

Ensemble forecasts of isolated and compound wind and precipitation extremes in Europe using HC-SWG (v3.1) and MA-SWG (v1.1) Stochastic Weather Generators

Meriem Krouma ¹ and Gabriele Messori ^{1,2,3}

¹Department of Earth Sciences, Uppsala University, Uppsala, Sweden

²Swedish Centre for Impacts of Climate Extremes (climes), Uppsala University, Uppsala, Sweden

³Department of Meteorology and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden

Correspondence: Meriem Krouma (meriem.krouma@geo.uu.se)

Abstract. Ensemble forecasts of extreme wind and precipitation provide essential information for early warning systems. In this study, we present two forecasting approaches that combine a stochastic weather generator (SWG) with atmospheric circulation analogs to forecast extreme precipitation and extreme wind speed in Europe. The first approach, which we term HC-SWG, combines ECMWF ensemble reforecasts with the stochastic weather generator to forecast extreme precipitation at
5 different locations in Europe. The second approach, which we term MA-SWG, uses multivariate atmospheric analogs as input to the SWG to forecast extreme 10m wind speed. These ensemble forecasts of precipitation and wind speed extremes display a higher forecast skill than ECMWF numerical reforecasts at lead times up to 10 days, using station data as the ground truth. As a final step, we evaluate the forecasted and observed frequencies of simultaneous and sequential precipitation and wind speed extremes in Europe, which are a class of high-impact compound events. Our forecasts yield comparable occurrence frequencies
10 to the observations.

1 Introduction

The isolated or compound occurrence of wind and precipitation extremes can result in large detrimental impacts on natural and socio-economic systems. Examples include ecosystems, agricultural production, and industry (Hao et al., 2022; Weinkle et al., 2018; Zscheischler et al., 2020). Wind and precipitation extremes can also cause fatalities and property losses, for
15 example, through extreme waves, storm surges, and flooding in low-lying coastal areas (Fang et al., 2021; Bevacqua et al., 2019). Improving the forecast of these and other extreme weather events, particularly at the medium range (1 to 10 days), is essential for issuing timely early warnings (Vitart et al., 2019).

Numerical weather prediction (NWP) models, based on a process-based modelling of the evolution of the atmosphere, have until recently been the dominant approach for weather forecasting (Magnusson et al., 2023). Their performance has improved
20 in the last decades thanks to more accurate initial conditions and parametrisations and higher resolution (Ben Bouallègue et al., 2019; Vitart et al., 2019; White et al., 2022). Recently, data-driven forecasting models have achieved comparable or better skill than NWP models (Bouallègue et al., 2024; Rasp et al., 2024).

However, both NWP and data-driven models face challenges in accurately forecasting extreme events (Haiden et al., 2023; Ben-Bouallegue, 2023; Domeisen et al., 2022). For instance, Olivetti and Messori (2024) highlighted the limitations of data-driven models in predicting cold extremes and noted significant regional variations in the forecast skill of different data-driven models. They ascribed this in part to the fact that the models are optimised for overall forecast skill at the cost of comparatively poorer performance for extreme events. The limitations of NWP arise from its constrained ability to resolve small-scale processes that influence meteorological variables such as precipitation or near-surface wind speed. Although substantial progress has been made in the parameterization of subgrid-scale phenomena, significant uncertainties remain within these schemes (Hersbach et al., 2020; Haiden et al., 2023). Another major challenge lies in the initialization of variables: while large-scale atmospheric fields can be accurately initialized using satellite and radiosonde observations, the initialization of surface variables is often hindered by data quality and incomplete global coverage of in situ surface data (Haiden et al., 2023). These challenges are compounded by the high computational costs associated with producing numerical forecasts at very high spatial resolution. This requires either high-resolution global models, or post-processing and downscaling to obtain high-resolution regional forecasts (Ben Bouallègue et al., 2023; Specq and Batté, 2020; Stevens et al., 2019; Alessi and DeGaetano, 2021). The latter significantly enhances forecast accuracy at local scales and for some challenging weather variables such as precipitation and wind. Machine learning or statistical techniques can also be combined with NWP models to correct forecast biases, downscale the forecasts and enhance forecast quality (Ben Bouallègue et al., 2023; Specq and Batté, 2020; Harris et al., 2022).

An alternative approach to NWP and data-driven models comes from stochastic weather generators (SWGs). These can be used as a forecasting tool or as a postprocessing tool, and can generate very large ensembles at a low computational cost (Ailliot et al., 2015; Yiou and Déandréis, 2019; Yiou, 2014; Brunner et al., 2021). SWGs have also been combined with circulation analogs – namely sets of similar states of the atmospheric circulation. This combined tool showed promising forecast skill for variables such as precipitation and temperature (Krouma et al., 2022, 2024; Atencia and Zawadzki, 2017; Blanchet et al., 2018) at subseasonal lead times of 25 to 30 days, as well as in forecasting climate indices such as the North Atlantic Oscillation and the Madden Julian Oscillation (Krouma et al., 2023; Yiou and Déandréis, 2019). SWGs are often used as a complement to numerical models, since they allow the simulation of large ensembles of variables at specific locations and can be calibrated at spatial scales that are generally not resolved by dynamical models. This makes them particularly suited for impact studies (Ailliot et al., 2015; Brunner et al., 2021).

In this study, we aim to use a SWG to produce ensemble forecasts of local extreme precipitation and wind speed events in Europe. We use two different forecasting approaches for the two variables. For extreme precipitation, we combine the SWG with the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble reforecasts (also known as hindcasts, or HC). This approach, which we term HC-SWG, uses analogs from the reforecasts, defined using 500 hPa geopotential height. It was tested in Krouma et al. (2024) to forecast subseasonal precipitation in Europe. Here, we apply it specifically to precipitation extremes. For extreme wind speed, we adopt the MA-SWG based on multivariate atmospheric analogs (MA). We developed the MA-SWG specifically to forecast the wind speed, after finding that analogs computed using a single atmospheric variable provided limited forecast skill for wind extremes. Additionally, extending the HC-SWG framework to jointly handle multiple

predictors and ensemble reforecasts for wind would substantially increase the model complexity, without a clear gain in forecast skill compared to the MA-SWG configuration adopted here.

The rest of the paper is structured as follows: Section 2 details the data used in our forecasts. Section 3 describes the forecasting process, including the circulation analogs computation and the two different versions of the SWG, and explains the verification metrics used to evaluate the forecast skill. The evaluation of the SWG ensemble forecasts, and their comparison to the ECMWF forecasts for precipitation and wind speed extremes as well as the compound forecast evaluation, are presented and discussed in Section 4. Section 5 outlines the main conclusions.

2 Data

We use daily data for precipitation and wind speed retrieved from the European Climate Assessment and Data (ECA&D) project for 9 locations across Europe (Bergen, Berlin, Brest, De Blit, Linköping, Madrid, Orly, Santander, and Stockholm) (Klein Tank et al., 2002) from 1960 to ~~2022~~-2022 (Fig.F1). The choice of those locations was based on: (i) ensuring diversity of meteorological conditions; and (ii) the availability of co-located observational data for precipitation and wind speed. The ECA&D data is used as ground truth.

We also use ERA5 reanalysis data (Hersbach et al., 2020), with a resolution of $0.25^\circ \times 0.25^\circ$ over 1960 to 2022. Hourly geopotential height at 500 hPa (Z500) and Sea Level Pressure (SLP) were used to obtain daily data over the region of $80^\circ\text{W} - 40^\circ\text{E}$, $90^\circ - 30^\circ\text{N}$. We consider this geographical domain to cover all the different analysis locations and to optimise computation time. For investigations focusing on specific locations or small regions, targeted domains could instead be used.

We further analyse reforecasts of Z500 collected from the ECMWF subseasonal to seasonal (S2S) database (Vitart et al., 2017) over the region of $80^\circ\text{W} - 40^\circ\text{E}$, $90^\circ - 30^\circ\text{N}$. The ECMWF reforecasts comprise an 11-member ensemble covering the past 20 years, initialised twice a week, and running up to 46 days lead time (Vitart et al., 2019). As initial conditions, the reforecasts use ERA5 and ORAS5 for the atmosphere and ocean, respectively. We consider the ensemble members at different lead times δ from 1 to 5 days. We chose the model version CY47R3, available from 2001 to 2021, with a horizontal resolution of 15 to 31 km and providing daily data, which contains ice and ocean initial conditions (Vitart et al., 2019). The ERA5 and S2S data are used to define analogs of the atmospheric circulation.

Finally, we considered ECMWF forecasts of precipitation and u and v components of 10 m wind from the THORPEX Interactive Grand Global Ensemble (TIGGE) database from 2017 to 2021 (Bougeault et al., 2010). We used the TIGGE database to evaluate our SWG forecasts, as it provides actual operational ensemble forecasts issued daily in near-real-time, with ~~higher~~-high spatial resolution (~~9~~-18 km) and more frequent initialisations (daily as opposed to bi-weekly) than the S2S database (Bougeault et al., 2010). This makes them ideal for verifying medium-range forecasts of extreme precipitation and extreme wind speed. The forecasts (referred to in the rest of the paper as ECMWF forecasts) for wind and precipitation have been bias-corrected ~~-For using a linear regression-based approach. For each variable and location, a linear model was fitted between the raw forecasts at the closest grid point and the corresponding station observations over the training period. The fitted regression parameters were then applied to all ensemble members, yielding a calibrated ensemble forecast. This approach~~

90 corrects systematic biases in both the mean and amplitude of the forecast while preserving ensemble spread. For comparison to the SWG forecasts, we considered ECMWF forecast data at the closest gridded points to the geographical coordinates of the studied stations as indicated in ECA&D.

We define extreme precipitation and extreme 10m wind speed as the precipitation (wind speed) that exceeds the empirical local (station-based) 95th percentile from the ECA&D data. For precipitation, the percentile was computed after excluding 95 values below 1mm/day. From the ECMWF forecasts, the same definition was applied, but using the 95th percentile of the climatological forecast distribution.

3 Methods

3.1 Forecasting tools: Analogs & SWG

To forecast precipitation and wind speed extremes over Europe, we leverage analogs of the atmospheric circulation and SWGs. 100 Here, we describe two configurations of the SWG. For extreme precipitation, we use the HC-SWG (Sect. 3.1.1), previously tested to forecast sub-seasonal precipitation by Krouma et al. (2024). HC-SWG combines the stochastic weather generator with NWP reforecasts. For extreme wind speed, we use the MA-SWG (Sect. 3.1.2), namely a stochastic weather generator combined with multivariate atmospheric analogs.

3.1.1 Extreme Precipitation forecast approach: HC-SWG

105 We use the ECMWF S2S ensemble reforecasts at lead times of $\delta = 1$ to 5 days to forecast extreme European precipitation (Figure 1). We first look in the ensemble reforecasts initialised at time t_0 and with lead time δ for analogs of the Z500 on a target date $t_0 + \delta$. We define analogs based on Euclidean distance, and consider only dates within a calendar window of 30 around the date of $t_0 + \delta$ yet in different years than $t_0 + \delta$. We then keep the $K = 20$ best analogs for each target day $t_0 + \delta$.

We next produce forecasts by generating random trajectories based on the identified analogs, following the procedure out- 110 lined in Krouma et al. (2024). The initialization point of our forecasts is set at $t_0 + \delta$, and each trajectory extends to time $t_0 + \delta + T$, with the lead time T ranging from 1 to 10 days. Beginning on day $t_0 + \delta$, we randomly select an analog t'_k among the $K = 20$ best analogs. The random selection of analogs of the day is carried out using weights that are proportional to the calendar difference between $t_0 + \delta$ and the analog dates, in order to ensure that time progresses (Yiou, 2014). We then replace $t_{0+\delta}$ with the selected analog of t'_k and repeat the operation T times.

115 The above process produces a random trajectory between $t_0 + \delta$ and $t_0 + \delta + T$. The procedure is repeated to simulate $N = 100$ trajectories, providing an initialised ensemble forecast. The SWG reforecasts are started every $\Delta t \approx T/2$ days between January 1st, 2002 and December 31st, 2021 (Figure 1a). The daily precipitation of each trajectory is time-averaged between $t_0 + \delta$ and $t_0 + \delta + T$. Hence, we obtain an ensemble of $N = 100$ forecasts of average precipitation over T days. From these average precipitation values, we define extremes as values in excess of the 95th percentile of the distribution for the full forecast period 120 (from 2002 to 2021). The HC-SWG in this paper has been improved compared to Krouma et al. (2024) by defining analogs

from the full ECMWF ensemble reforecast of Z500 compared to the use of the ensemble reforecast mean in Krouma et al. (2024). We additionally test the sensitivity of the HC-SWG forecast to δ instead of using a fixed δ value (Appendix D).

We illustrate the procedure with an example, where we generate an ensemble forecast of extreme precipitation starting on $t_0 = \text{February } 20^{th}, 2020$, with a forecast lead time $T = 6$ days. We set $\delta = 5$ days, so our starting point is the Z500 reforecast initialised on the 20th February, 2020 and with a lead time of 5 days, corresponding to $t_0 + \delta = \text{February } 25^{th}, 2020$. As first step, we identify the $K = 20$ best analogs of the Z500 field on February 25^{th} (within a ± 30 -day calendar window, excluding dates in 2020), and randomly select one analog weighted by calendar date similarity—for example, February 23^{rd} , 2011. We repeat this process for February 23^{rd} , 2011. We take the reforecasts initialised on February 23^{rd} , 2011, with lead time $\delta = 5$ days, and find the $K = 20$ best analogs for February 28th, 2011 (excluding dates in 2011) and select one randomly—for example, March 2^{nd} , 2008. We continue this iterative process until we have a timeseries of 6 dates. For each step t_k , we use the Z500 field to estimate daily extreme precipitation. The Z500 is used as a large-scale predictor to identify circulation patterns associated with extreme precipitation. The HC-SWG does not directly estimate precipitation from Z500 but conditions the selection of analogs on similar Z500 configurations, thereby linking large-scale dynamics to local extremes. We then calculate the mean precipitation over these 6 days to produce one forecast: the average of precipitation over February 20^{th} , 2020 to February 25^{th} , 2020. This entire procedure is repeated $N = 100$ times, each time generating a different random analog sequence, to create an ensemble of 100 forecasts of precipitation for the period February $20^{th} \sim 26^{th}$, 2020.

3.1.2 Extreme Wind forecast approach: MA-SWG

The multivariate analogs SWG relies on analogs computed using daily averages of Z500 and SLP from ERA5 reanalysis data (Figure 1b). These variables provide information on both the mid-tropospheric and surface large-scale circulation (Buizza and Leutbecher, 2015). We also tested adding the Z250 as a third variable, but found that this degraded our forecast skill.

