Received: 20 April 2018 | Revised: 18 June 2018 | Accepted: 11 July 2018 | Published on: 4 October 2018

DOI: 10.1002/qj.3380

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society

Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble

Florian Pantillon[1] | Sebastian Lerch[2,3] | Peter Knippertz[1] | Ulrich Corsmeier[1]

[1] Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2] Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Correspondence
F. Pantillon, KIT IMK-TRO, Kaiserstr. 12, 76128 Karlsruhe, Germany.
Email: florian.pantillon@kit.edu

Windstorms associated with low-pressure systems from the North Atlantic are the most important natural hazards for central Europe. Although their predictability has generally improved over the last decades, forecasting wind gusts is still challenging, due to the multiple scales involved. One of the first ensemble prediction systems at convection-permitting resolution, COSMO-DE-EPS, offers a novel 2.8-km dataset over Germany for the 2011–2016 period. The high resolution allows representation of mesoscale features that are barely captured by global models, while the long period allows both investigation of rare storms and application of statistical post-processing. Ensemble model output statistics based on a truncated logistic distribution substantially improve forecasts of wind gusts in the whole dataset. However, some winter storms exhibit uncharacteristic forecast errors that cannot be reduced by post-processing. During the passage of the most severe storm, gusts related to a cold jet are predicted relatively well at the time of maximum intensity, whereas those related to a warm jet are poorly predicted at an early phase. Wind gusts are overestimated during two cases of frontal convection, which suggests that even higher resolution is needed to resolve fully the downward mixing of momentum and the stabilization resulting from convective dynamics. In contrast, extreme gusts are underestimated during a rare case involving a possible sting jet, but this arises from the representation of the synoptic rather than the mesoscale. The synoptic scale also controls the ensemble spread, which is inherited mostly from the initial and boundary conditions. This is unsurprising, but leads to high forecast uncertainty in the case of a small, fast-moving cyclone crossing the model domain. These results illustrate how statistical post-processing can help identify the limits of predictability across scales in convection-permitting ensemble forecasts. They may guide the development of regime-dependent statistical methods to improve forecasts of wind gusts in winter storms further.

KEYWORDS

central Europe, COSMO-DE-EPS, extratropical cyclone, frontal convection, predictability, statistical post-processing, sting jet

1 | INTRODUCTION

Extratropical cyclones are important components of the climate system (Catto, 2016). However, the most intense of them—known as winter storms or cyclonic windstorms—are a threat to populations in regions where they frequently occur: for example, on the US West Coast (Mass and Dotson, 2010) and East Coast (Layer and Colle, 2015). Winter storms are even the most important natural hazard over central Europe, where tens of fatalities and billions of euros in damages were caused by extreme storms such as *Lothar* in December 1999 (Wernli *et al.*, 2002) or *Kyrill* in January 2007 (Fink *et al.*, 2009).

European winter storms typically form over the extratropical North Atlantic, although some originate in hurricanes that underwent extratropical transition (Browning *et al.*, 1998). Their intensification is driven mostly by synoptic-scale dynamics, but the strongest gusts recorded during the passage of storms are often due to embedded mesoscale features. Among them, sting jets (Browning, 2004) have received growing interest internationally and in the British Isles in particular, where they have been shown to be frequent features of intense storms (Hart *et al.*, 2017). Convective lines embedded in cold fronts are equally responsible for extreme gusts over the United Kingdom (Earl *et al.*, 2017) and have been exemplified over central Europe by storm *Kyrill* (Ludwig *et al.*, 2015) and even qualified as derechoes (Gatzen *et al.*, 2011). Low-level jets associated with the warm and cold conveyor belts of a cyclone—warm and cold jets, respectively—are more common and can also produce strong, albeit less extreme, gusts (Martínez-Alvarado *et al.*, 2014; Hewson and Neu, 2015), while dry intrusions behind the cold front are responsible for gusts in some extratropical cyclones (Raveh-Rubin and Wernli, 2016).

Representing all these highly dynamic mesoscale features is a challenge for large-scale weather and climate models due to their coarse resolution (Hewson and Neu, 2015). Modelling sting jets requires a horizontal grid spacing of about 10 km and vertical levels separated by about 200 m in the mid-troposphere (Coronel *et al.*, 2016). Capturing convective lines embedded in cold fronts requires even finer horizontal grid spacings of no more than a few km to represent convection explicitly (Ludwig *et al.*, 2015). Climate models can rely on dynamical downscaling to improve the representation of storms (Born *et al.*, 2012), but numerical weather predictions are constrained by errors at the synoptic scale, which exhibit large case-to-case variability and limit useful forecasts to 2–4 days ahead (Pantillon *et al.*, 2017). The representation of vertical stability in the warm and cold sectors of storms is a further challenge and can lead to systematic errors in wind forecasts (Layer and Colle, 2015). Finally, subgrid-scale parametrizations are required to mimic the formation of gusts by the downward transport of high momentum in the boundary layer (Panofsky *et al.*, 1977; Brasseur, 2001). Only large-eddy simulations are able to—at least partly—resolve the formation of gusts (Heinze *et al.*, 2017), but they are not affordable for operational weather forecasts yet.

