Selective ensemble mean technique for severe European windstorms

Sebastian Scher | Gabriele Messori

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

Correspondence
Sebastian Scher, Department of Meteorology and Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden.
Email: sebastian.scher@misu.su.se

We show that ensemble forecasts of extreme European windstorms can be improved up to a lead time of 36–48 hr by sub-selecting ensemble members based on their performance at very short lead times (up to 12 hr). This applies to both the ensemble mean position of the cyclone centre and the ensemble windstorm footprint over the continent. A number of ensemble forecasts, including those from the ECMWF Ensemble Prediction System, are initialised every 12 hr and disseminated several hours after initialisation; therefore our approach has the potential to provide improved forecasts in an operational context. The analysis is performed on GEFS reforecast data.

KEYWORDS
ensemble forecasts, forecasting (methods), mid-latitudes, severe weather

1 | INTRODUCTION

Windstorms are a leading source of insured losses in Europe (Schwierz et al., 2010) and frequently lead to human casualties (Kahn, 2005). Correctly forecasting these events is therefore of crucial importance. In operational practice, ensemble forecast systems are routinely used for this task.

The skill of ensemble forecasts has been steadily improving since their inception, and recent ensemble systems can successfully predict the track and footprint of severe European storms up to 2–4 days ahead (Pantillon et al., 2017). The full ensemble provides probabilistic forecast information, but this is not suitable for all applications and a number of end users rely on the ensemble mean. There are many known methods for improving forecasts with statistical post-processing (e.g., Glahn and Lowry, 1972; Gneiting et al., 2005). These methods all require historical datasets for calibration. Recently, other types of post-processing have been proposed that do not need calibration data. Du and Zhou (2011) showed that weighted means of ensemble members can outperform the standard ensemble mean, especially in the short range (∼1 day). Qi et al. (2014) later applied a similar concept to forecasts of tropical hurricanes and developed a selective ensemble mean technique. They argue that because there is a time-lag of several hours between the initialisation time of the ensemble forecasts and the actual delivery time of the forecast data, at dissemination time new observations are already available. One can therefore compare the first hours of the ensemble forecast to reality and retain only the best-performing ensemble members. In operational practice, the “real” position of a cyclone might, for example, be inferred from measurements of surface pressure by ground stations or from satellite images, both of which have much shorter latency than ensemble forecasts. If the sub-selection is performed on sufficiently short lead times – ideally very close to the dissemination time of the forecast – this would provide an operationally valuable forecast improvement. If, for example, the dissemination time is 8 hr after initialisation and the sub-selection is done at a lead time of 8 hr, the method can be applied immediately after the forecasts have become available.

The method of Qi et al. (2014) was further investigated in Du et al. (2016), who also considered time-lagged ensembles consisting of 2–3 consecutive ensemble forecast initialisations. Switching from a single ensemble initialisation to the time-lagged ensemble resulted in a small improvement in the forecasts of the full ensemble mean, but worsened the mean of the sub-selection. The mean of the members selected from the most recent ensemble initialisation provided the best forecast.

Here, we adapt the approach developed in Qi et al. (2014) and test it on a set of extreme European windstorms. We specifically verify whether such an approach could in theory...
be applied in an operational context. We base our analysis on a homogeneous ensemble reforecast dataset – specifically the National Oceanic and Atmospheric Administration’s (NOAA) Global Ensemble Forecast System (GEFS) reforecasts (Hamill et al., 2013).

2 | DATA

2.1 | Storm database

We use the storms listed in the Extreme Wind Storms (XWS) Catalogue provided by Roberts et al. (2014). The XWS Catalogue contains 50 extreme European windstorms over the period 1979–2012. The storms are selected because they either caused large insured losses or were of similar intensity to those that did (but, for example, did not affect any large urban areas, thus limiting damage to property). The same database has recently been used by Pantillon et al. (2017). Since the reforecast data used here are available only from December 1984 onwards (see below), our study is limited to the 44 storms in the 1985–2012 period.

2.2 | Forecast and validation data

Operational weather forecast models are updated regularly (roughly every 6 months in the case of the European Centre for Medium-Range Weather Forecasts [ECMWF] model). This makes analyses of historical storms difficult, as conclusions drawn from old versions of models are not necessarily valid for the current version. Since our study focuses on extremes, obtaining a sufficiently large sample of storms requires selecting back over a long period – much longer, in fact, than the typical update cycle of the models. Reforecasts (sometimes also called hindcasts) mitigate this problem. They provide forecasts initialised from historical data using the same model configuration as in current forecasts (or at least a relatively recent version). Even though the forecast model and the assimilation systems are consistent across the dataset, the quality and amount of available observations can change over time. Nevertheless, reforecasts provide the most consistent set of ensemble forecasts available.