We first compute Empirical Orthogonal Functions (EOFs) from anomalies of Z500 and SLP, analysing the two separately. The anomalies are defined relative to the daily mean climatology over the period from 1960 to 2022. We apply a cosine-of-latitude weighting during the EOF analysis to account for grid-area variations. We keep the s principal components that contain 90% of the variance ($s = 6$ for Z500, and $s = 10$ for SLP). Therefore, we have 16 daily time series from 1960 to 2022 corresponding to the principal components of the selected variables. Analogues for each target day are computed from this timeseries data, finding the closest tuples of values to the tuple of the target day. We again consider a window of 30 calendar days around the target day and exclude analogs in the same year as the target day. Then, we generate ensemble forecasts of wind speed using these analogs, following the same procedure described for the HC-SWG (Sect. 3.1.1). The difference is that in the MA-SWG we do not use reforecast data, and hence there is no δ . We again average the wind speed over the forecast lead time.

We take as example an initialization date $t_0 = 20^{th}$ February, 2020 and a forecast lead time $T = 6$ days. As a first step, we find the $K = 20$ best analogs of t_0 in the principal component space, and select one randomly (weighted by calendar difference), for example, 22^{nd} February, 2011. We next repeat this process iteratively to generate the rest of the trajectory up

to T days. We then compute the average wind speed over the forecast period and repeat the whole process $N = 100$ times to
155 produce an ensemble of 100 forecasts of average wind speed between February 20th and February 26th, 2020.

3.2 Compound extreme forecast

We based the compound extreme forecast of wind speed and precipitation in the 9 studied locations on the ensemble forecasts
generated from the HC-SWG and the MA-SWG (Figure 1c). To do so, each forecast ensemble is represented as a grid of binary
values, where “1” indicates an extreme event and “0” indicates no extreme event (Sutanto et al., 2020). The compound extreme
160 forecast is derived by identifying overlapping occurrences of extremes in both precipitation and wind speed ensembles. We
also checked the cases of sequential extremes, namely extreme precipitation events and extreme wind speeds occurring in
succession within lags of 1 to 5 days of one another.

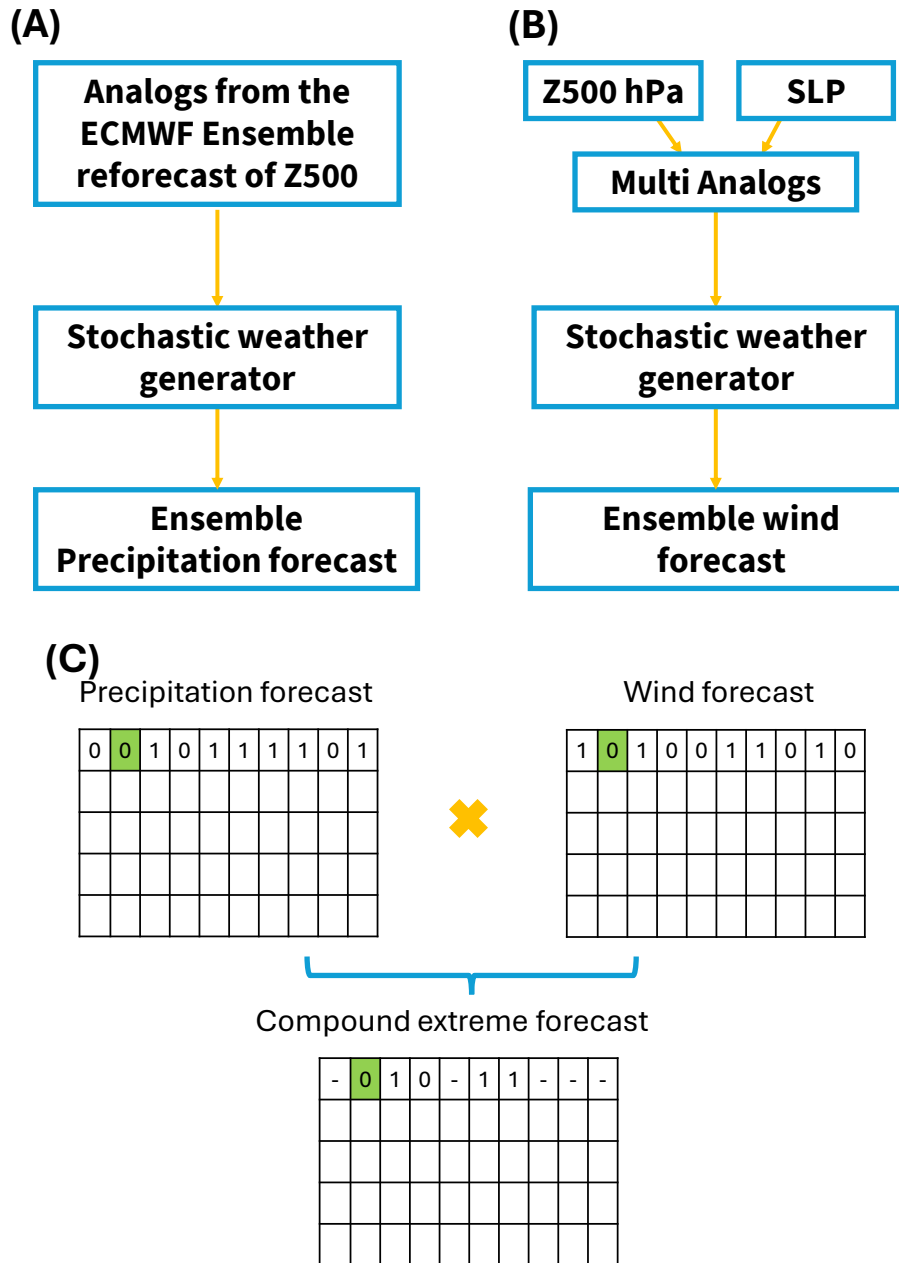


Figure 1. Illustration of the forecast process. (a) The HC-SWG used to forecast extreme precipitation; (b) the MA-SWG used to forecast wind speed, and (c) the ensemble forecast of the co-occurrence of precipitation and wind speed extremes.

3.3 Forecast Evaluation

We evaluate the ability of the HC-SWG and MA-SWG to forecast extreme precipitation and wind speed, respectively, by using the Symmetric Extremal Dependence Index (SEDI) and the Peirce Skill Score (PSS). These two metrics are particularly suited for rare events (Ferro and Stephenson, 2011; Magnusson et al., 2014; Stephenson et al., 2008a). Unlike traditional scores, they emphasise event discrimination and forecast skill under low base-rate conditions and are less influenced by class imbalance, thus providing a more reliable assessment of the model's ability to detect extremes (Stephenson et al., 2008a).

The SEDI accounts for hits (H), false alarms (F), misses (M), and correct rejections (C), and is defined as:

$$SEDI = \frac{\log(H/M) - \log(F/C)}{\log(H/M) + \log(F/C)}. \quad (1)$$

We assessed forecast skill beyond random chance using the PSS (Stephenson, 2000; Manzato, 2007). The PSS ranges from 0 to 1, where $PSS = 1$ corresponds to a perfect forecast, and $PSS = 0$ indicates that the forecast performs no better than random chance. We calculated the PSS as:

$$PSS = POD - FAR, \quad (2)$$

where POD is the Probability of Detection and FAR is the False Alarm Ratio, which in this context is sometimes referred to as POFD (Probability Of False Detection) in the literature. The FAR measures the frequency of false alarms relative to the total number of forecasted extreme events, while the POD quantifies the fraction of observed extreme events that were correctly forecasted. These metrics are defined as:

$$FAR = \frac{F}{H + F}, \quad (3)$$

$$POD = \frac{H}{H + M}. \quad (4)$$

High values of POD and FAR indicate overprediction of extremes, low values of POD and FAR suggest missing extremes, and high values of POD with low values of FAR indicate good forecasting skill (Wilks, 2011b). ~~We compute SEDI~~

~~We compute the SEDI and the PSS to evaluate forecast skill in the upper tail of the precipitation (wind speed) distribution using quantile thresholds of 20Q, 70Q, and PSS considering different thresholds, the 20th, 70th, and 90th 90Q. These thresholds correspond to quantiles of the distributions of extreme precipitation (and extreme wind speed), to evaluate the ability of the HC-SWG (MA-SWG) to forecast the most extreme values of the extreme distribution, and not of the full precipitation (wind speed) distribution. Thus, the 20Q threshold corresponds to relatively moderate extremes while the 70Q and 90Q thresholds highlight the most extreme precipitation events. These conditional quantiles are used to assess forecast skill at different levels of extremeness and should therefore not be interpreted as percentiles of all precipitation (wind speed extremes) values. Both SEDI and PSS were applied to evaluate the forecast skill of the SWGs.~~

Finally, we compared the ensemble forecasts of the HC-SWG and the MA-SWG to the ECMWF ensemble forecast using the Brier skill score (BSS). The BSS is computed between the Brier score (see Appendix C) of the HC-SWG (MA-SWG) forecast

and the Brier score of the ECMWF precipitation (wind speed) forecast, which we consider as a benchmark, as follows:

$$195 \quad BSS = 1 - \frac{BS_{SWG_s}}{BS_{ECMWF}}. \quad (5)$$

Values above 0 indicate that the SWG forecasts are better than ECMWF forecasts; a value of zero indicates equal performance of the different forecasts, and negative values indicate that the ECMWF forecasts outperform the SWG forecasts (Hersbach, 2000).

200 To investigate further the difference between the SWGs forecasts and the ECMWF forecasts, we compute the cumulative distribution functions (CDFs) of both the SWG and ECMWF ensemble forecasts.

4 Results

4.1 Evaluation of the extreme precipitation forecasts

We evaluate the HC-SWG's forecasting skill for extreme precipitation over Europe. We focus here on the results using $\delta = 5$ days. Results showing the sensitivity of the forecast performance to different δ , [the lead time of the Z500 ensemble reforecast - used as input to the HC-SWG](#), are provided in Appendix D.

210 The HC-SWG reproduces closely the time series of the observed extreme precipitation amounts from 2002 to 2021 at lead times of up to 10 days. Figure 2 shows the results for Linköping (Sweden). The HC-SWG forecasts are particularly good for the moderate extreme events, while they display an overestimation of the most extreme precipitation values (upper tails in Figure 2b, d, f). This behavior is likely linked to the stochastic nature of the HC-SWG, which may require further calibration to refine its ability to predict the highest extreme precipitation values accurately. As the lead time T increases, the amplitude of both the observed and forecasted extreme precipitation events decreases. Indeed, longer forecast horizons imply averaging extreme precipitation across more days, and thus lead to a smoothing effect.

215 The other stations that we consider present results in line with those for Linköping (Figure 3). We again find a strong forecast performance for moderate extreme events and widespread overestimation for the most extreme events, with the discrepancy growing larger at longer lead times (Figure 3). There is some variability across stations, with some (e.g. Bergen) displaying larger forecast errors while others (e.g. Berlin) display better agreement. Nonetheless, the qualitative overestimation pattern for the most extreme events is similar across all stations.

220 Overall, the results illustrate the ability of the HC-SWG to accurately forecast heavy precipitation events at medium-range timescales, with some variations in performance across different stations and a clearly degraded performance for the most extreme events.

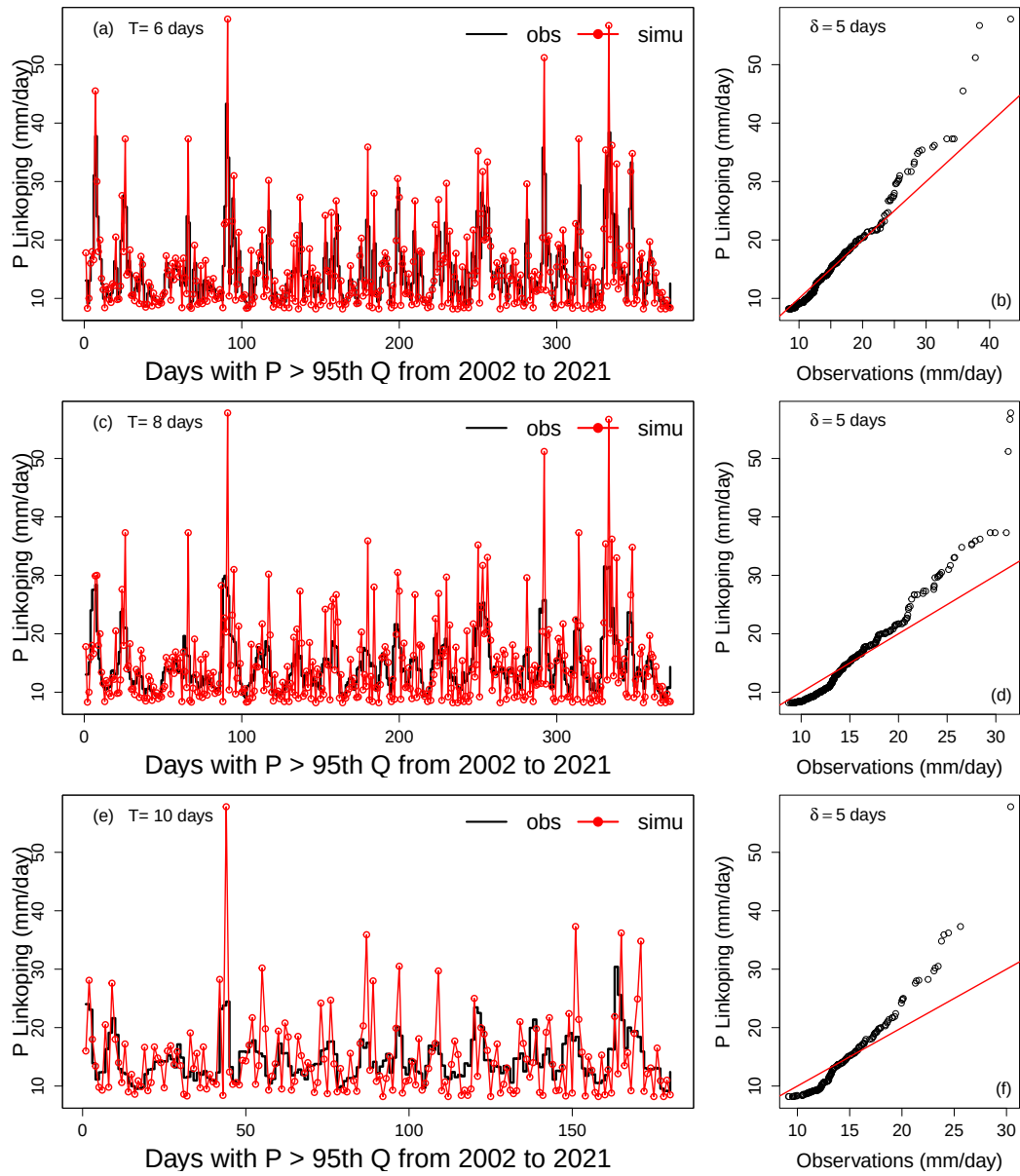


Figure 2. Comparison of observed and forecasted extreme precipitation events ($P > 95\text{th } Q$) using the HC-SWG at lead times $T = 6, 8$ and 10 days for Linköping. (a, c, e) display observed (black) and forecasted (red, defined as the median of the 100 members) mean extreme precipitation values (mm/day) from 2002 to 2021. (b, d, f) present scatter plots comparing observations and forecasts, with the red diagonal lines representing a perfect 1:1 relationship.

