

We thank reviewer 1 for their feedback and comments. In this document, we provide answers for their questions.

Reviewer 1

The paper is interesting. It extends an approach developed by Krouma et al. (2024) to wind and precipitation forecasts.

Major comments

The reason why precipitation and wind speed are treated differently (with different versions of the SWG) is unclear and should be explained. This seems to add an unnecessary complication to the manuscript.

⇒ We thank the reviewer for this comment. Precipitation and wind speed are treated with different SWGs because the HC-SWG was specifically developed and validated for precipitation and then extreme precipitation, using Z500 as a predictor whose relevance has been extensively demonstrated in previous work (Krouma et al., 2022, 2024).

In contrast, extreme wind speed is driven by different physical mechanisms and requires different large-scale predictors. We tested forecasts of wind speed using one single atmospheric variable and they did not work. We now specify this in the manuscript. Extending the HC-SWG framework to jointly handle multiple predictors and ensemble reforecasts for wind would significantly increase the model complexity without clear added value. For this reason, we developed a dedicated SWG for wind speed, better tailored to its dynamics. This distinction and its rationale are clarified in the manuscript (L.54-58, p.2-3).

A greater care should be used in the figures. Figure 2 mentions "Santander" on the right column, while it should be "Linköping". The colors for Fig. 3 and 4 (among others) should be consistent across stations. Figure D1 is unreadable. What are the colors for? Figure E2 is very hard to read.

⇒ Thank you for mentioning that, we corrected the legend for Figure 2. We verified the consistency of the colors among all figures (Fig 3, 4, 8 and 9). We changed the colors for Figure D1 for better readability, and now explain that they provide a visual illustration of the value of the Brier scores.

Who is likely to use such systems, which seem to require a complex layer of simulations on top of already complex datasets of ensemble forecasts?

⇒ This is an interesting question. Such systems are primarily intended for operational forecasting centres and risk management agencies. Several of these have the technical competence needed to implement approaches such

as the ones we propose, and would then be able to provide the results to local stakeholders involved in impact-based decision making (e.g., emergency managers, infrastructure operators), who in many cases may lack the ability to implement the approaches themselves. By providing calibrated probabilistic information on extremes and compound events at specific locations, the framework can support early warning systems and local adaptation planning, complementing existing ensemble forecast products. We added this in the introduction (l. 45-48 p.2)

Specific comments

Why are the procedures described in Secs. 3.1.1 and 3.1.2 relevant to extremes?

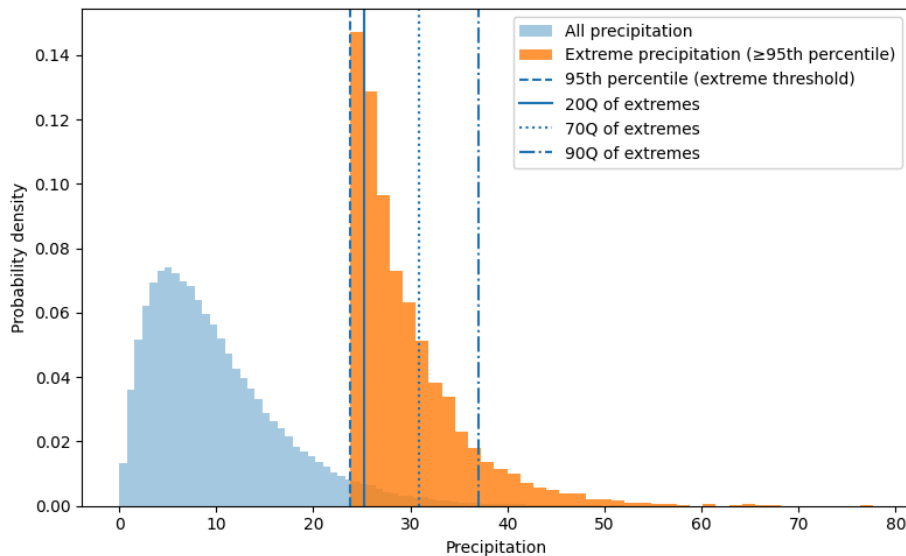
⇒ The described procedures in Secs 3.1.1 and 3.1.2 explain how each SWG works and is used to forecast each variable. For example for the HC-SWG, we use the whole ensemble reforecast of the ECMWF to find good analogs, which was specifically useful for the extreme precipitation also using different lead time of the reforecast. Similarly, for the wind extremes we explain the inclusion of different atmospheric variables in the SWG instead of one. We included these descriptions to ensure reproducibility of our analysis and to make the study self-contained.

Please provide a map of the 9 stations (the reader might not be so familiar with European geography).

⇒ A map showing the 9 stations has been added, in Appendix F.

What do 20Q, 70Q or 90Q mean? This is not stated in the text. If 20Q is the 20th quantile of precipitation, this cannot be considered as extreme.

⇒ Apologies for the confusion, but what we do is that we consider the quantiles of the extremes (and that's what 20Q, 70Q and 90Q refer to) and not the quantile of the full distribution as shown in this example below.



So the quantiles are computed conditionally on extreme precipitation, rather than from the full precipitation distribution. Extreme precipitation is first defined as values exceeding the 95th percentile of daily precipitation at each station. The resulting subset therefore represents only the upper tail of the precipitation distribution. The 20Q, 70Q, and 90Q are then calculated *within this conditional distribution of extreme precipitation* with lower quantiles corresponding to relatively moderate extremes and higher quantiles highlighting the most intense precipitation events. Importantly, these quantiles should be interpreted as percentiles of the extreme precipitation distribution and not as percentiles of all precipitation values. We explained this better in the text (l.184-190, p.8).