Two commonly used reforecast datasets are the NOAA’s GEFS reforecast system version 2 (Hamill et al., 2013) and the ECMWF hindcasts (Hagedorn et al., 2008; 2012). The ECMWF hindcasts have the advantage of a higher spatial resolution than the GEFS reforecasts (O640 compared to T254). However, they are only initialised twice a week, while the GEFS reforecasts are initialised daily. Therefore, the use of GEFS reforecasts allows for a more comprehensive analysis of each storm.

The GEFS reforecast system version 2 uses the 2012 version of the GEFS ensemble prediction system, and is run with 10 members and one control forecast at a horizontal resolution of T254 for the first 7 forecast days. Forecasts are initialised daily at 0000 UTC and are available from December 1984 to the present. We use 1° × 1°, 6-hourly surface pressure and surface wind speed data. The temporal resolution is chosen to match that of the ECMWF’s ERA-Interim reanalysis (Dee et al., 2011), which we use as a tool to verify “reality”, or validation data.

ERA-Interim provides 6-hourly data from 1979 up to 2 months from the present. Here we use data with a horizontal resolution of 0.75° × 0.75°.

3 | WINDSTORM ANALYSIS

3.1 | Storm tracking

To identify the storms of interest both in ERA-Interim and in the reforecasts, the storm tracking algorithm from Pinto et al. (2005) was used. This algorithm is based on the algorithm of Murray and Simmonds (1991), and its improved version in Simmonds et al. (1999). In short, the algorithm works as follows. Cyclones are identified as surface pressure minima or minima in surface pressure gradients close to maxima in the Laplacian of surface pressure. Sequential cyclone positions are then connected up to form tracks by using a maximum likelihood method. Merging and splitting is not allowed in the scheme. For more details, the reader is referred to Murray and Simmonds (1991), Simmonds et al. (1999) and Pinto et al. (2005). Tracking was done on the ERA-Interim 0.75° grid, with GEFS reforecasts being regridded from 1° × 1° with bicubic interpolation (as suggested in Pinto et al., 2005). Storm tracking always maintains a certain degree of subjectiveness, and different tracking algorithms yield different results. It is often difficult to decide which approach will be the best (Raible et al., 2008; Neu et al., 2012). However, using the same algorithm for both the forecasts and the validation data at least avoids errors due to discrepancies between tracking methodologies. We note that the same approach was used in Pantillon et al. (2017).

3.2 | Identification of storms in reforecasts

For each storm in the XWS Catalogue, we first identify the corresponding storm in ERA-Interim. The challenge is then to select the tracks that correspond to the different realisations of the historical XWS storm in each GEFS reforecast. We begin by searching for the $t = 0$ hr track (i.e., the storm position at the model initialisation time) in each ensemble member which is closest to the track position in ERA-Interim. If no track is found within $3^\circ$ of latitude and longitude of the ERA-Interim track, the ensemble member is discarded. If suitable tracks are found in less than 6 (out of 10) members, the whole forecast is discarded and not considered further. This happens in around 50% of the cases, where by “case” we refer to one initialisation time for a given storm (typically 2 cases per storm pass the pre-selection). The above approach is different from those adopted in earlier studies. Pantillon et al. (2017) considered the distance of the forecast track to the real track at $t = 24$ hr,
while the more rigorous approach of Froude et al. (2007) used the distance along the whole track. However, in this study we aim to present an approach which could be applied in an operational context. We therefore opt to use \( t = 0 \) hr data for the initial storm identification and \( t = 12 \) hr data for the ensemble sub-selection. Further, we only consider initialisation times at which the historical storm was strong (relative vorticity \( > 4 \times 10^{-5} \text{ s}^{-1} \)), as given by Leckebusch et al. (2008). This is similar to Qi et al. (2014) and Du et al. (2016), who only used initialisation times where the storm had at least tropical storm strength. A final constraint is that we can only consider initialisation times at 0000 UTC (the initialisation time of the GEFS reforecasts).