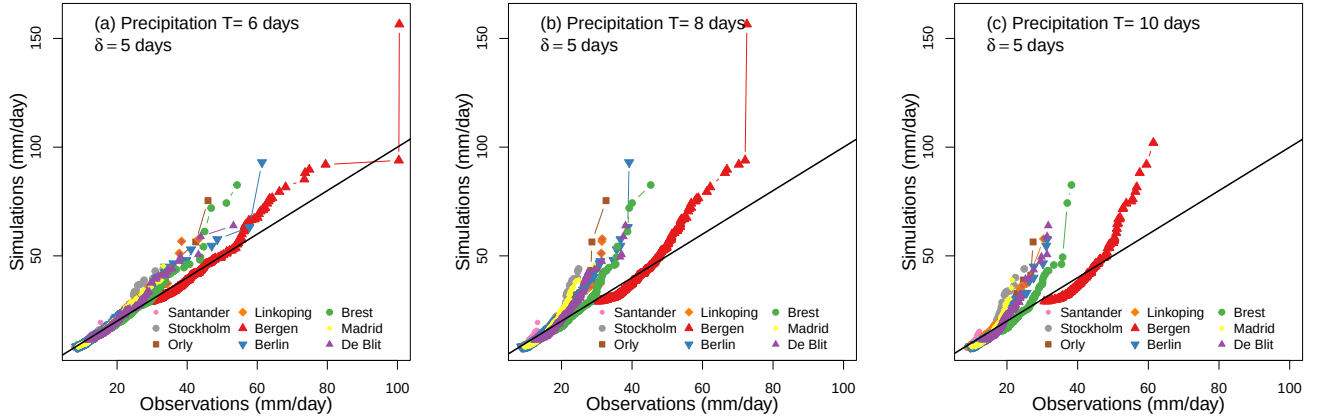


Figure 3. Comparison of observed and forecasted extreme precipitation events using the HC-SWG at lead times $T = 6$ (a), 8 (b) and 10 (c) days for all stations considered here. The panels present scatter plots comparing observed and forecasted extreme precipitation values (mm/day) from 2002 to 2021 on days exceeding the local 95th percentile. The black diagonal lines represent a perfect 1:1 relationship.

To quantify the performance of the HC-SWG forecasts, we compute the PSS and SEDI for different quantiles of extreme precipitation going from the 20th to the 90th quantiles (Figure 4). PSS quantifies the added value of the HC-SWG forecasts compared to random forecast. For all stations, PSS remains close to one across all the percentiles and lead times (Figure 4 a, c, e). Indeed, the forecasts have relatively low FAR and relatively high POD, resulting in high PSS values and indicating a good forecast skill. The forecasts display a relatively stable PSS for moderate (exceeding the 70th quantile of the distribution of extreme precipitation) and most extreme events (exceeding the 90th quantile of the distribution of extreme precipitation) for different lead times.

SEDI accounts for hits, false alarms, misses and correct rejections. There is a slight degradation of performance with increasing lead time, but this is highly variable across stations, with a number of stations showing higher SEDI values at longer lead times (Figure 4, b, d, f). SEDI values remain relatively positive even for the most extreme events. Indeed, even though the magnitude of these events is overpredicted in the forecasts, they qualify as exceeding a given quantile in both the forecasts and observations, and thus do not count as false alarms.

These results again indicate a strong forecast performance of the HC-SWG, albeit with some lead-time and location dependence.

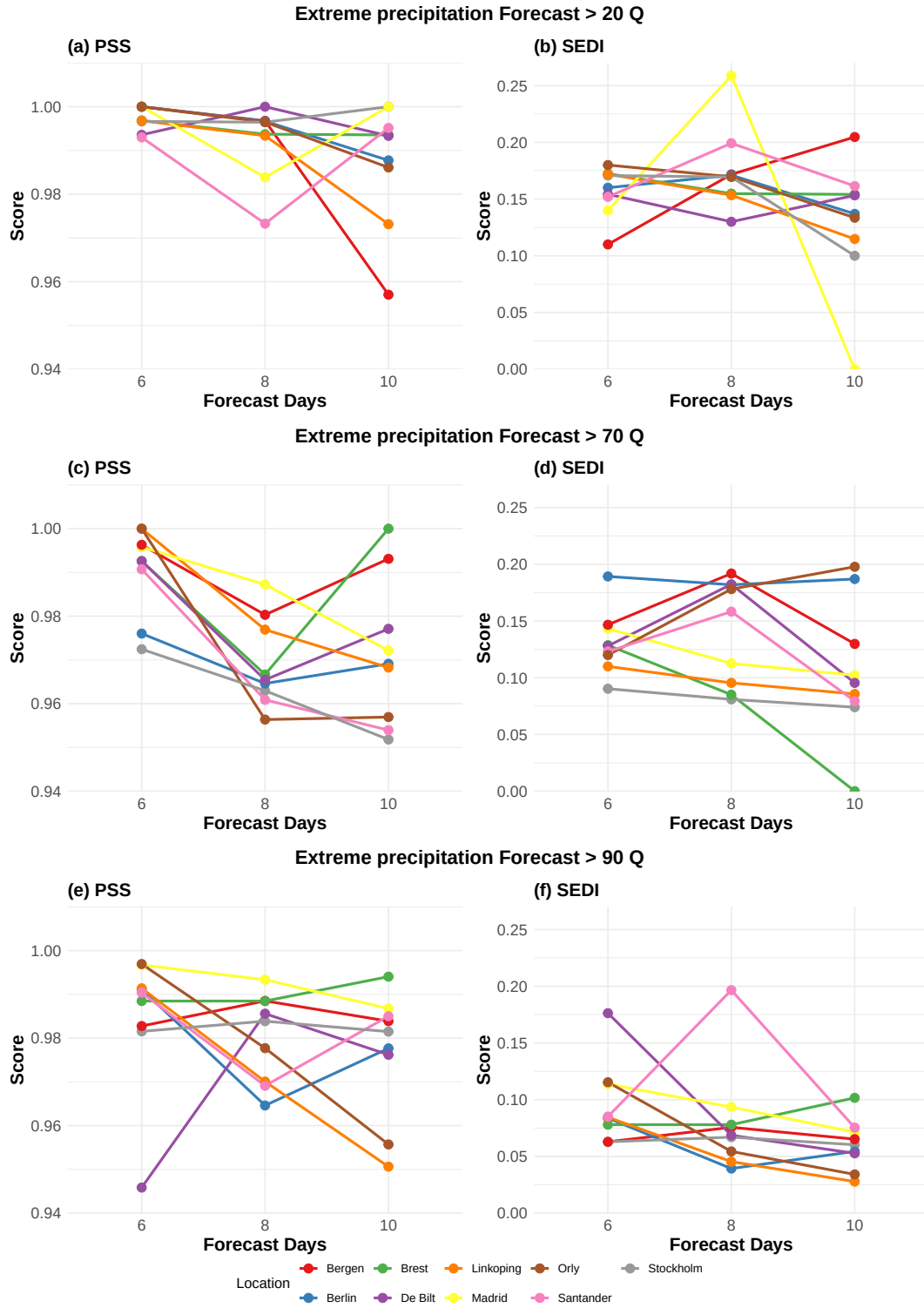


Figure 4. Extreme precipitation forecast skill for HC-SWG evaluated using PSS (a, c, e) and SEDI (b, d, f) at lead times $T = 6$ (a), 8 (b) and 10 for all stations considered here. We consider separately events above the 20th, 70th and 90th quantiles of the extreme precipitation days (i.e. > 95Q of the full distribution).

235 We next compare the HC-SWG ensemble forecast of extreme precipitation to the ECMWF forecasts for the 9 studied stations using the BSS at different lead times T (Figure 5). The BSS values are between 0.4 and 0.98 for all lead times going from 6 to 10 days, which indicates that HC-SWG ~~outperforms~~ shows improvement against ECMWF forecasts for extreme precipitation. The BSS values decrease with lead time and are spatially dependent.

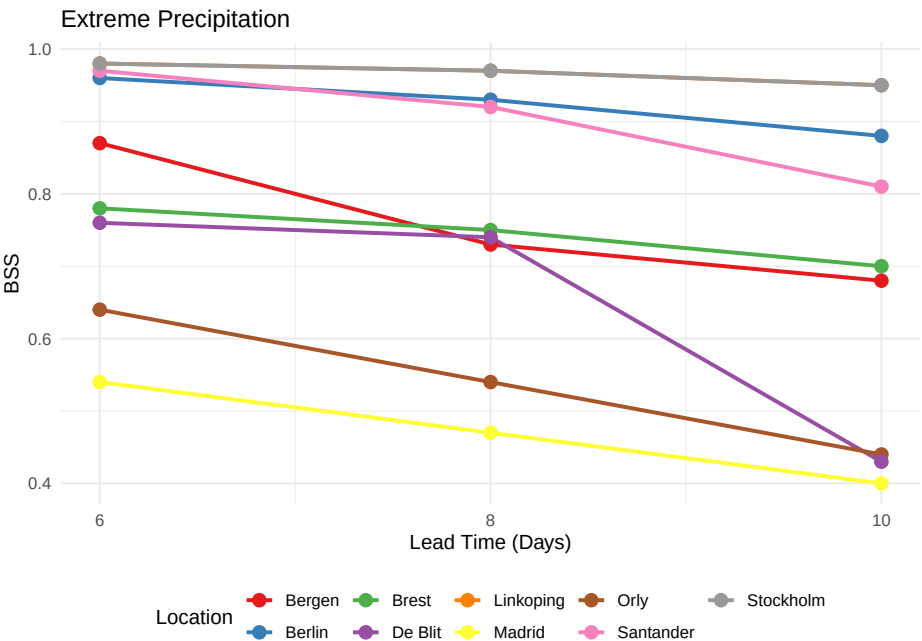


Figure 5. BSS between HC-SWG and the ECMWF forecasts of extreme precipitation for different locations across Europe at different lead times, going from $T = 6$ days to $T = 10$ days, from 2017 to 2021.

240 To better understand the differences between the two forecasts, we consider the CDFs of the forecasted versus observed extreme precipitation. We use Stockholm and Brest as example stations (Figure 6). The ECMWF forecast shows very steep CDFs, indicating an under-dispersed ensemble that does not sufficiently capture precipitation variability, leading to an overconfident forecast. In contrast, the HC-SWG forecasts follow the observed CDFs more closely, preserving the distribution’s spread and better representing extremes. As the forecast lead time T increases from 6 to 10 days, the ECMWF forecasts remain tightly clustered around a narrow range of precipitation values, suggesting the ensemble struggles to account for increased uncertainty

245 at longer lead times. Meanwhile, the HC-SWG forecast continues to align well with observations and provides a more reliable probabilistic representation of precipitation. This suggests that incorporating a stochastic approach like HC-SWG can improve ensemble forecast spread and better capture precipitation extremes. For the rest of the stations, we show results in Table B1 in Appendix B.

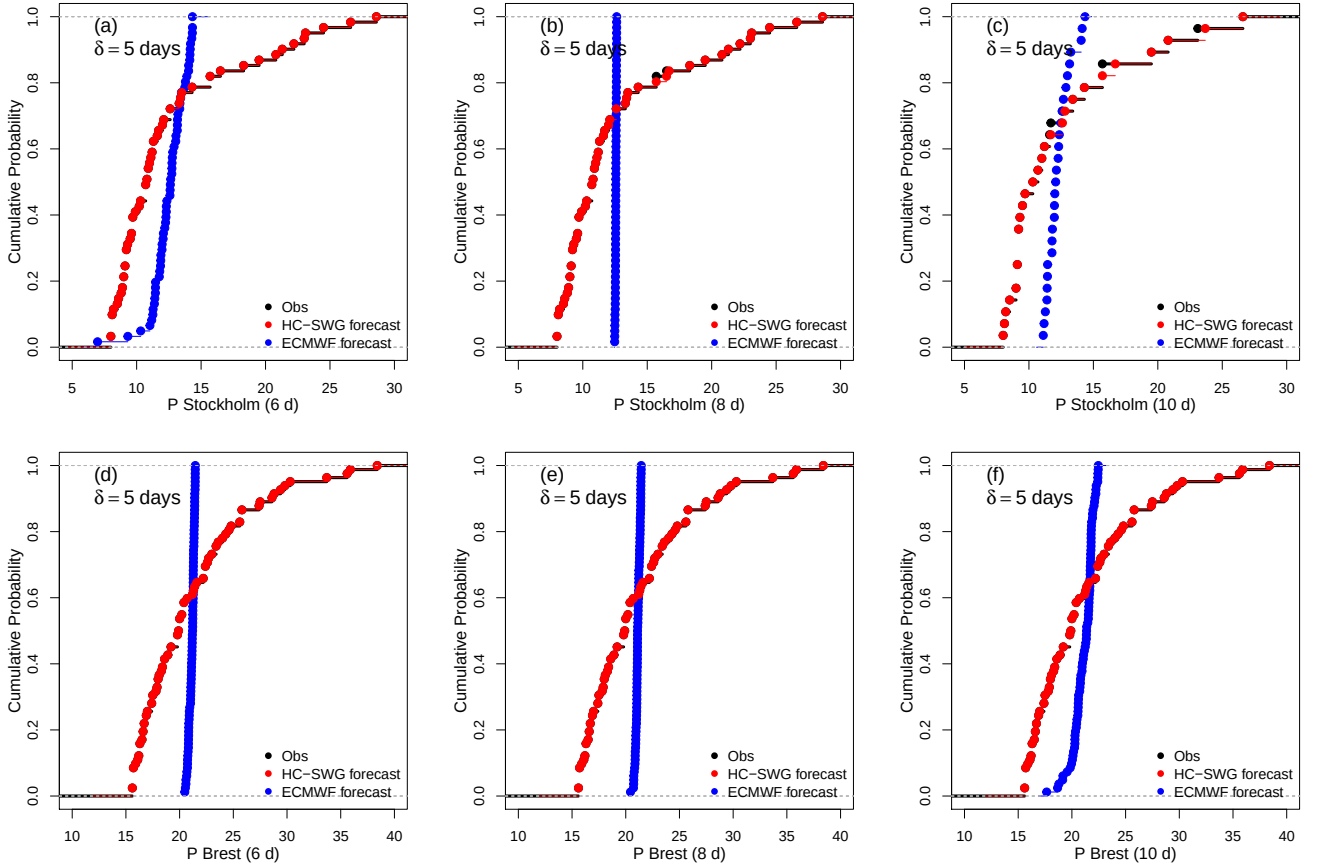


Figure 6. Cumulative Distribution Functions (CDFs) of observed (black) and forecasted (HC-SWG, red; ECMWF, blue) extreme precipitation. We consider Stockholm (a–c) and Brest (d–f) at lead times $T = 6$ (a, d), 8 (b, e) and 10 (c, f).