The "outperform" verb looks like an exaggeration. Since the SWGs are based on the forecast products, the least that can be achieved is an improved performance. The improvement for extremes is not clearly outlined.

⇒ We agree with the reviewer that the term “outperform” may be too strong in particular for the HC-SWG as it is built upon ensemble reforecast products. The HC-SWG role is to improve forecast and representation of extremes rather than to fundamentally outperform the underlying forecasts. We therefore revised the wording to “shows improvement against” or “improving forecast skill compared to” to clarify the added value of the SWGs.

We thank reviewer 2 for their feedback and comments. In this document, we address their questions.

Reviewer 2

This manuscript discusses statistical methods for the prediction of extreme precipitation and extreme wind and compound events.

The statistical methods are based on stochastic weather generator (SWG) using analogues from ERA5. The verification focuses on 9 locations in Europe.

Major comments

The verification is based on scores applied to deterministic and probabilistic forecasts.

SEDI and PSS are scores to assess deterministic forecasts. Could the author describe more thoroughly to which forecasts these scores are applied to? This is not clear from the text.

⇒ The SEDI and PSS are applied to binary exceedance forecasts derived from the SWG outputs, where events are defined relative to extreme thresholds (Figures 6 and 8). Probabilistic verification (e.g., BSS) is applied to ensemble-based forecasts. This distinction is now clarified in the manuscript (l.189-192, p. 8).

ECMWF forecasts are used as a reference for computing BSS. ECMWF forecasts are not calibrated as illustrated in Figures 6 and 9, despite that it is claimed that they are in line 82. For scores focusing on extreme events, it is crucial to compare calibrated forecasts.

⇒ The ECMWF forecasts used as a reference are bias-corrected using a linear regression-based approach applied over the full distribution of precipitation and wind speed at the nearest grid point to each station (explained now in l.194-199, p.8). This is a standard calibration approach, yet may not optimally calibrate extremes, which likely affects the comparison with the SWGs. We explicitly acknowledge this limitation in the revised manuscript (l.326-328, p.23).

Specific comments

Line 79, "9km". Could you double check the resolution of ECMWF ensemble forecasts in TIGGE over the period 2017-2021?

⇒ Thank you for spotting this mistake. The resolution for the data before 2021 is 18 Km. It is 9 km only in the new model which we do not use. This was corrected.

Line 82, "bias corrected". Could you indicate how?

⇒ The ECMWF ensemble forecasts were bias-corrected 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 corrects systematic biases in both the mean and amplitude of the forecast while preserving ensemble spread. This clarification has been added to the manuscript (l.84-90, p.3-4).

Line 85, "local". How do you build a climatology? Which period do you use? Is it local in space and time?

⇒ Local here refers to a specific station. We have 9 locations over Europe: Bergen, Berlin, Brest, De Blit, Linköping, Madrid, Orly, Santander, and Stockholm. To define the extreme values of precipitation or wind speed, we consider the 95th quantile for each time series of precipitation at each specific location, as the forecast is done separately for each station and also as extreme values might differ across locations. We now specify that the period is from 1960 to 2022 at the beginning of the Section. We don't consider the climatology here.

Line 85, "excluding values below 1mm/day". Why? Why not considering higher percentiles instead?

⇒ Daily precipitation can contain a lot of zeros as is the case of southern European locations. In order to retrieve the high percentiles corresponding to extreme precipitation we exclude values below 1mm/day. Precipitation below 1mm/day is considered as drizzle or no precipitation. This is a relatively standard threshold to use, see. e.g. Douville et al. (2023) and the papers cited therein.

Douville, H., Chadwick, R., Saint-Lu, M., & Medeiros, B. (2023). Drivers of dry day sensitivity to increased CO₂. *Geophysical Research Letters*, 50(14), e2023GL103200.

Line 110, "95th percentile of the distribution for the full forecast period". What is forecast period again? How is defined the climatology? Is it a forecast climatology?

⇒ The forecast period is from 2002 to 2021, which we now specify in the text (l.120, p.4). The forecast is not climatology, and we don't use climatology here at all.

Line 121, "we use the Z500 field to estimate daily extreme precipitation". I don't understand this step. Could you please elaborate?

⇒ Z500 is used as a large-scale predictor to identify circulation patterns associated with extreme precipitation. The 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 now explain this in the text on l.131-133 p.5.

Line 187, "using $\delta = 5$ ". Could you recall what δ stands for?

⇒ δ stands for the lead time of the Z500 ensemble reforecast. This was added (l.203-204, p.9).

Figure 2, $P > 95\text{th } Q$. What does P refer to? Observation, or forecast, or both?

⇒ P refers to precipitation, and we use it to indicate both observed and forecasted values. This was added to the figure.

Figure 4, "forecasted" to be replaced by "forecast".

⇒ This was corrected.

Line 428, "forecast bias". The definition in the text and Equation A1 do not match.

⇒ Thank you, we corrected the definition to match the equation. Here is the correct definition: "the forecast bias was defined as the mean relative bias, computed as the difference between simulated and observed extremes normalized by the observed value".