Note that the “pre-selection” described above precedes the performance-based member selection. If this initial step is not performed, the full ensemble could potentially include too many storms that are unrealistic from the very start and thus be excessively penalised relative to the sub-selected ensemble. If one particular ensemble member does not have any track close to the real track, simply searching for the closest track might yield a track associated with a different weather system. Therefore, pre-selection is necessary. The group of members that pass the pre-selection process will hereafter be referred to as the “full ensemble”, which can consist of 6–10 members. In Figure 1 we illustrate two cases, one where the identification of the historical storm in the forecast is straightforward (all members retained during pre-selection; see Figure 1a), and one where a member is discarded in the pre-selection procedure (grey dotted line; see Figure 1c). We note that the difficulties in identifying storms in forecasts are unavoidable, and would also occur in an operational context.

### 3.3 Storm Severity Index (SSI) footprint

From an impacts point of view, correctly forecasting the footprint of the destructive surface winds associated with a cyclone is more valuable than forecasting the exact cyclone track, although the two are obviously correlated. To quantify the destructiveness associated with a given cyclone, we use an adapted version of the Storm Severity Index (SSI) developed by Leckebusch et al. (2008). The SSI is based on wind power above a threshold \( v_{\text{ref}} \):

\[
\text{SSI} = \sum_{ij} \max \left(0, \frac{v_{ij}}{v_{\text{ref},ij}} - 1 \right)^3,
\]

where \( v_{ij} \) is the 10 m wind speed at gridpoint \((i,j)\) and \( v_{\text{ref},ij} \) is the 98th percentile of the full-year climatology of the wind speed at point \((i,j)\). The rationale behind the definition of the SSI is that individuals and infrastructure are adapted to local wind conditions. Widespread damage or large numbers of casualties will generally occur when the wind is beyond a “normal” range for the region (Klawa and Ulbrich, 2003).

In their SSI, Leckebusch et al. (2008) then summed the index over time. We, however, are interested in forecast performance at individual lead times. We therefore consider SSI values at individual times. To avoid including the footprints of nearby storms in our calculation, we limit the SSI to a circle of 5° around the storm center. For the purpose of evaluating the forecasts, the storm center is taken to be that of ERA-Interim. The percentile \( v_{\text{ref}} \) was computed separately for ERA-Interim and for the reforecasts, but over the same period (1985–2016).

### 3.4 Member selection

Our selection method is based on that of Qi et al. (2014) and Du et al. (2016). Qi et al. (2014) proposed different methods, of which we use the most successful, called SEAV by the authors and POS-SEL (position selection) here. The method compares the tracks of a given storm in all ensemble members (that passed pre-selection) to the verification track at a lead time of \( t = 12 \) hr. For each ensemble member, the error in the track position is computed and only those members with an error below the average error of all members are selected. These are then used to compute a new ensemble mean for the later lead times, as depicted in Figure 2. A real case is shown in Figure 1. The left panels show the full ensemble and the corresponding ensemble mean for Storm Vivian at two different stages of the storm. The right panels show the members that were selected for the sub-ensemble and the corresponding new ensemble mean.

In their study, Qi et al. (2014) only considered performance in forecasting the location of the cyclone centre at \( t = 12 \) hr. Here, we extend this approach, and use a second selection method based not on the quality of the cyclone centre forecast but on the quality of the forecast of the cyclone central pressure and thus on the intensity of the cyclone. We term this PMIN-SEL (pressure minimum selection). As for the selection based on track forecasts, we only retain those members which have below-average error.

In addition to our two main methods, we test two other selection methods. The first method selects based on the quality of the SSI forecasts at \( t = 12 \) hr. We define the error in the SSI as the Euclidean distance between the SSI in ERA-Interim and the SSI of each ensemble member. We then define the mean SSI error of the ensemble as the mean of the errors of the individual ensemble members. The second method selects based on the correctness of the storm translation vector in the first 12 hr of the forecast. These additional methods will be called SSI-SEL (SSI selection) and VEC-SEL (translation vector selection), respectively. SSI-SEL is sketched in Figure S1 of the Supporting Information.

