonal prediction of North American temperature and precipitation. *Monthly Weather Review*, 147(2), 607–625. <https://doi.org/10.1175/MWR-D-18-0156.1>
- Toth, Z. and Buizza, R. (2019). Weather forecasting: what sets the forecast skill horizon?, In: A. W. Robertson and F. Vitart (eds), *Sub-Seasonal to Seasonal Prediction*, pp. 17–45. Amsterdam: Elsevier, <https://doi.org/10.1016/B978-0-12-811714-9.00002-4> (to appear in print).
- Tripathi, O.P., Charlton-Perez, A., Sigmund, M. and Vitart, F. (2015) Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environmental Research Letters*, 10(10), 104007. <https://doi.org/10.1088/1748-9326/10/10/104007>
- Van Schaeybroeck, B. and Vannitsem, S. (2015) Ensemble post-processing using member-by-member approaches: theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141(688), 807–818. <https://doi.org/10.1002/qj.2397>
- Vautard, R. (1990) Multiple weather regimes over the North Atlantic: analysis of precursors and successors. *Monthly Weather Review*, 118(10), 2056–2081. [https://doi.org/10.1175/1520-0493\(1990\)118<2056:MWROTN>2.0.CO;2](https://doi.org/10.1175/1520-0493(1990)118<2056:MWROTN>2.0.CO;2)
- Vigaud, N., Robertson, A.W. and Tippett, M.K. (2017) Multimodel ensembling of subseasonal precipitation forecasts over North America. *Monthly Weather Review*, 145(10), 3913–3928. <https://doi.org/10.1175/MWR-D-17-0092.1>
- Vijverberg, S., Kraaijeveld, B., van der Wiel, K., Schmeits, M. and Coumou, D. (2020) Sub-seasonal statistical forecasts of eastern United States extreme temperature events. *Monthly Weather Review*. submitted
- Vitart, F. (2017) Madden–Julian Oscillation prediction and teleconnections in the S2S database. *Quarterly Journal of the Royal Meteorological Society*, 143(706), 2210–2220. <https://doi.org/10.1002/qj.3079>
- Vitart, F. and Robertson, A.W. (2018) The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmospheric Science*, 1(1), 3. <https://doi.org/10.1038/s41612-018-0013-0>
- Weigel, A.P., Baggenstos, D., Liniger, M.A., Vitart, F. and Appenzeller, C. (2008) Probabilistic verification of monthly temperature forecasts. *Monthly Weather Review*, 136(12), 5162–5182. <https://doi.org/10.1175/2008MWR2551.1>
- Wheeler, M.C., Zhu, H., Sobel, A.H., Hudson, D. and Vitart, F. (2017) Seamless precipitation prediction skill comparison between two global models. *Quarterly Journal of the Royal Meteorological Society*, 143(702), 374–383. <https://doi.org/10.1002/qj.2928>
- Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, International Geophysics Series Vol. 1003rd edition). Amsterdam: Academic Press.
- Wilks, D.S. (2018). Univariate ensemble postprocessing, *Statistical Postprocessing of Ensemble Forecasts*, pp. 49–89. Amsterdam: Elsevier, DOI 10.1016/B978-0-12-812372-0.00003-0, (to appear in print).
- Wilks, D.S. and Vannitsem, S. (2018). Uncertain forecasts from deterministic dynamics, *Statistical Postprocessing of Ensemble Forecasts*, pp. 1–13. Amsterdam: Elsevier.
- World Meteorological Organization (2015) *Seamless Prediction of the Earth System: From Minutes to Months*. Geneva: World Meteorological Organization.

- Wulff, C.O. and Domeisen, D.I.V. (2019) Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. *Geophysical Research Letters*, 46(20), 11520–11529. <https://doi.org/10.1029/2019GL084314>
- Yadav, P. and Straus, D.M. (2017) Circulation response to fast and slow MJO episodes. *Monthly Weather Review*, 145(5), 1577–1596. <https://doi.org/10.1175/MWR-D-16-0352.1>
- Yang, Z. and Villarini, G. (2019) Examining the capability of reanalyses in capturing the temporal clustering of heavy precipitation across Europe. *Climate Dynamics*, 53(3–4), 1845–1857. <https://doi.org/10.1007/s00382-019-04742-z>
- Ying, Y. and Zhang, F. (2017) Practical and intrinsic predictability of multiscale weather and convectively coupled equatorial waves during the active phase of an MJO. *Journal of the Atmospheric Sciences*, 74(11), 3771–3785. <https://doi.org/10.1175/JAS-D-17-0157.1>
- Yiou, P., Goubanova, K., Li, Z.X. and Nogaj, M. (2008) Weather regime dependence of extreme value statistics for summer temperature and precipitation. *Nonlinear Processes in Geophysics*, 15(3), 365–378. <https://doi.org/10.5194/npg-15-365-2008>
- Yoo, C., Johnson, N.C., Chang, C.-H., Feldstein, S.B. and Kim, Y.-H. (2018) Subseasonal prediction of wintertime East Asian temperature based on atmospheric teleconnections. *Journal of Climate*, 31, 9351–9366. <https://doi.org/10.1175/JCLI.D.17.0811.1>

How to cite this article: van Straaten C, Whan K, Coumou D, van den Hurk B, Schmeits M. The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Q J R Meteorol Soc.* 2020;146:2654–2670. <https://doi.org/10.1002/qj.3810>