4.2 Evaluation of the extreme wind speed forecasts

250 Unlike HC-SWG, the MA-SWG approach has not been previously tested in the literature. We therefore first test the ability of the MA-SWG to forecast the wind speed on all days, and find that the forecasts provide considerable added value when compared to climatology (Appendix E). We next consider the forecast skill for extreme wind speed only. As for extreme precipitation, we first consider the performance at one example station, here Santander (Figure 7).

At all T lead times, the MA-SWG reproduces well the timing of the extreme events. However, it overestimates their magnitude, particularly for the most intense events (Figure 7 b, d, f). This is more pronounced for $T = 3$ days (Figure 7 b) than for
255 $T = 5$ and 10 days (Figure 7 d, f). The MA-SWG thus captures the temporal occurrence of extremes, but it tends to exaggerate their intensity, as we also saw for HC-SWG and extreme precipitation.

We note a similar tendency of MA-SWG to overestimate the most extreme wind speeds also at the other studied stations (Figure 8). At $T = 3$ days (Figure 8a), the overestimation is most visible for Santander, Stockholm and Linköping, and similar

260 patterns are visible at $T = 5$ and 10 days (Figure 8 b, c). However, the overall forecast bias appears to decrease in average with forecast lead time T , pointing to a stability in tail reliability for longer-range wind forecasts (see Figure A1 in Appendix A).

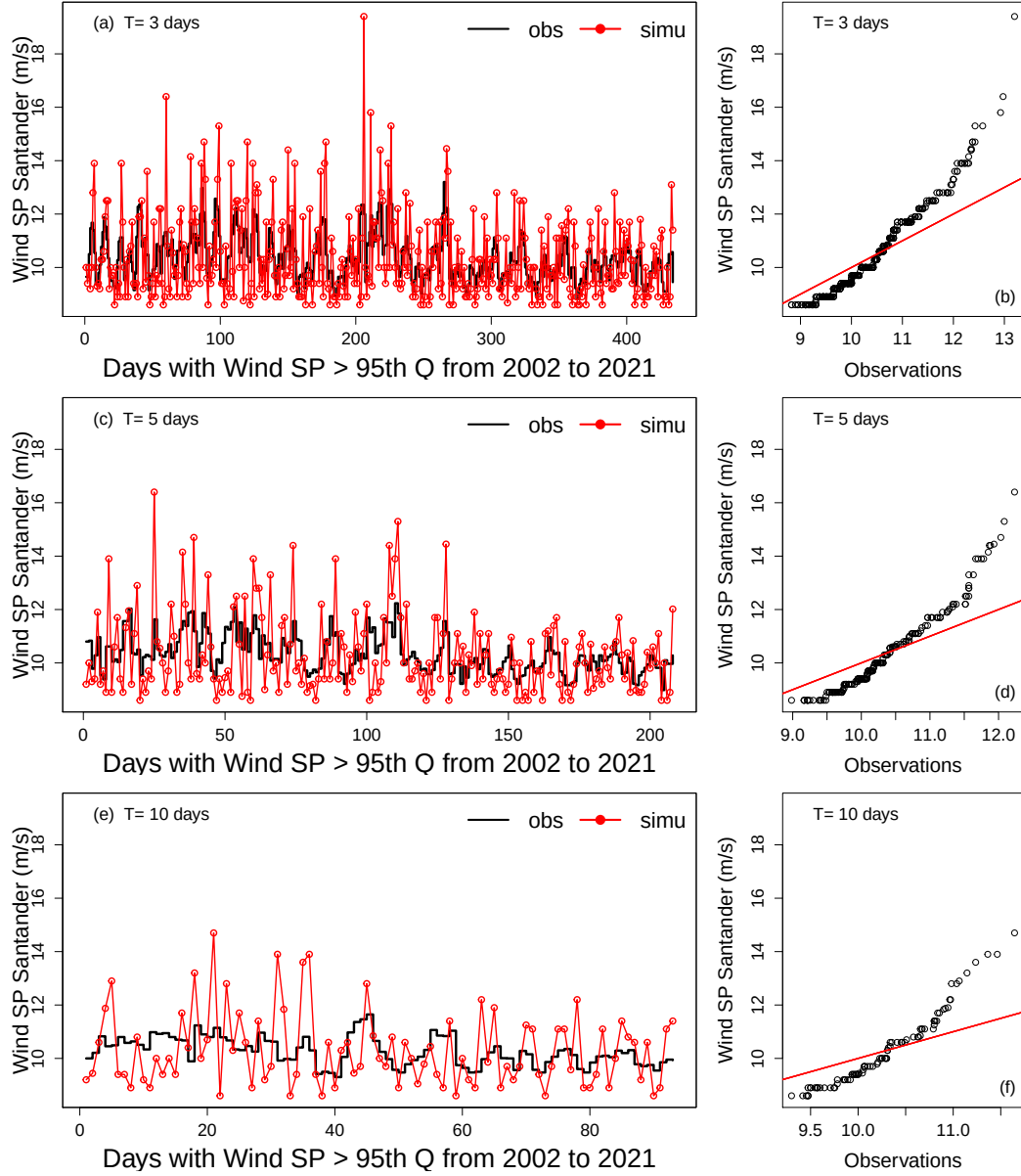


Figure 7. Comparison of observed and forecasted extreme wind speed events using the MA-SWG at lead times $T = 3, 5$ and 10 days for Santander. (a, c, e) display time series of observed (black) and forecasted (red) mean wind speeds (m/s) from 2002 to 2021 for forecasts whose mean wind speed exceeds the local 95th percentile. (b, d, f) present scatter plots comparing observations and forecasts, with the red diagonal lines representing a perfect 1:1 relationship.

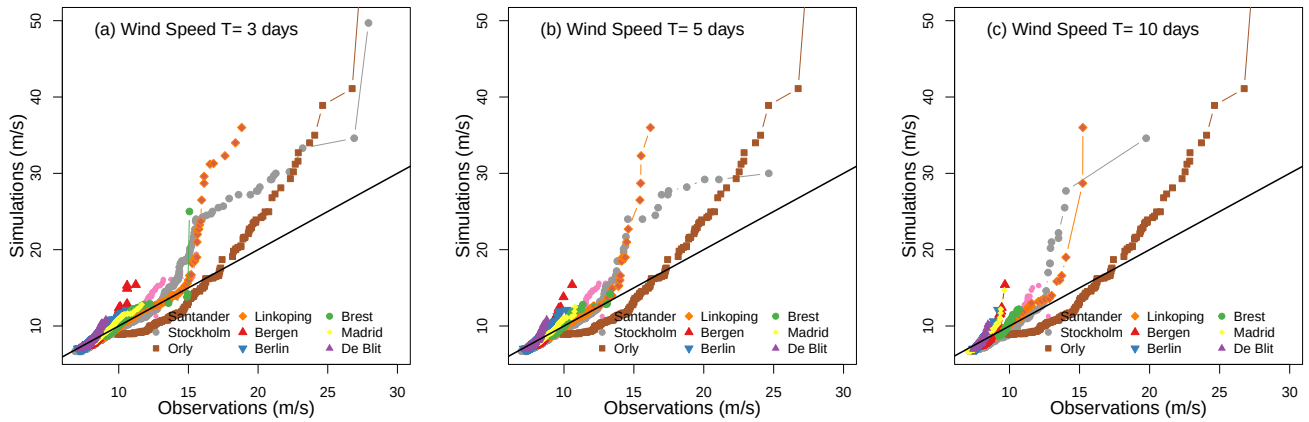


Figure 8. Comparison of observed and forecasted extreme wind speed events using the MA-SWG at lead times $T = 3$ (a), 5 (b) and 10 (c) days for all stations considered here. The panels present scatter plots comparing observed and forecasted wind speed values (m/s) from 2002 to 2021 on days exceeding the local 95th percentile. The black diagonal lines represent a perfect 1:1 relationship.

We next compute PSS and SEDI for the extreme wind forecasts (Figure 9). The PSS scores (Figure 9a, c, e) remain consistently high across most stations and lead times, with values close to 1. This indicates that the forecasts are highly skillful in forecasting both extreme and very extreme wind events. Same as for the extreme precipitation forecasts, the PSS for wind speed remains stable. Madrid (yellow line in Figure 9a, c, e) is a clear outlier and displays systematically lower scores than any of the other stations. This may arise from the fact that Madrid displays a higher wind speed variability compared to the other stations.

The SEDI scores (Figure 9b, d, f) vary widely across stations. For stations such as Bergen, they show a steep decline at longer lead times. This suggests that, while the forecasts remain skilful in detecting extreme wind events (as shown by the high PSS values), their reliability diminishes beyond 5 days, in line with increasing uncertainty in the evolution of synoptic and mesoscale atmospheric features at extended lead times. SEDI also declines for higher percentiles, indicating that while the MA-SWG correctly forecasts extreme wind speed events in general, its ability to confidently predict the most severe cases is reduced. For more moderate extremes (> 70 th quantile), the SEDI varies on average from 0.74 to 0.30 (Figure 9 d), while the average for extremes (> 90 th quantile) moves from 0.34 to 0.26 (Figure 9 f) from 3 to 10 days.

Overall, the results indicate that while the forecasts maintain strong discrimination skill (PSS) for extreme wind events, their reliability (SEDI) varies across stations and decreases for the most extreme cases.

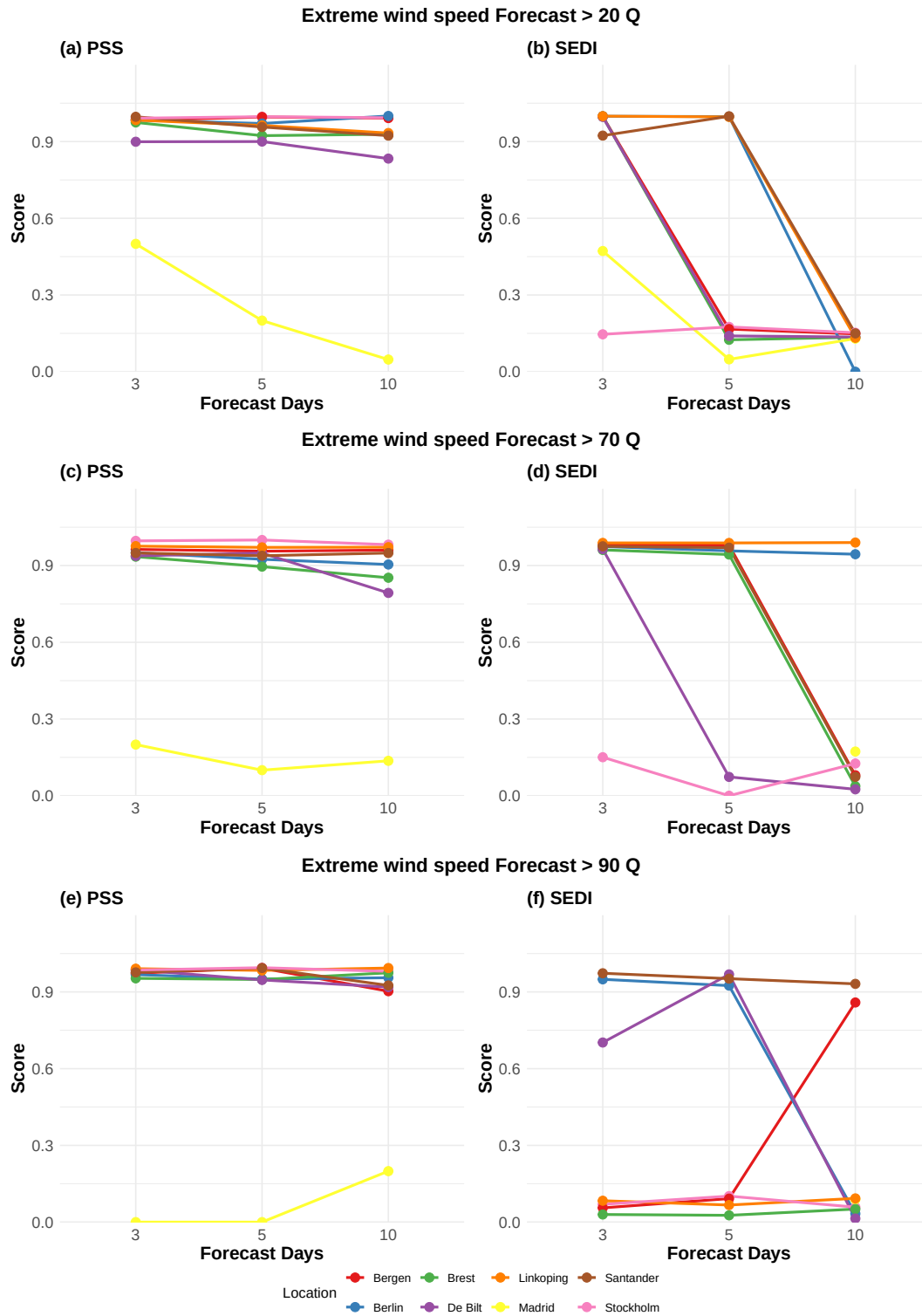


Figure 9. Extreme wind speed forecast skill for MA-SWG evaluated using PSS (a, c, e) and SEDI (b, d, f) at lead times $T = 3$ (a), 5 (b) and 10 for all stations considered here. We consider separately events above the 20th, 70th and 90th quantiles of the extreme wind speed days (i.e. > 95Q of the full distribution).

We next compare the MA-SWG and ECMWF forecasts for 10 m wind speed extremes at different lead times using the BSS (Figure 10). The BSS for all the studied stations is positive, which indicates that MA-SWG ~~outperforms ECMWF~~ shows improvement against the ECMWF forecast. The magnitude of the improvement varies across stations, and generally decreases with increasing lead time. At $T = 3$ days, values range between 0.62 and 0.97, while they decrease to between 0.41 and 0.9 at $T = 10$ days (Figure 10). This result should be interpreted in light of potential limitations associated with our bias-correction method applied to the ECMWF wind speed forecasts.

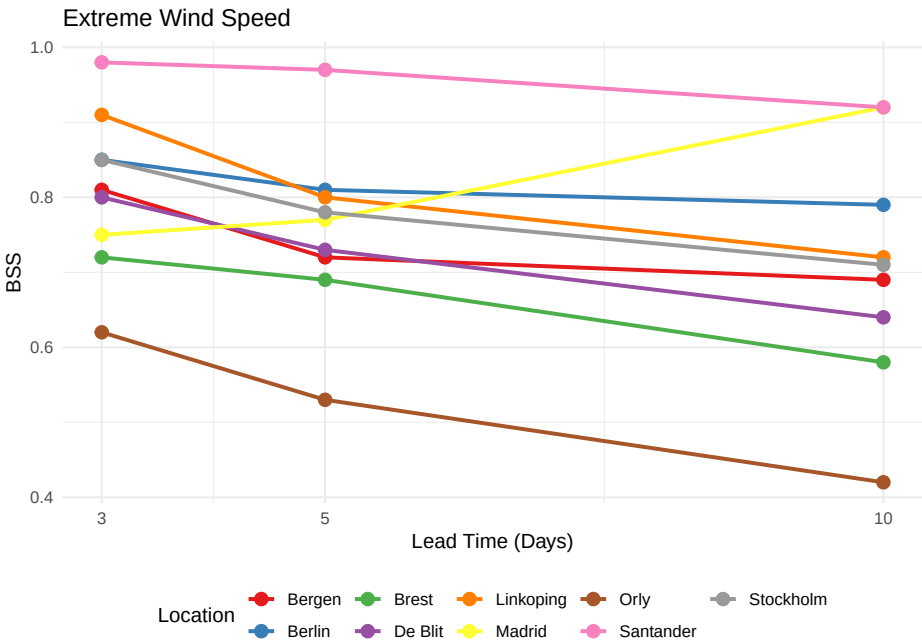


Figure 10. BSS between the MA-SWG forecast and the ECMWF forecast of extreme wind speed for different locations across Europe at different lead times, going from $T = 3$ days to $T = 10$ days, from 2017 to 2021.