### 4 Analysis of results

We first present the results obtained using the track-based sub-selection method (POS-SEL). Figure 3a shows the mean position error at different lead times for both the full ensemble (blue bars) and the sub-selected ensemble (POS-SEL;
FIGURE 1  Examples of initial selection and subsequent performance selection of storms in the ensemble forecasts for Storm Vivian at two different stages during its lifetime. The upper panels show a stage at which suitable tracks are found in all members, and no pre-selection is necessary. The lower panels show a stage at which the initial selection is more difficult. One member is thus discarded in the pre-selection (dotted grey line). The left panels show the individual ensemble members of the full ensemble plus the mean of the full ensemble. The right panels show the individual members of the sub-selected ensemble and the mean of the sub-selected ensemble. In all panels, the track from ERA-Interim is also shown. The grey dots indicate the positions of the individual members at \( t = 12 \) hr and \( t = 24 \) hr. The black circles indicate the ERA-Interim track at \( t = 12 \) hr and \( t = 24 \) hr [Colour figure can be viewed at wileyonlinelibrary.com]
There is little or no correlation between the forecast error for short lead times and the performance for long lead times. This indicates that there is no improvement in the track position quality in the improving cases. In some cases, the method makes the forecasts worse. For the sub-selection at shorter lead times, it is noted that while the sub-selection at shorter lead times reduces the average error, it does not make all forecasts better. In some cases, the method makes the forecasts worse. For a lead time of 18 hr, around 70% of the forecasts improve, while the remaining 30% become worse. At lead times of 24 and 30 hr, around 55% and 65% improve, respectively. From 36 hr onwards, more forecasts decrease than increase in quality (Figure 4). From 36 to 48 hr, however, the typical increase in forecast position quality in the improving cases is larger than the typical decrease in the worsened cases, therefore leading to a mean forecast improvement when averaging over all cases.

To summarise, our method on average increases the performance of the forecasts for short lead times, but decreases the performance for long lead times. This indicates that there is little or no correlation between the forecast error for short and long (>36 hr) lead times, even when following the storm. Thus the forecasts at longer lead times do not benefit from the additional information on quality at 12 hr, while at the same time they suffer from the smaller ensemble size. This is because, in general, a larger ensemble should lead to a better ensemble mean. Another important factor explaining why the sub-selection only leads to improvements up to 36 hr might be the predictability limit of windstorms in the GEFS reforecasts. Figure 3a) suggests that the error begins to saturate at around 36–48 hr. Thus – on average – at this lead time the ensemble members have spread out over the full spectrum of climatologically possible tracks, and sub-sampling cannot lead to any further improvement.

We next assess the impact of the POS-SEL approach on the performance of the model in predicting the SSI. The mean error in the SSI of the POS-SEL sub-ensemble does not show any significant differences in the mean error of the full ensemble (Figure 3b,d). Thus, selection based on track performance at t = 12 hr significantly improves the forecasts of storm position but does not improve forecasts of the SSI. This indicates that better forecasts of a storm’s position do not necessarily imply better forecasts of its destructiveness.

We note that the SSI errors display a diurnal cycle, with peaks at 36 and 60 hr lead times. These are at 1200 UTC (since all forecasts were initialised at 0000 UTC) and correspond to the early afternoon, when the strongest winds of the day occur due to daytime heating and may lead to larger forecast errors. This largely masks the increase in error with lead time.

We proceed to our second method, namely selection based on quality of the storm intensity at a lead time of 12 hr, approximated by the central pressure of the cyclones. In this case, the sub-ensemble leads to significantly better SSI forecasts at lead times of up to 36–48 hr (Figure 5). The improvement of SSI forecasts over land points only is less than for all points (Supporting Information, Figure S2). This might be due to the higher wind speeds over the sea, where an exceedance of the threshold in the SSI computation would lead to higher values because of the third power of the wind speed. The position forecasts also improve, albeit not as much as when selecting on position performance.

Selecting directly on SSI performance (SSI-SEL) leads to better forecasts of the SSI, but to little or no improvement in the quality of the track forecasts, depending on lead time (Supporting Information, Figure S3). We find that the method’s performance in track position is essentially in line with the full ensemble at t = 12 hr and t = 24 hr and leads to a small but significant improvement in the track position at t = 36 hr, but rapidly degrades thereafter and is significantly worse than the full ensemble mean at longer lead times. The improvement in SSI forecasts is comparable to the PMIN-SEL method. Therefore, the SSI-SEL method is clearly outperformed by the PMIN-SEL method. The same holds for selection based on the storm’s propagation vector (VEC-SEL) (Supporting Information, Figure S4). The ineffectiveness of VEC-SEL might
Finally, we tested our methods from a decision-maker's viewpoint by computing the equitable threat score (ETS), a widely used verification metric (e.g., Wilks, 2006; see Appendix S1 in the Supporting Information for a detailed description) for wind speed forecasts from the full ensemble as well as for POS-SEL and PMIN-SEL. We used both a gridpoint-dependent threshold, namely the 98th percentile of the local climatology, analogous to the computation of the SSI, and a fixed threshold of 18 m/s, which is the threshold at which the German weather service (DWD) issues warning level 2 (DWD, 2018). The result for the 98th percentile threshold is shown in Figure 6. It can be seen that, for most lead times, both sub-selecting methods lead to small but significant improvements in the ETS. For some lead times, there are no significant differences. For the fixed threshold, absolute values of the ETS were lower, but the relative differences between the methods were similar (not shown). One key point...
382