The predictability of wind gusts during winter storms is investigated here in an ensemble prediction system (EPS) running at convection-permitting resolution. Global EPSs have long shown better performance than deterministic forecasts for early warnings of extreme events such as winter storms (Buizza and Hollingsworth, 2002). Convection-permitting EPSs have a shorter history, as they have been developed in recent years at national weather services, and their focus has been mainly on summer convective precipitation (Schwartz *et al.*, 2015; Raynaud and Bouttier, 2017; Hagelin *et al.*, 2017). In cases of strong synoptic forcing, their uncertainty is expected to be mostly inherited from the larger scale (Keil *et al.*, 2014), but their potential has also been shown for predicting mesoscale features such as snowbands in winter storms (Greybush *et al.*, 2017). The convection-permitting EPS of the Deutscher Wetterdienst (DWD) was one of the first of its kind to become operational (Gebhardt *et al.*, 2008; Peralta *et al.*, 2012). It offers a novel six-year dataset, which encompasses several cases of intense winter storms involving the main mesoscale features: warm jets, cold jets, convective lines, and even a rare sting jet. The long time period further allows the use of statistical post-processing methods to calibrate forecasts and identify systematic model errors. The combination of detailed case studies and statistical analysis thus brings a new perspective on predicting gusts in winter storms. It may both aid understanding of issues related to the representation of specific mesoscale features in a convection-permitting ensemble forecast and guide the development of physically based ensemble post-processing methods that take these features into account.

The article is structured as follows. Section 2 describes the model forecasts, their evaluation, the post-processing methods, and the selection of storms based on observations. Section 3 presents the results for the predictability of gusts, first for the whole dataset and then for 10 severe winter storms, before it details case studies of storms with poor predictability. Section 4 concludes the article with a discussion.

2 | DATA AND METHODS

2.1 | Model forecasts

This article makes extensive use of the operational EPS at convection-permitting resolution of the DWD. The EPS is based on the Consortium for Small Scale Modelling operational forecast for Germany (COSMO-DE: Baldauf *et al.*, 2011), which runs on a rotated grid with 2.8 km horizontal spacing and 50 vertical levels. The resulting COSMO-DE-EPS system contains 20 members using initial and boundary conditions downscaled from the global models of four centers (European Centre for Medium-Range Weather Forecasts, DWD, National Centers for Environmental Prediction, and Japan Meteorological Agency) combined with five sets of physical perturbations (Gebhardt *et al.*, 2008; Peralta *et al.*, 2012). Forecasts up to 21 hr lead time have been run every 3 hr in pre-operational mode since December 9, 2010 and they became operational on May 22, 2012. As expected for an operational system, COSMO-DE-EPS was frequently updated, including an upgrade from COSMO version 4 to version 5 on December 11, 2013 and a switch in the driving DWD global forecast from the former GME to the new ICON model on January 20, 2015. Nevertheless, available COSMO-DE-EPS forecasts for the 2011–2016 period maintained an overall consistent design. A more radical change occurred on March 21, 2017. Lateral boundaries are now driven by the global ICON-EPS system, whereas initial conditions are given by a kilometer-scale ensemble data assimilation system (KENDA: Schraff *et al.*, 2016). This new design is promising, but the available time period is still too short for a statistical analysis and is thus not used here.

Model gusts are output hourly as maximum values over the last hour and are issued from a subgrid-scale parametrization in COSMO, which estimates a turbulent component added to the resolved 10 m wind speed (Schulz, 2008). Following Panofsky *et al.* (1977), the friction velocity u^* is scaled by empirical factors to depict the turbulent component. This approach is comparable to using turbulent kinetic energy and delivers similar results for extratropical storms over Germany in COSMO (Born *et al.*, 2012). In contrast to coarser models, convective gusts are assumed to be explicitly represented here and thus do not require a further component in the gust parametrization. Model gusts were archived for the 2011–2016 period with a limited number of surface and atmospheric variables for the purpose of training statistical models. Corresponding observations of hourly wind gusts were recorded hourly at 175 SYNOP stations of the DWD surface network over Germany and are essential for the verification of model forecasts.