As for the HC-SWG extreme precipitation forecasts, we compare the CDFs of the MA-SWG and ECMWF wind speed forecasts to the CDFs of the observations, showing Stockholm and Brest as examples (Figure 11). For both stations, the CDFs of observations and the MA-SWG forecasts show a close agreement. The ECMWF forecasts instead underestimate the most extreme wind speed values and overestimate the lowest values, the latter in particular for Stockholm at 3 and 5 days (Figure 11 a, b). For Brest, the CDFs of the ECMWF forecasts are generally closer to observations, but they keep underestimating the most extreme values. As for the extreme precipitation, the results for the other stations are represented in Table B1 in Appendix B.

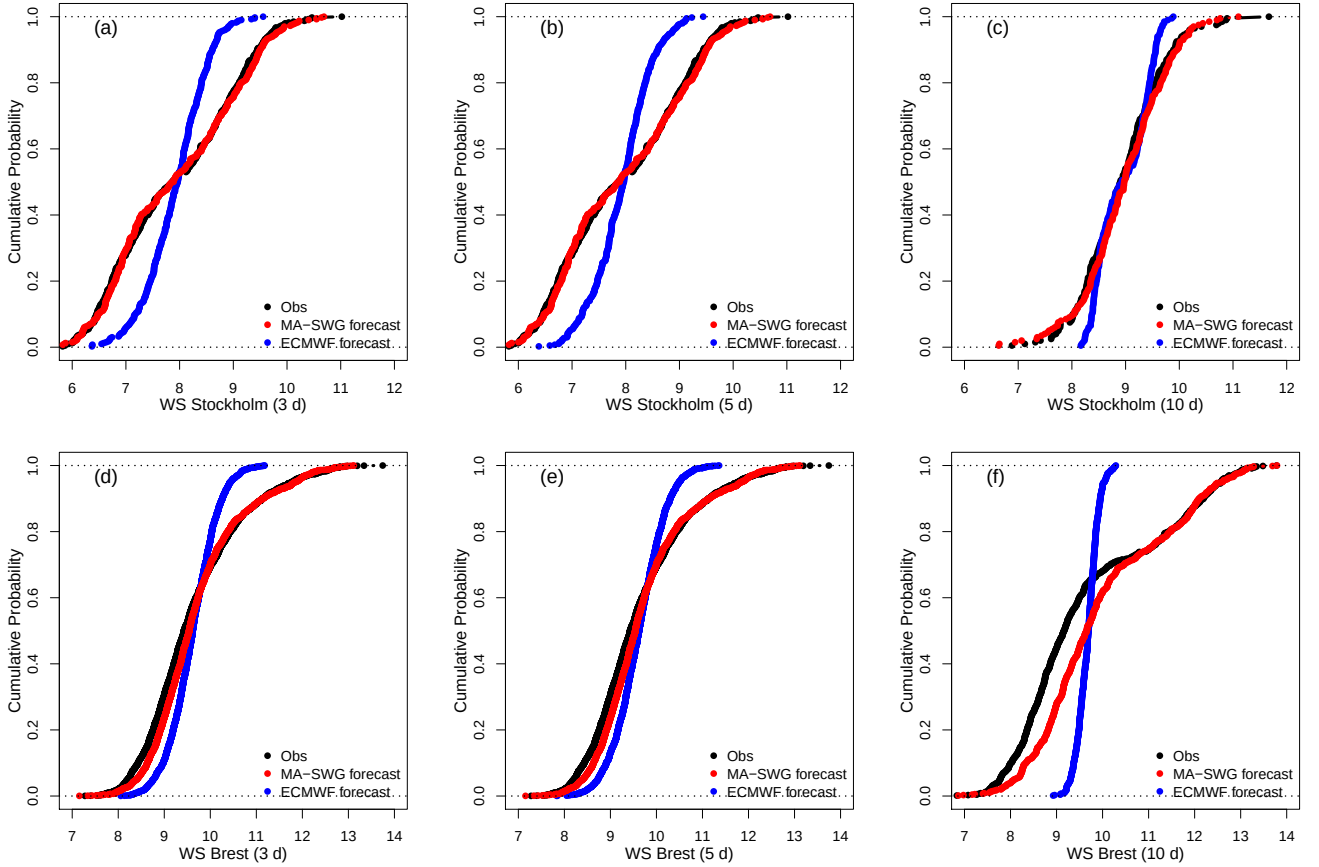


Figure 11. Cumulative Distribution Functions (CDFs) of observed (black) and forecasted (MA-SWG, red; ECMWF, blue) extreme wind speed. We consider Stockholm (a–c) and Brest (d–f) at lead times $T = 3$ (a, d), 5 (b, e) and 10 (c, f).

290 4.3 Assessment of Compound forecasts

We now evaluate the capacity of the SWG to forecast extreme precipitation and extreme wind speed events, which occur simultaneously or sequentially, following the procedure described in Section 3.1.

We find that, for a forecast of $T = 10$ days, the SWG is able to reproduce very closely the observed frequency of occurrence of simultaneous extreme precipitation and wind speed, as shown in Table 1. In particular, the SWG correctly identified three of
 295 the four stations that display no such events in observations, and reproduces a very low frequency of occurrence for the fourth station.

We performed a similar analysis for sequential events. Figure 12 shows the number of extreme precipitation events followed by extreme wind, and extreme wind events followed by extreme precipitation, using time windows ranging from 1 to 5 days. The number of sequential events by definition increases with longer time windows, as these allow more chances for one type
 300 of event to follow the other. The results reveal strong spatial variability in the number of both types of events. For instance,

Table 1. Simultaneous occurrences of extreme precipitation and extreme wind speed events from 2002 to 2021 in observations and SWG forecasts. The percentages are relative to the total number of forecasted or observed extreme precipitation and wind speed events. The total number of extremes at each location in the SWG forecasts is indicated in parentheses.

Location	Simultaneous events - SWG forecast (%)	Simultaneous events - observations (%)
Bergen	5.73 (n = 146)	5.81
Berlin	4.78 (n = 125)	6.48
Brest	5.68 (n = 88)	6.50
De Blit	4.80 (n = 83)	4.55
Orly	2.40 (n = 90)	0
Linköping	0 (n = 168)	0
Madrid	0 (n = 143)	0
Stockholm	0 (n = 170)	0
Santander	33.30 (n = 144)	36.60

locations like Brest, Bergen, De Bilt, and Santander (Figure 12 (a, c, d, i)) show a higher number of extreme wind events followed by extreme precipitation events, consistent with Atlantic-driven storm systems that often bring strong winds before heavy rain. In contrast, Berlin, Linköping, Madrid and Orly (Figure 12 (b, g, e, f)) show comparable or higher numbers of extreme precipitation events followed by extreme wind speed than extreme wind speed events followed by extreme precipitation. Finally, Stockholm (Figure 12 (h)) shows very few sequential events.

We next compared the sequential extremes in SWG forecasts to those in observations (Figure 13). Compared to observations, the SWGs capture the general frequency of occurrence of sequential events reasonably well, especially for Stockholm, Orly, Madrid and Bergen (Figure 13 (c) and (f)). Overestimations are notable in locations like Santander and Brest for both types of sequential events, particularly for longer time windows (3–5 days). De Bilt shows a systematic overestimation of extreme windspeed events followed by extreme precipitation (Figure 13f). There is a single instance of SWG forecasts underestimating the frequency of occurrence of sequential extremes, namely De Bilt for rain extremes followed by wind speed extremes for a time window of two days (Figure 13c).

The above results suggest that the SWG forecasts, while displaying some biases, nonetheless capture key regional features of compound wind-precipitation extremes. They thus offer a promising approach for simulating and predicting compound hazards.

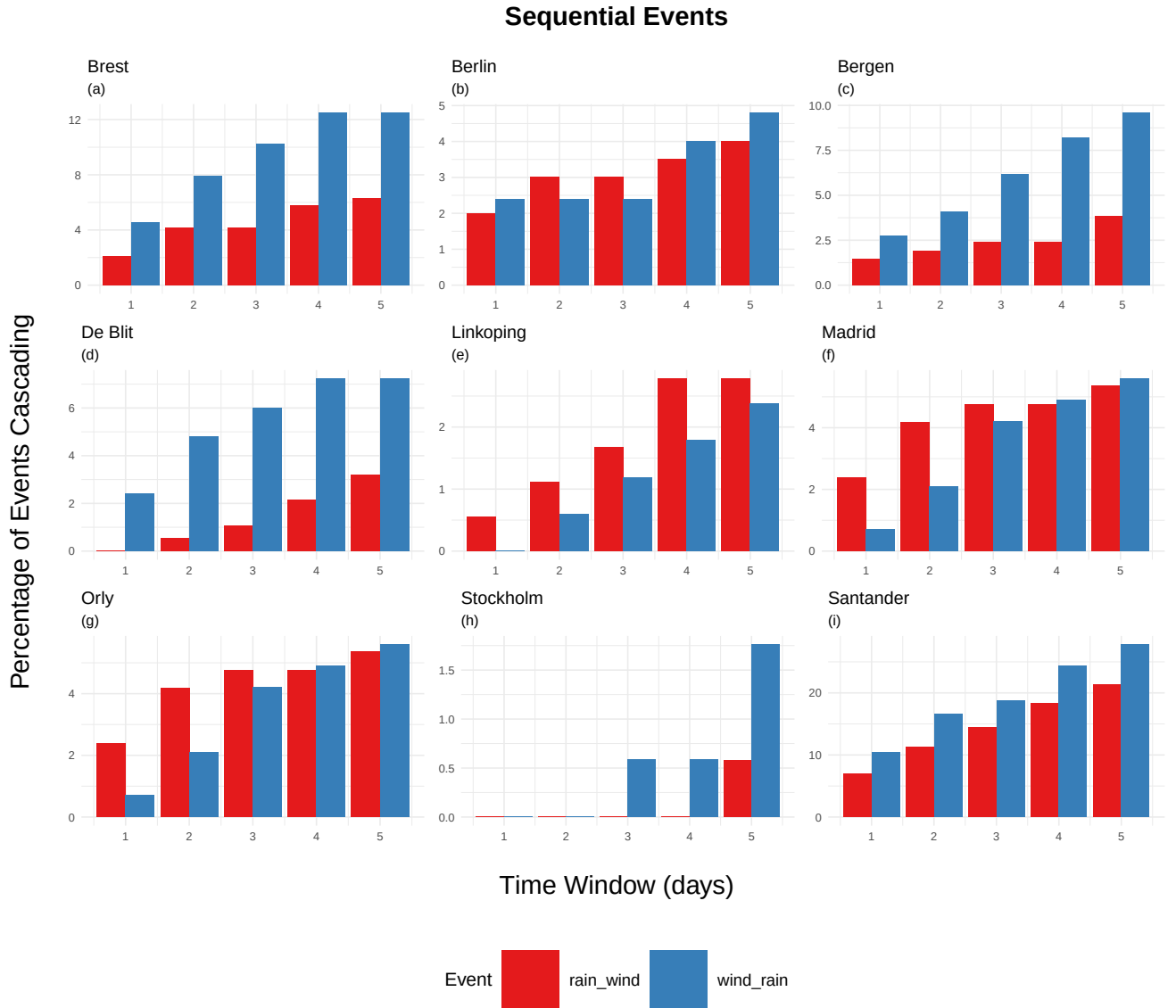


Figure 12. Percentage of extreme precipitation or wind events which are followed by the other extreme, as forecasted by the SWG at lead time $T = 10$ days. Red represents events where extreme precipitation is followed by extreme wind, and blue represents events where extreme wind is followed by extreme precipitation. We consider time windows (lags between the two extreme weather events) from 1 to 5 days. The y-ranges differ between panels.

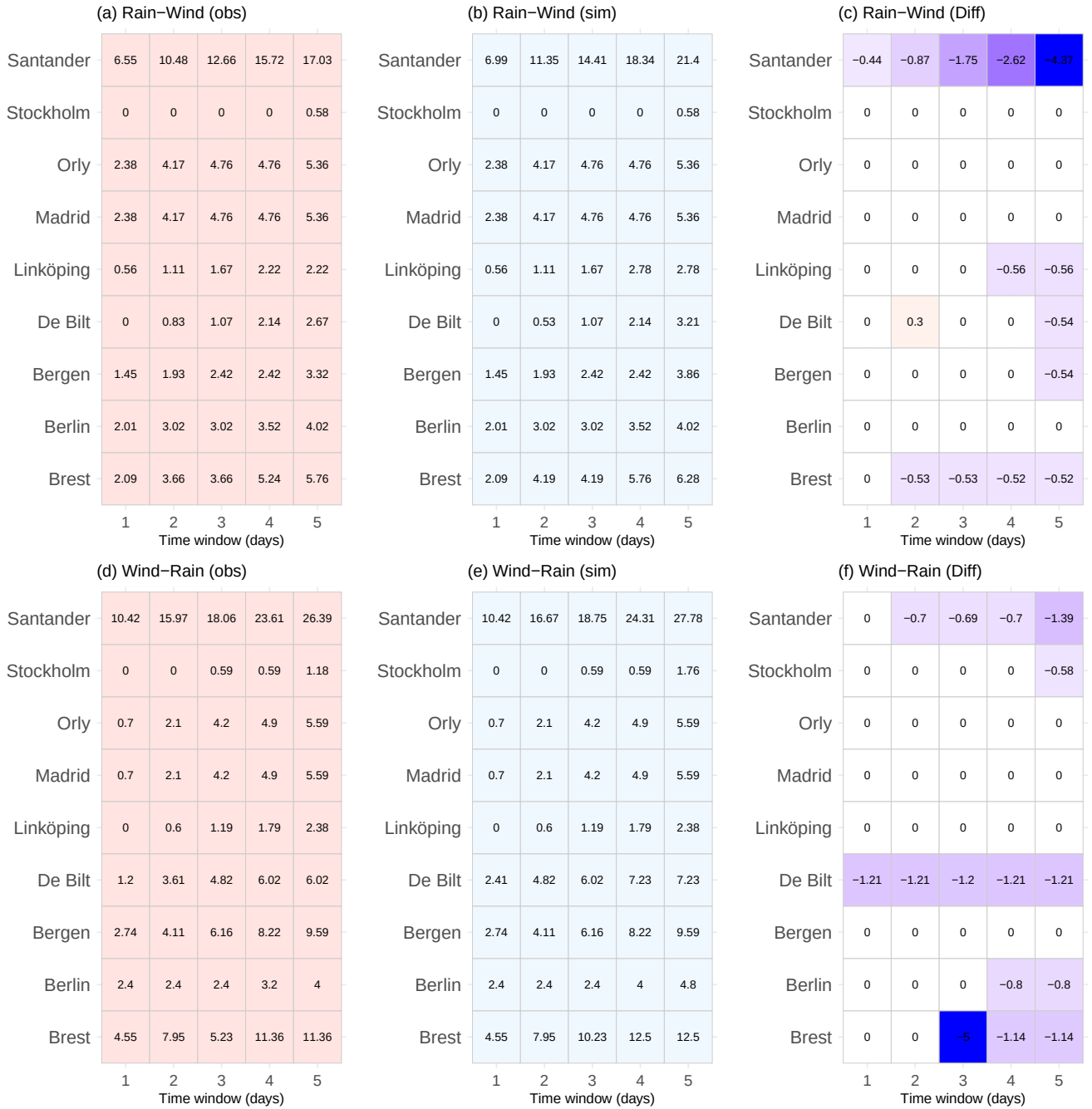


Figure 13. Comparison of the percentage of sequential extreme events as identified in observations (a,d) and simulated using the SWGs at lead time $T = 10$ days (b,e) across the studied locations for varying time windows (1 to 5 days). (a) and (d) display the observed percentages of Rain→Wind and Wind→Rain sequential events, respectively. (b) and (e) display the corresponding percentages in the SWG forecasts. (c) and (f) present the differences between observed and simulated percentages (Obs - Sim). Positive values indicate an underestimation by the SWG forecasts, while negative values indicate an overestimation.