SCHER AND MESSORI

FIGURE 5  Mean error in position (a) and SSI (b) of the full ensemble (blue) and of the members selected using the PMIN-SEL method (orange). The numbers in parentheses indicate the number of cases analysed. (c) and (d) Differences between the error of the full ensemble and that of PMIN-SEL. Error bars show the 15–85 (black) and 5–95 (grey) percentiles estimated via bootstrapping. Negative values indicate better performance of PMIN-SEL. SSI is unitless [Colour figure can be viewed at wileyonlinelibrary.com]

to note in the ETS analyses is the drop in ETS below zero beyond 36 hr, which strengthens the findings from above that the limit of predictability of storms in the GEFS reforecasts is around 36 hr.

5  |  DISCUSSION AND CONCLUSIONS

We investigated the possibility of using a selective ensemble mean method to improve forecasts of extreme European windstorms. This was inspired by a similar approach developed for tropical hurricanes by Qi et al. (2014). The sub-selection was based on the error of the individual ensemble members in forecasting the position of the cyclone, or the cyclone’s central pressure, at a lead time of 12 hr. Applying this approach to GEFS reforecasts led to significant forecast improvements at short range (up to 36–48 hr).

In detail, the sub-selection based on position only led to improvements in the forecasted position of the cyclone centres for lead times of up to 36 hr, but did not increase the quality of the wind footprint forecasts. The selection based on central pressure led to an improvement in the forecasts of wind footprints as well as some improvement in the forecasts of the tracks, albeit smaller than when selecting based on position. The fact that selecting on central pressure also increases the skill in storm track position makes physical sense, since storm intensity is dynamically linked to the structure of the surrounding atmosphere, which exerts a strong influence on storm movement. We note that the improvement in the forecast of the storm tracks is more short-lived than that found by Qi et al. (2014) for tropical cyclones, where the technique increases performance up to very long lead times (120 hr). One reason for this may be that the typical propagation speed of a tropical cyclone is slower than that of a midlatitude cyclone. The average propagation speed of the
cyclones in our study is ∼50 km/h (not shown), whereas tropical cyclones usually move at ∼17 km/h at 20°N and ∼27 km/h at 35°N (NOAA, 2014). Thus, during the same forecast time window, a midlatitude cyclone moves much farther, which could explain why the lead times up to which the sub-selection improves the forecasts is shorter. Qi et al. (2014) did not assess any measure other than cyclone centre position, and we therefore cannot compare our results for SSI forecasts to those for tropical hurricanes.

We saw that the largest increase in forecast quality of the cyclone track is obtained when directly selecting on position error. However, selecting on cyclone centre pressure yields a larger increase in quality of the SSI, and also some increase in the quality of the cyclone centre position. Thus, deciding which method is best depends on the intended application. In most situations, PMIN-SEL would be the preferred method as it improves both the track and the SSI performance. However, if for some specific application a highly accurate forecast of the centre of the cyclone were needed, then the POS-SEL method would yield the best results. In terms of the ETS, the only significant difference between the two approaches before the ETS score falls to near-zero favours the POS-SEL approach.

While our sub-selection cannot beat the performance of the ensemble forecast initialised after the one we consider (not shown), it can still provide an improved forecast in the intervening period between the dissemination of successive initialisations. ECMWF forecasts – often recognised as being the world’s best medium-range forecasts (e.g., Hagedorn et al., 2012) – are initialised every 12 hr, and a typical dissemination time might be 8 hr. Our sub-selection can be performed 12 hr after the initialisation of the forecast (or even 8 hr after depending on data availability), and is therefore available well before the following forecast is disseminated (see Figure S5 in the Supporting Information). In the cases where our sub-selection is not possible because there are too few sub-selected storms, the full ensemble mean forecast can be used. We should also note that, since our sub-selection is done at a lead time of 12 hr, the effective forecast improvement time is 12 hr less than that shown in the figures (i.e., up to 24–36 hr
from the time of sub-selection). While this is a relatively short time-scale to warn for an extreme storm, many types of warnings are frequently disseminated at even shorter notice. For example, hail or heavy precipitation warnings in nowcasting systems are often issued several hours to 15 min before the event (e.g., Forster et al., 2012). Therefore, storm warnings at a 24 hr lead time can have potential value to end users.