2.2 | Forecast evaluation

Several methods to evaluate COSMO-DE-EPS and post-processed forecasts are introduced here in a general form. Probabilistic forecasts should aim to maximize sharpness subject to calibration (Gneiting *et al.*, 2007). While sharpness refers to the concentration of the predictive distribution, calibration refers to the statistical consistency between the forecast distribution and corresponding observations. Specifically, consider probabilistic forecasts $F_{s,t}$ at station s and time t, and corresponding observations $y_{s,t}$. Calibration of ensemble forecasts can be assessed via verification rank histograms summarizing the distribution of ranks of the observation $y_{s,t}$ when it is pooled with the ensemble forecast $F_{s,t} = \{x_1^{s,t}, \dots, x_m^{s,t}\}$ (Hamill, 2001; Gneiting *et al.*, 2007; Wilks, 2011). Calibrated forecasts result in uniform histograms and deviations from uniformity indicate systematic errors such as biases or lack of spread. For continuous forecast distributions with cumulative distribution function (CDF) $F_{s,t}$ and observation $y_{s,t}$, histograms of the probability integral transform (PIT) $F_{s,t}(y_{s,t})$ provide continuous analogs of verification rank histograms. Verification rank and PIT histograms are usually shown for aggregates over stations s and times t.

For comparative model assessment, proper scoring rules allow simultaneous evaluation of calibration and sharpness (Gneiting and Raftery, 2007). A scoring rule assigns a numerical score to a pair of probabilistic forecast F and corresponding realizing observation y, and is called proper if the expected score is optimized if the true distribution of the observation is issued as forecast (see Gneiting and Raftery, 2007, for details). Here, scoring rules are considered to be negatively oriented, with smaller scores indicating better forecasts. A popular proper scoring rule is the continuous ranked probability score (CRPS; Matheson and Winkler, 1976):

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - 1(y \le z))^2 \, dz, \tag{1}$$

where F denotes the CDF of the forecast distribution with finite first moment, y denotes the observation, and $1(y \le z)$ is an indicator function that is 1 if $y \le z$ and 0 otherwise. The integral in Equation 1 can be computed analytically for ensemble forecasts and a variety of continuous forecast distributions (for example, Jordan *et al.*, 2017). The continuous ranked probability skill score (CRPSS) is further defined as

$$\mathrm{CRPSS}(F, y) = 1 - \frac{\mathrm{CRPS}(F, y)}{\mathrm{CRPS}(F_{\mathrm{ref}}, y)},$$

where F_{ref} denotes the CDF of a reference forecast. The CRPSS is positively oriented and can be interpreted as relative improvement over the reference. The CRPSS is usually computed as a skill score of CRPS averages.

Finally, in order to assess forecast quality for extreme events, the Brier score (BS: Brier, 1950),

$$\mathrm{BS}_z(F, y) = (F(z) - 1(y \le z))^2,$$

is computed for high thresholds z. The Brier score is a proper scoring rule for probabilistic forecasts of binary events, and the CRPS in Equation 1 corresponds to the integral over the Brier score at all real-valued thresholds (Hersbach, 2000). As before, the Brier skill score (BSS),

$$\mathrm{BSS}(F, y) = 1 - \frac{\mathrm{BS}(F, y)}{\mathrm{BS}(F_{\mathrm{ref}}, y)},$$

allows us to assess improvements relative to a reference forecast F_{ref}.

2.3 | Statistical post-processing

Ensemble forecasts typically show systematic biases and lack calibration. Calibration here refers to the statistical consistency between predictions and observations; a probabilistic forecast is called calibrated if observations cannot be distinguished from a random draw from the predictive distribution. Ensemble forecasts thus require some form of statistical post-processing. The non-homogeneous regression or ensemble model output statistics (EMOS) approach proposed by Gneiting *et al.* (2005) is followed here. In this approach, the forecast distribution is given by a single parametric distribution with parameters depending on the ensemble forecasts through suitably chosen link functions. EMOS models have been developed for a variety of weather variables, such as temperature, pressure, wind speed, and precipitation. However, work on wind gusts is sparse. Thorarinsdottir and Johnson (2012) propose a truncated Gaussian distribution EMOS model based on ensemble forecasts of wind speed and gust factors. Outside the EMOS framework, Oesting *et al.* (2017) develop a spatial post-processing model for extreme wind gusts utilizing conditional simulation procedures from extreme value theory, an approach that was also followed by Friederichs *et al.* (2018). Staid *et al.* (2015) compare statistical models for wind gust prediction at offshore locations based on predictors from output of global weather models. They only consider deterministic forecasts, but the methodology might be extended towards ensemble post-processing by including summary statistics from ensemble predictions.