5 Conclusions

This study presents and evaluates two ensemble forecasting approaches based on stochastic weather generators: the HC-SWG and the MA-SWG. We used the first to forecast extreme precipitation, and the second to forecast extreme wind speed across different locations in Europe. Both approaches integrate analogs of the large-scale atmospheric circulation with a stochastic weather generator to produce ensemble forecasts at medium-range lead times of up to 10 days.

The HC-SWG uses analogs from the ECMWF Z500 ensemble reforecasts. The MA-SWG uses multivariate analogs defined from the ERA5 reanalysis of Z500 and SLP. The two approaches thus differ fundamentally in their input sources. The HC-SWG indirectly benefits from flow-dependent information and information on ensemble spread as provided by the reforecasts. In contrast, the MA-SWG relies on the long historical record of the ERA5 reanalysis and on multi-variable patterns to capture large-scale circulation features.

Both SWG approaches show strong skill in forecasting extreme events up to 10 days ahead, ~~outperforming~~ improving on the ECMWF forecasts across different evaluation metrics and lead times. While potential limitations in the bias correction of ECMWF precipitation and wind speed forecasts should be acknowledged, they are unlikely to affect the qualitative improvement of SWGs over ECMWF. The HC-SWG displays high Peirce Skill Score (PSS) and Symmetric Extremal Dependence Index (SEDI) values across locations. MA-SWG shows greater spatial variability and generally positive SEDI values. The SWG forecasts also display a strong performance in reproducing the observed frequency of simultaneous and sequential extreme precipitation and wind speed extremes.

Notwithstanding their strong performance, the SWG forecasts still show limitations. Indeed, the forecast skill can be location dependent, likely due to local processes that are not well captured by the large-scale circulation analogs. Both SWG methods also overestimate the intensity of the most extreme events. For HC-SWG, this could be related to the relatively short timespan covered by the ECMWF reforecasts, limiting the availability of good circulation analogs for these rare cases. Future work could explore calibration strategies to reduce such biases and extend the SWG approach to other compound event types or geographical regions.

To conclude, we find that SWG forecasts ~~outperform~~ provide improved forecast skill compared to a set of recent numerical forecasts for extreme wind and precipitation in Europe, and correctly reproduce the frequency of compound wind and precipitation extremes. This highlights the potential of SWG forecasts for use in early-warning applications of compound hazards, which pose a key challenge for current forecasting tools.

References

- Ailliot, P., Allard, D., Monbet, V., and Naveau, P.: Stochastic weather generators: an overview of weather type models, 156, 101–113, 2015.
- 345 Alessi, M. J. and DeGaetano, A. T.: A comparison of statistical and dynamical downscaling methods for short-term weather forecasts in the US Northeast, *Meteorological Applications*, 28, e1976, 2021.
- Atencia, A. and Zawadzki, I.: Analogs on the Lorenz attractor and ensemble spread, *Monthly Weather Review*, 145, 1381–1400, 2017.
- Ben-Bouallegue, Z.: Seamless prediction of high-impact weather events: a comparison of actionable forecasts, *arXiv preprint arXiv:2312.01673*, 2023.
- 350 Ben Bouallègue, Z., Magnusson, L., Haiden, T., and Richardson, D. S.: Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events, *Quarterly Journal of the Royal Meteorological Society*, 145, 1741–1755, 2019.
- Ben Bouallègue, Z., Cooper, F., Chantry, M., Düben, P., Bechtold, P., and Sandu, I.: Statistical Modeling of 2-m Temperature and 10-m Wind Speed Forecast Errors, *Monthly Weather Review*, 151, 897 – 911, <https://doi.org/10.1175/MWR-D-22-0107.1>, 2023.
- Bevacqua, E., Maraun, D., Vourdoukas, M. I., Voukouvalas, E., Vrac, M., Mentaschi, L., and Widmann, M.: Higher probability of compound
- 355 flooding from precipitation and storm surge in Europe under anthropogenic climate change, *Science advances*, 5, eaaw5531, 2019.
- Blanchet, J., Stalla, S., and Creutin, J.-D.: Analogy of multiday sequences of atmospheric circulation favoring large rainfall accumulation over the French Alps, *Atmos. Sci. Lett.*, 19, e809, <https://doi.org/10.1002/asl.809>, 2018.
- Bouallègue, Z. B., Clare, M. C., Magnusson, L., Gascon, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Drams, J. S., Lang, S. T., et al.: The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an
- 360 operational-like context, *Bulletin of the American Meteorological Society*, 105, E864–E883, 2024.
- Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D. H., Ebert, B., Fuentes, M., Hamill, T. M., Mylne, K., et al.: The THORPEX interactive grand global ensemble, *Bulletin of the American Meteorological Society*, 91, 1059–1072, 2010.
- Brunner, M. I., Gilleland, E., and Wood, A. W.: Space–time dependence of compound hot–dry events in the United States: assessment using a multi-site multi-variable weather generator, *Earth System Dynamics*, 12, 621–634, 2021.
- 365 Buizza, R. and Leutbecher, M.: The forecast skill horizon, *Quarterly Journal of the Royal Meteorological Society*, 141, 3366–3382, 2015.
- Domeisen, D. I., White, C. J., Afargan-Gerstman, H., Muñoz, Á. G., Janiga, M. A., Vitart, F., Wulff, C. O., Antoine, S., Ardilouze, C., Batté, L., et al.: Advances in the subseasonal prediction of extreme events: relevant case studies across the globe, *Bulletin of the American Meteorological Society*, 103, E1473–E1501, 2022.
- Fang, Z., Wang, Y., Peng, L., and Hong, H.: Predicting flood susceptibility using LSTM neural networks, *Journal of Hydrology*, 594, 125–134, 370 2021.
- Fawcett, T.: An introduction to ROC analysis, *Pattern Recognition Letters*, 27, 861–874, <https://doi.org/10.1016/j.patrec.2005.10.010>, 2006.
- Ferro, C. A. T. and Stephenson, D. B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events, *Weather and Forecasting*, 26, 699–713, <https://doi.org/10.1175/WAF-D-10-05030.1>, 2011.
- Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., and Prates, F.: Evaluation of ECMWF forecasts, including the 2023 upgrade, 375 <https://doi.org/10.21957/d47ba5263c>, 2023.
- Hao, Z., Hao, F., Xia, Y., Feng, S., Sun, C., Zhang, X., Fu, Y., Hao, Y., Zhang, Y., and Meng, Y.: Compound droughts and hot extremes: Characteristics, drivers, changes, and impacts, *Earth-Science Reviews*, 235, 104–121, 2022.
- Harris, L., McRae, A. T., Chantry, M., Dueben, P. D., and Palmer, T. N.: A generative deep learning approach to stochastic downscaling of precipitation forecasts, *Journal of Advances in Modeling Earth Systems*, 14, e2022MS003120, 2022.

- 380 Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *WEATHER AND FORECASTING*, 15, 559–570, [https://doi.org/https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2), 2000.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., and Schepers, D.: The ERA5 global reanalysis, *Quat. J. Roy. Met. Soc.*, 146, 1999–2049, 2020.
- Klein Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L.,
385 Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and Petrovic, P.: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment, *International Journal of Climatology*, 22, 1441–1453, <https://doi.org/10.1002/joc.773>, 2002.
- 390 Krouma, M.: Ensemble forecasts of isolated and compound wind and precipitation extremes in Europe using a stochastic weather generator, <https://doi.org/10.5281/zenodo.16531845>, <https://doi.org/10.5281/zenodo.16531845>, 2025.
- Krouma, M., Yiou, P., Déandreis, C., and Thao, S.: Assessment of stochastic weather forecast of precipitation near European cities, based on analogs of circulation, *Geoscientific Model Development*, 15, 4941–4958, <https://doi.org/10.5194/gmd-15-4941-2022>, 2022.
- Krouma, M., Silini, R., and Yiou, P.: Ensemble forecast of an index of the Madden–Julian Oscillation using a stochastic weather generator
395 based on circulation analogs, *Earth System Dynamics*, 14, 273–290, <https://doi.org/10.5194/esd-14-273-2023>, 2023.
- Krouma, M., Specq, D., Magnusson, L., Ardilouze, C., Batté, L., and Yiou, P.: Improving subseasonal forecast of precipitation in Europe by combining a stochastic weather generator with dynamical models, *Quarterly Journal of the Royal Meteorological Society*, 2024.
- Laurila, T. K., Sinclair, V. A., and Gregow, H.: Climatology, variability, and trends in near-surface wind speeds over the North Atlantic and Europe during 1979–2018 based on ERA5, *Int. J. Climatol*, 41, 2253–2278, 2021.
- 400 Magnusson, L., Richardson, D., and Haiden, T.: Verification of Extreme Weather Events: Discrete Predictands, Tech. Rep. 731, ECMWF Technical Memoranda, <https://doi.org/10.21957/1iq131n2c>, 2014.
- Magnusson, L., Prudhomme, C., Di Giuseppe, F., Di Napoli, C., and Pappenberger, F.: Operational multiscale predictions of hazardous events, in: *Extreme Weather Forecasting*, pp. 87–129, Elsevier, 2023.
- Manzato, A.: A note on the maximum Peirce skill score, *Weather and Forecasting*, 22, 1148–1154, 2007.
- 405 Olivetti, L. and Messori, G.: Advances and prospects of deep learning for medium-range extreme weather forecasting, *Geoscientific Model Development*, 17, 2347–2358, 2024.
- Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., et al.: Weatherbench 2: A benchmark for the next generation of data-driven global weather models, *Journal of Advances in Modeling Earth Systems*, 16, e2023MS004 019, 2024.
- 410 Specq, D. and Batté, L.: Improving subseasonal precipitation forecasts through a statistical–dynamical approach : application to the southwest tropical Pacific, *Clim Dyn*, p. 15, 2020.
- Stephenson, D. B.: Use of the “odds ratio” for diagnosing forecast skill, *Weather and Forecasting*, 15, 221–232, 2000.
- Stephenson, D. B., Casati, B., Ferro, C. A. T., and Wilson, C. A.: The Extreme Dependency Score: A Non-Vanishing Measure for Forecasts of Rare Events, *Meteorological Applications*, 15, 41–50, <https://doi.org/10.1002/met.125>, 2008a.
- 415 Stephenson, D. B., Coelho, C. A., and Jolliffe, I. T.: Two extra components in the Brier score decomposition, *Weather and Forecasting*, 23, 752–757, 2008b.

- Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D., et al.: DYAMOND: the DYNAMics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains, *Progress in Earth and Planetary Science*, 6, 1–17, 2019.
- 420 Sutanto, S. J., Vitolo, C., Di Napoli, C., D’Andrea, M., and Van Lanen, H. A.: Heatwaves, droughts, and fires: Exploring compound and cascading dry hazards at the pan-European scale, *Environment international*, 134, 105 276, 2020.
- Toth, Z., Talagrand, O., Candille, G., and Zhu, Y.: Probability and ensemble forecasts, *Forecast verification: A practitioner’s guide in atmospheric science*, 137, 163, 2003.
- Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., Hendon, H.,
 425 Hodgson, J., Kang, H.-S., Kumar, A., Lin, H., Liu, G., Liu, X., Malguzzi, P., Mallas, I., Manoussakis, M., Mastrangelo, D., MacLachlan, C., McLean, P., Minami, A., Mladek, R., Nakazawa, T., Najm, S., Nie, Y., Rixen, M., Robertson, A. W., Ruti, P., Sun, C., Takaya, Y., Tolstykh, M., Venuti, F., Waliser, D., Woolnough, S., Wu, T., Won, D.-J., Xiao, H., Zaripov, R., and Zhang, L.: The Subseasonal to Seasonal (S2S) Prediction Project Database, *Bulletin of the American Meteorological Society*, 98, 163–173, <https://doi.org/10.1175/BAMS-D-16-0017.1>, 2017.
- 430 Vitart, F., Balsamo, G., Bidlot, J.-R., Lang, S., Tsonevsky, I., Richardson, D., and Alonso-Balmaseda, M.: Use of ERA5 to Initialize Ensemble Re-forecasts, 841, *ECMWF Technical Memoranda*, <https://doi.org/10.21957/w8i57wuz6>, 2019.
- Weinkle, J., Landsea, C., Collins, D., Musulin, R., Crompton, R. P., Klotzbach, P. J., and Pielke Jr, R.: Normalized hurricane damage in the continental United States 1900–2017, *Nature sustainability*, 1, 808–813, 2018.
- White, C. J., Domeisen, D. I., Acharya, N., Adefisan, E. A., Anderson, M. L., Aura, S., Balogun, A. A., Bertram, D., Bluhm, S., Brayshaw,
 435 D. J., et al.: Advances in the application and utility of subseasonal-to-seasonal predictions, *Bulletin of the American Meteorological Society*, pp. 1–57, 2022.
- Wilks, D. S.: *Statistical methods in the atmospheric sciences*, vol. 100, Academic press, 2011a.
- Wilks, D. S.: *Statistical methods in the atmospheric sciences*, vol. 100, Academic press, 2011b.
- Yiou, P.: AnaWEGE: a weather generator based on analogues of atmospheric circulation, *Geoscientific Model Development*, 7, 531–543,
 440 <https://doi.org/10.5194/gmd-7-531-2014>, 2014.
- Yiou, P. and Déandréis, C.: Stochastic ensemble climate forecast with an analogue model, *Geosci. Model Dev.*, 12, 723–734, <https://doi.org/10.5194/gmd-12-723-2019>, 2019.
- Zamo, M. and Naveau, P.: Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts, *Mathematical Geosciences*, 50, 209–234, 2018.
- 445 Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R. M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M. D., et al.: A typology of compound weather and climate events, *Nature reviews earth & environment*, 1, 333–347, 2020.