We have also found that the lead time limit of usability of our method (36–48 hr) seems to coincide with the saturation of the storm position error of the GEFS reforecasts for the storms analysed in this study. This is not surprising, as a post-processing technique (such as our method) cannot add additional information to the forecasts, but can only help to better exploit the available information. Therefore, our method only improves predictability in the cases where there is some skill in the first place. However, the limitation by the predictability of the underlying forecast system suggests that in a more robust ensemble which has a longer predictability horizon – for example, ECMWF’s ENS, which has repeatedly been shown to outperform the GEFS ensemble (Buizza et al., 2005; Froude, 2010; Hagedorn et al., 2012; Korfe and Colle, 2018) – the proposed method might also lead to forecast improvements at longer lead times.

In our study, we use gridded reanalysis data as the ground truth, which would not be not available in real time. In an operational context, surface pressure data from ground stations or from satellite images could be used instead, but would present some challenges. Surface observations have high quality but are sparse, and would thus have to be interpolated. Satellite images are not sparse, but cannot measure pressure directly. However, storm position could, for example, be determined from cloud patterns. We plan to test both the surface observation approach and the satellite approach in a future study.

Our results suggest that selective ensemble mean techniques can be fruitfully applied in the midlatitudes, and encourage further research on the topic. While the full ensemble contains the highest degree of forecast information, a number of end users rely on the ensemble mean as a forecast. An improvement of the ensemble mean via sub-selecting members therefore provides an important step forward. Future research should include the use of different models (for example, ECMWF’s current operational forecast model) and different meteorological parameters and extreme event types. Furthermore, it would be interesting to compare our method with the already widely used ensemble calibration methods, and explore potential synergies between the different approaches. Finally, future research could investigate whether performance-based selection methods are also suitable for improving probabilistic forecasts. The fact that we successfully adapted a method developed for tropical cyclone prediction to a midlatitude context also suggests that a closer collaboration between the respective scientific communities might be valuable.

ACKNOWLEDGEMENTS

We would like to thank Erland Källén for his valuable input in discussions. We also thank Robert Mureau and Rodrigo Caballero for their helpful input in the planning phase of the research. Further, we would like to thank two anonymous reviewers for their extensive and insightful comments which helped to improve our study. S. Scher and G. Messori were funded by the Department of Meteorology of Stockholm University and by Vetenskapsrådet grant no. 2016-03724. The authors declare that they have no conflicts of interest.

DATA

XWS data: ©Met Office, University of Reading and University of Exeter. Licensed under Creative Commons CC BY 4.0 International Licence: http://creativecommons.org/licenses/by/4.0/deed.en_GB. The data can be accessed at http://www.europeanwindstorms.org/. GEFS reforecast data can be accessed at ftp://ftp.cdc.noaa.gov/Projects/Reforecast2. ERA-Interim data can be accessed at http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/.

AUTHOR CONTRIBUTION

S. Scher had the initial idea for the study and analysed the data. Both authors designed the study, interpreted the results and drafted the manuscript.

ORCID

Sebastian Scher  https://orcid.org/0000-0002-6314-8833
Gabriele Messori  https://orcid.org/0000-0002-2032-5211

REFERENCES

Buizza, R., Houtekamer, P.L., Pellerin, G., Toth, Z., Zhu, Y. and Wei, M. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. Monthly Weather Review, 133(5), 1076–1097. https://doi.org/10.1175/MWR2905.1.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Holm, E.V., Isaksen, L., Källberg, P., Kohler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N. and Vitart, F. (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society, 137(656), 553–597. https://doi.org/10.1002/qj.828.

Du, J. and Zhou, B. (2011) A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. Monthly Weather Review, 139(10), 3284–3303. https://doi.org/10.1175/MWR-D-10-05007.1.

Du, Y., Qi, L. and Cao, X. (2016) Selective ensemble-mean technique for tropical cyclone track forecast by using time-tagged ensemble and multi-centre ensemble in the western North Pacific. Quarterly Journal of the Royal Meteorological Society, 142(699), 2452–2462. https://doi.org/10.1002/qj.2838.