Appendix A: Forecast Bias

We evaluated the SWG forecast bias for extreme wind speed (Figure A1 a) and extreme precipitation (Figure A1 b) across the studied European locations at different lead times. The forecast bias ~~was defined as the mean difference between the simulated (sim) and observed (obs) extreme precipitation or extreme wind speed~~ corresponds to the mean relative bias, i.e. the difference between simulated and observed extremes normalized by the observed value, computed at each lead time T and for each station as follows: ~~follow~~:

$$Bias = \frac{\overline{sim} - \overline{obs}}{\overline{obs}} \quad (A1)$$

For extreme wind speed, MA-SWG forecasts show a positive bias across most locations and lead times, indicating a systematic overestimation (Figure A1 a). This overestimation is particularly notable for Santander, while it is moderate for all the other stations. In contrast, the bias for extreme precipitation is more variable, both spatially and with lead time (Figure A1 b). Several locations, including Brest, Berlin, and Bergen, show a strongly negative bias at longer lead times (notably at $T = 10$ days), suggesting underestimation of high precipitation amounts. Others, like Madrid, De Bilt and Stockholm, maintain near-zero or slightly positive biases. The MA-SWG thus appears to overpredict wind extremes, while the HC-SWG results are more variable, yet the forecasts tend to underpredict precipitation extremes as the forecast lead time increases.

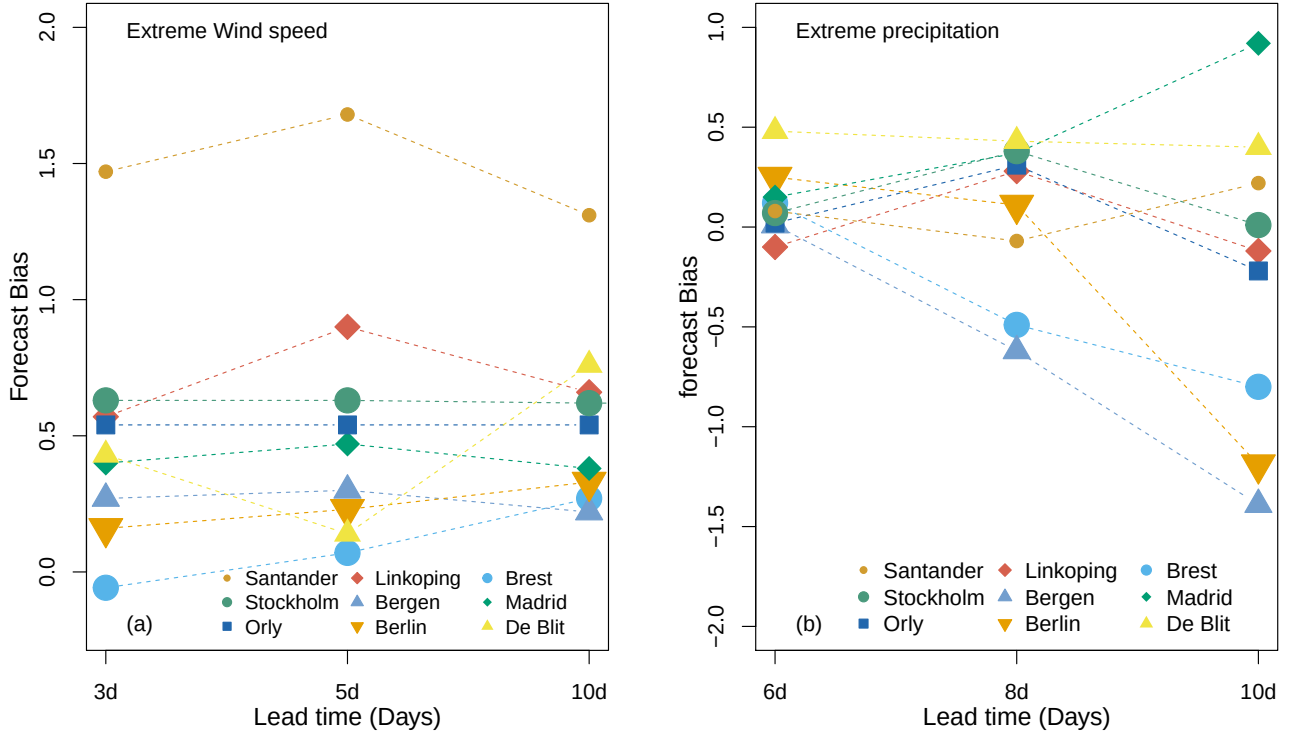


Figure A1. Bias of the MA-SWG and HC-SWG forecasts for extreme wind speed (a) and extreme precipitation (b), computed against observations for different lead times T from 3 to 10 days for all stations considered here.

Appendix B: Comparison of the SWGs forecast to the ECMWF forecast

We compare the performance of the SWGs forecast to the ECMWF forecast for all stations using D values from the Kolmogorov-Smirnov test (Table B1).

D quantifies the maximum distance between two CDFs, with higher values indicating larger distances. At all lead times and for all stations, MA-SWG and HC-SWG display lower D than ECMWF forecast (Table B1), indicating very similar distributions between SWGs forecasts and observed extremes. However, for the ECMWF forecast, D values systematically increase with lead times, either for extreme precipitation or extreme wind speed, in particular at $T = 10$ days, showing large differences between the distribution of the ECMWF forecast and the observations.

Appendix C: Additional forecast evaluation metrics

To evaluate wind speed forecasts in Appendix E, we use the Continuous Ranked Probability Skill Score (CRPSS) and temporal correlation. This mirrors the evaluation performed in Krouma et al. (2024) for precipitation forecasts. We also use the area

Table B1. Comparison of the CDFs of extreme wind speed forecasts from MA-SWG, extreme precipitation forecasts from HC-SWG, to the respective ones from the ECMWF forecast for different lead times T days with observations, using Kolmogorov-Smirnov D_{KS} values. The D values are determined between each forecast and observations.

	D_{KS} for MA-SWG forecast			D_{KS} for ECMWF forecast			D_{KS} for HC-SWG forecast			D_{KS} for ECMWF forecast		
	$T = 3$	$T = 5$	$T = 10$	$T = 3$	$T = 5$	$T = 10$	$T = 6$	$T = 8$	$T = 10$	$T = 6$	$T = 8$	$T = 10$
Bergen	0.02	0.03	0.04	0.33	0.52	0.71	0.01	0.01	0.04	0.11	0.25	0.28
Berlin	0.04	0.06	0.07	0.61	0.72	0.73	0.01	0.02	0.02	0.22	0.24	0.54
Brest	0.03	0.04	0.04	0.35	0.61	0.79	0.01	0.01	0.02	0.19	0.25	0.32
De Blit	0.05	0.05	0.08	0.78	0.86	0.87	0.01	0.01	0.03	0.3	0.23	0.36
Orly	0.03	0.05	0.09	0.25	0.36	0.37	0.01	0.05	0.07	0.14	0.19	0.35
Linköping	0.02	0.05	0.9	0.57	0.86	0.91	0.01	0.01	0.05	0.16	0.27	0.27
Madrid	0.03	0.09	0.11	0.81	0.91	0.90	0.01	0.02	0.05	0.26	0.27	0.40
Stockholm	0.05	0.07	0.07	0.61	0.62	0.91	0.01	0.01	0.02	0.26	0.57	0.47
Santander	0.02	0.04	0.05	0.53	0.59	0.69	0.01	0.01	0.02	0.19	0.23	0.23

under the Receiver Operating Characteristic (ROC) curve to evaluate the forecasts of extreme wind speed, again in line with the evaluation conducted in Krouma et al. (2024) for extreme precipitation.

To compute CRPSS we first compute the Continuous Ranked Probability Score (CRPS), which serves as a quadratic metric to measure discrepancies between the forecasted CDF and the empirical CDF derived from observed data (Wilks, 2011a; Zamo and Naveau, 2018). The CRPS is defined as:

$$CRPS(P, x_a) = \int_{-\infty}^{+\infty} (P(x) - \mathcal{H}(x - x_a))^2 dx, \quad (C1)$$

where x_a represents the observed values of x within the period $[t_0, t_0 + T]$, P is the cumulative distribution function of x from the ensemble forecast, and $\mathcal{H}$ denotes the Heaviside function, defined as $\mathcal{H}(y) = 1$ if $y \geq 0$ and $\mathcal{H}(y) = 0$ otherwise. A perfect forecast yields a CRPS value of 0. As the CRPS depends on the variable's unit, it is beneficial to normalize it relative to the CRPS of a reference forecast, such as persistence or climatology. The Continuous Ranked Probability Skill Score (CRPSS) expresses the percentage improvement over such a reference forecast (Hersbach, 2000), given by:

$$CRPSS = 1 - \frac{\overline{CRPS}}{\overline{CRPS}_{ref}}, \quad (C2)$$

Here, $\overline{CRPS}$ is the mean CRPS of the SWG forecast and $\overline{CRPS}_{ref}$ is the mean CRPS of climatology.

The Area Under the ROC Curve (AUC) quantifies the discrimination skill of a forecast, measuring how well it differentiates between event and non-event occurrences. Higher AUC indicates a superior ability to distinguish between events and non-events (Fawcett, 2006; Toth et al., 2003), with values near 0.5 representing no skill (random chance). We evaluated the AUC for two wind speed thresholds, respectively the 70th and 90th quantiles, considering events below those thresholds as non-events.

490 The Brier score (BS) evaluates the accuracy of probabilistic forecasts by computing the mean squared difference between the forecast probabilities of a given event and the observed binary outcomes (Stephenson et al., 2008b; Wilks, 2011b). It is given by:

$$BS = \frac{1}{N} \sum_{i=1}^N (f_i - o_i)^2, \quad (C3)$$

where N is the number of forecasts, f_i is the predicted probability of the event occurring, and o_i is the observed outcome (1 if the event occurred, 0 otherwise). A lower Brier Score indicates better forecast accuracy. We compute the Brier skill Score (BSS) to evaluate the sensitivity of HC-SWG forecast skill to different reforecasts δ .

The use of these different skill scores provides a comprehensive evaluation of how HC-SWG and MA-SWG predict extreme precipitation events, wind speed, and extreme wind speed events.

Appendix D: Forecast evaluation as a function of reforecast lead time for extreme precipitation with HC-SWG

500 To assess the sensitivity of the extreme precipitation forecast to the different δ lead times of Z500 analogs used as inputs to the HC-SWG (Appendix D), we used the Brier Skill Score (BSS) that we compute against the climatology. Figure D1 illustrates the sensitivity of HC-SWG forecast skill for extreme precipitation, as measured by the BSS, to both the precipitation forecast lead time (T , from 2 to 10 days) and the hindcast lead time of Z500 (δ , from 1 to 5 days), which is used as input to the SWG.

The forecasts show better probabilistic skill (higher BSS values against climatology) at lead times T closer to δ values. As T increases for the same δ , the BSS decreases, implying that forecasts initialized with older Z500 hindcasts lose accuracy, particularly at longer precipitation T lead times. The degradation in skill highlights the importance of using Z500 states that are as close as possible to the forecast initialization time to optimize the performance of the SWG. These results emphasize that both lead time and the choice of atmospheric conditions used for determining the analogs play a crucial role for the forecast skill of extreme precipitation, with large-scale atmospheric states close to the forecast initialisation date providing better predictive information compared to more distant ones.

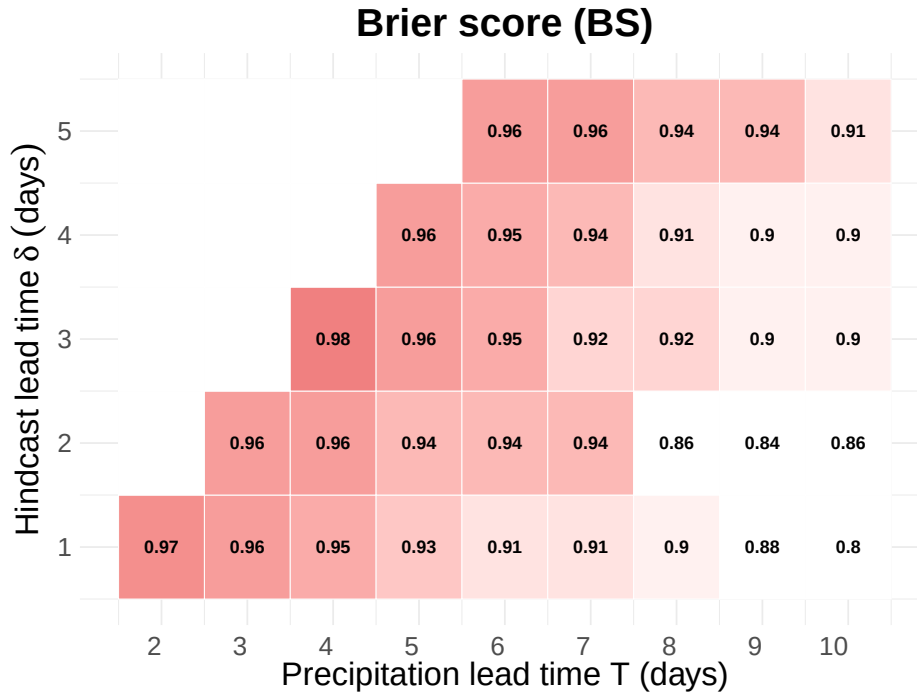


Figure D1. Brier Scores of the HC-SWG forecast performance for extreme precipitation at different forecast lead times T and for different δ lead times of the ECMWF Z500 ensemble reforecasts.

Appendix E: Forecast evaluation for wind speed and wind speed extremes with MA-SWG

We start by evaluating the forecast skill of MA-SWG for wind at 10m, as the SWG was not tested in any previous studies to forecast the wind in Europe.

Figure E1 presents the CRPSS for wind forecasts at different lead times (1, 3, 5, 10, and 20 days) relative to a climatological reference for summer (JJA, Figure E1 a) and winter (DJF, Figure E1 b). In both seasons, forecast skill decreases with increasing lead time, with the sharpest decline occurring between 1-day and 5-day lead times, followed by a more gradual reduction at longer lead times.

CRPSS values are always positive, indicating that the MA-SWG forecasts systematically provide an added value relative to climatology. Overall, CRPSS values are higher in winter than in summer, indicating that wind forecasts tend to be more reliable during DJF, whereas JJA forecasts show high variance across locations with CRPSS values varying between 0.93 and 0.58 at $T = 1$ day. The seasonal difference in forecast skill may be attributed to the fact that atmospheric conditions are more predictable in winter, and the correlation between atmospheric circulation and surface variables is stronger in that season (Laurila et al., 2021). This is particularly relevant given that we are using Z500 and SLP as predictors. Some locations, such as Bergen, Linköping and Santander, consistently show higher CRPSS values than the other stations in both seasons. Others, such as Stockholm and Berlin, systematically exhibit amongst the lowest skills. These location-specific variations may be due

to regional climate differences and local topography. Overall, wind forecasts perform better in winter, with skill declining with increasing lead time T and a pronounced location dependence in both seasons.

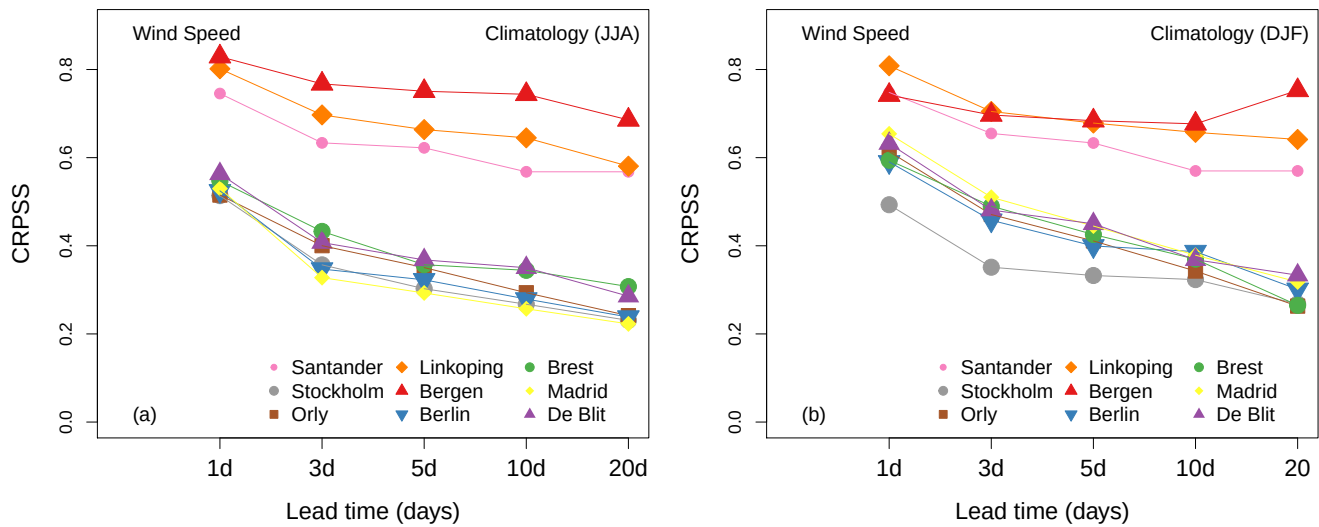


Figure E1. CRPSS with respect to climatology for the MA-SWG forecasts of wind speed. Forecasts for (a) JJA and (b) DJF for lead times of $T = 1, 3, 5, 10$ and 20 days, for all stations considered here.

Figure E2 provides information corresponding to Figure E1 but for the correlation between wind forecasts and observations. In both seasons, at short lead times (1 day), most locations show high correlation values (0.6–0.8), indicating strong predictive skill. The correlation decreases with increasing lead time, with a sharp drop between 1-day and 5-day lead times. Winter (DJF) forecasts generally exhibit higher correlations than summer (JJA) forecasts, supporting previous results in Figure E1. Indeed, summer forecasts display relatively low correlations already at 3 day lead times, likely due to increased atmospheric instability and more localised high-wind events compared to winter. Bergen and Santander tend to exhibit amongst the highest correlations in JJA, whereas Orly and Berlin emerge as displaying amongst the lowest correlations in DJF. Overall, winter wind forecasts are more consistent with observations than summer forecasts, with location-dependent variations becoming more pronounced as lead time increases.

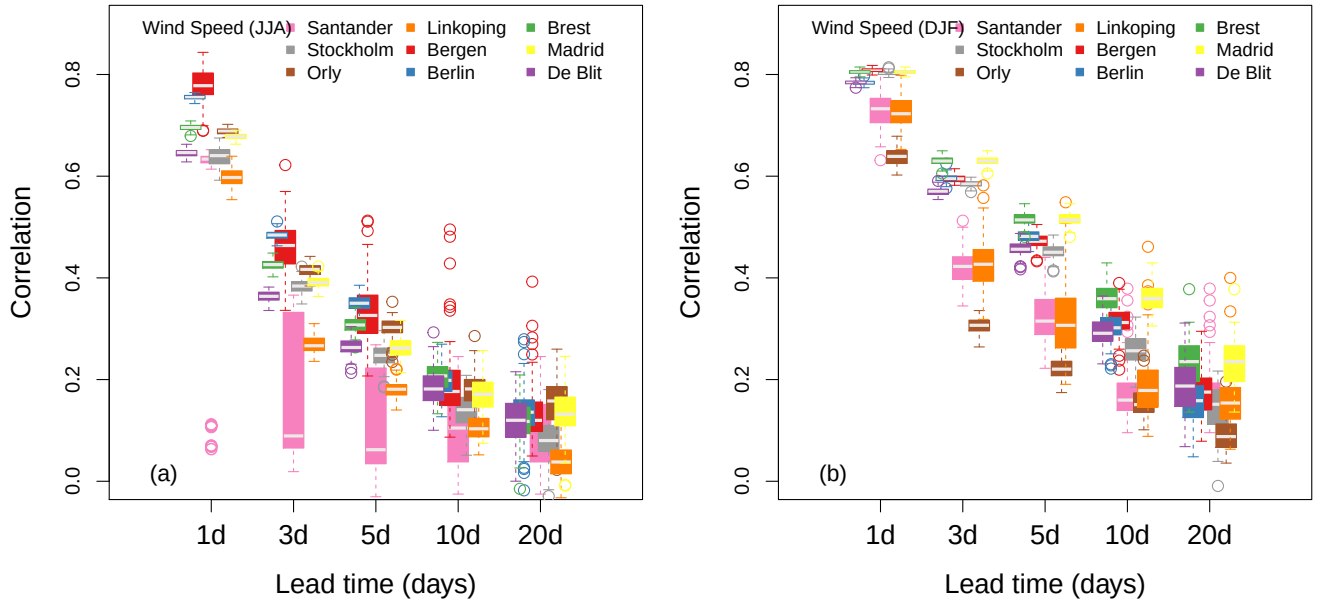


Figure E2. Rank Correlation between the MA-SWG forecast of wind speed and observations. Forecasts for (a) JJA and (b) DJF for lead times of $T = 1, 3, 5, 10$ and 20 days, for all stations considered here.

Figure E3 presents AUC values for wind speed forecasts exceeding $70Q$ (Figure E3 (a)) and $90Q$ (Figure E3 (b)) at the same lead times and stations as in the previous figures. In both cases, forecast skill decreases with increasing lead time, particularly for the $90Q$ threshold. AUC values are generally higher for the $70Q$ threshold than for the $90Q$ threshold, suggesting better predictability for more moderate wind events. Nonetheless, the spread between locations is higher for the lower wind threshold. Some locations, such as Berlin, Linköping and Orly, maintain relatively high skill levels over time, whereas others such as Bergen and Madrid show a more rapid degradation in AUC, particularly beyond 5–10 days. For the $90Q$ threshold, at longer lead times (10–20 days), several locations have AUC values close to or below 0.5, suggesting that forecasts provide little added value over random chance. Overall, wind forecasts are more skillful for lower wind speed thresholds (wind speed $> 70Q$) and shorter lead times, while forecasts for higher wind speed extremes (wind speed $> 90Q$) come with greater uncertainty.

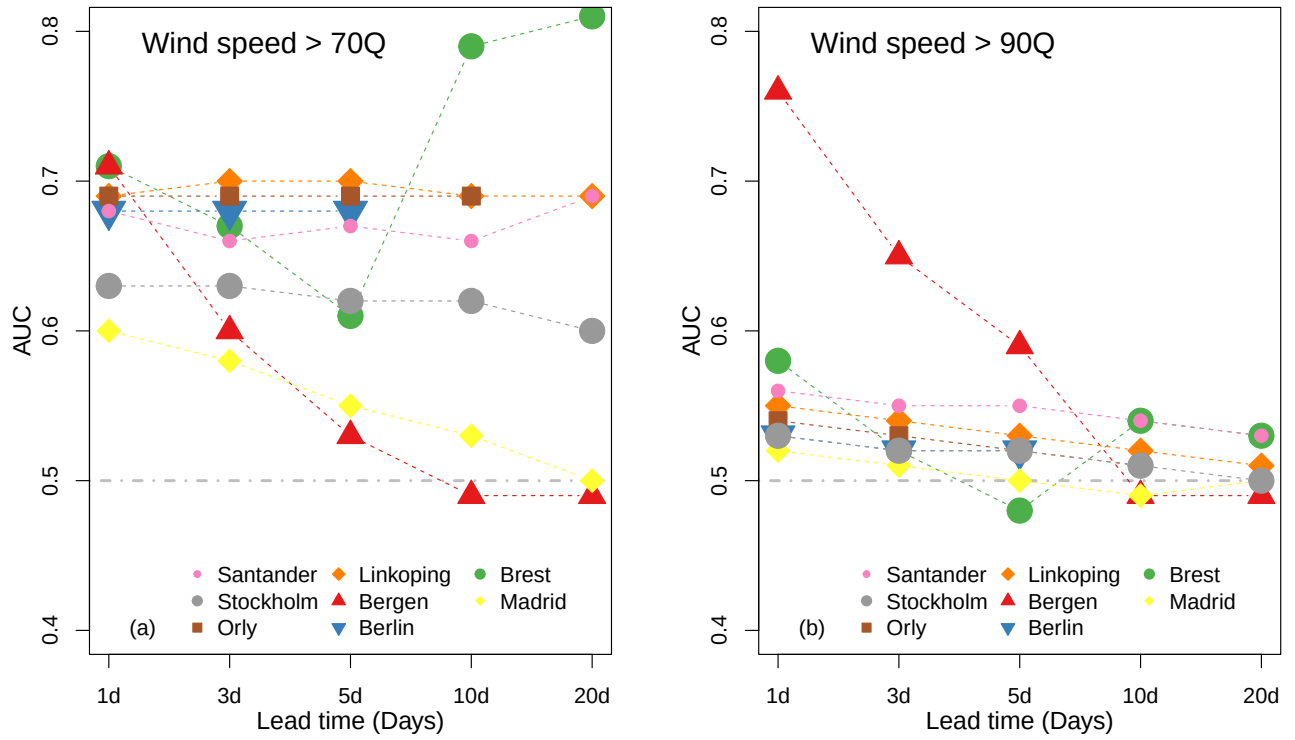


Figure E3. AUC for the MA-SWG wind speed forecasts. (a) For wind speed events exceeding the 70 Q threshold, and (b) for wind speed events exceeding the 90 Q threshold. The dashed grey lines represent $AUC = 0.5$.

Appendix F: Forecasted stations

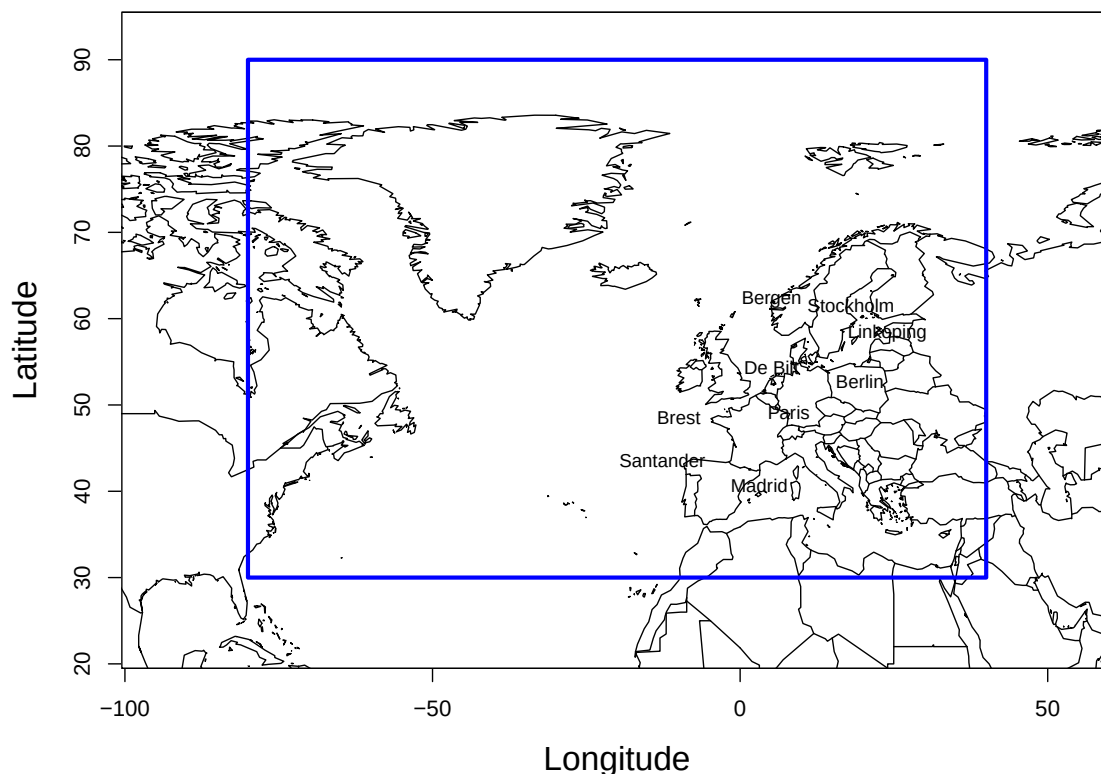


Figure F1. Map of the 9 considered stations over Europe. The blue rectangle shows the area over which we compute analogs.

Code and data availability. The code is available at <https://doi.org/10.5281/zenodo.16531845> (Krouma, 2025), together with the input data files for the SWG. The provided data files include daily precipitation for the studied stations (as an example dataset), as well as daily wind data from the ECA&D database (Klein Tank et al., 2002). We also include an example of Z500 analogs at $\delta = 5$. Z500 data can be retrieved
550 from the Copernicus Climate Data Store at the following link: <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-pressure-levels/>.

Author contributions. MK designed and performed the analyses. GM co-designed the analyses. MK wrote the first draft of the manuscript. Both authors contributed to refining and updating the manuscript.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

555 *Acknowledgements.* This study was funded by the European Union’s H2020 research and innovation programme under ERC grant no. 948309 (CENÆ project). G. Messori also acknowledges support from the Swedish Research Council Vetenskapsrådet (grant no. 2022-06599). We acknowledge ECMWF and the Copernicus Climate Change Service for granting access to the ERA5 and S2S reforecast data. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725