Here, an EMOS model for wind gusts is built on earlier works for wind speed by Messner *et al.* (2014) and Scheuerer and Möller (2015). The conditional distribution of wind gust y given ensemble forecasts $x_1, \ldots, x_m$ is modeled as

$$y|x_1, \ldots, x_m \sim \mathcal{L}_{[0,\infty)}(y|\mu, \sigma),$$

where $\mathcal{L}_{[0,\infty)}$ denotes a logistic distribution truncated at 0 with location $\mu \in \mathbb{R}$, scale $\sigma > 0$, and probability density function

$$f(z) = \frac{1+e^{\frac{\mu}{\sigma}}}{e^{\frac{\mu}{\sigma}}} \cdot \frac{e^{-\frac{z-\mu}{\sigma}}}{\sigma\left(1+e^{-\frac{z-\mu}{\sigma}}\right)^2} \qquad \text{for } z \geq 0,$$

and $f(z) = 0$ otherwise. The location parameter is modeled as a linear function of the ensemble mean $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$,

$$\mu = a + b\,\bar{x},$$

and the squared scale parameter is modeled as a linear function of the ensemble variance $s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$,

$$\sigma^2 = c + d\,s^2.$$

Alternative EMOS models for wind speed proposed by Lerch and Thorarinsdottir (2013), Baran and Lerch (2015), and Scheuerer and Möller (2015) have also been tested here and indicate only minor differences in predictive performance.

The EMOS model parameters a, b, c, d are estimated by minimizing the mean CRPS over a rolling training period consisting of forecasts and observations from the previous n days. Two variants of the model with different spatial composition of the training set are considered. The *global* model composites data from all stations to form a single training set, from which a single set of coefficients a, b, c, d for all stations is estimated. By contrast, the *local* model considers only forecast cases from the single observation station of interest and generates a different set of coefficients for each station. The local model accounts for spatial variability of the forecast errors, but requires longer training periods. In both variants, only previous EPS model runs with the same initialization time and forecast lead time are used for model estimation. Following common practice from the post-processing literature, the training period lengths are chosen by testing different values to minimize the CRPS. Here, this leads to setting the training period length to $n = 30$ days for the global model and $n = 100$ days for the local model, although the influence of different training period lengths is generally small.

Note that the focus here is on devising a post-processing model that is sufficiently simple to allow for straightforward interpretation of the model deficiencies and potential improvements in predictability of wind gusts. Alternative, more demanding modeling and estimation approaches (for example, Junk *et al.*, 2015; Dabernig *et al.*, 2017; Lerch and Baran, 2017) may result in improvements in predictive performance but impede inference on some features of EPS model error characteristics, such as station-specific biases.

2.4 | Selection of storms

Gust observations from the DWD surface network over Germany are available as daily maximum values since several decades ago and as early as the 1950s for some stations. They are used here to identify significant storms from a climatological perspective. Strong gusts are responsible for most damages within winter storms and their impact increases nonlinearly with strength but also depends on the vulnerability of the infrastructure, which is usually adapted to local climate conditions. Following Klawa and Ulbrich (2003), these factors are taken into account by defining the Storm Severity Index (SSI) as

$$\mathrm{SSI} = \sum_{\text{stations } s} \left\{ \left(\frac{v_{\max,s}}{v_{98,s}} - 1 \right)^3 \right\}_{v_{\max,s}>v_{98,s}} \qquad (2)$$

where $v_{\max,s}$ is the daily maximum wind gust and $v_{98,s}$ its local 98th climatological percentile at station s. Values of $v_{98,s}$ are extreme at some mountain stations and they are generally higher in coastal regions than over the mainland, due to higher exposure to wind gusts (see, for example, Figure S1 in File S1). The SSI is computed as a sum over all stations where $v_{\max,s} > v_{98,s}$ and thus depends on the number of stations reporting gusts. This number has remained stable in the past years, but it has increased dramatically in previous decades and must be corrected for in longer time series.

The most severe storms of the 2011–2016 period are selected based on the SSI computed from all stations of the DWD surface network reporting daily maximum gusts. Days with SSI $>$ 1 are listed in Table 1. The SSI is a relative rather than absolute value, as it depends on the number of stations with available data (Equation 2). The threshold SSI $>$ 1 is thus arbitrary, but it appears a reasonable value to select significant winter storms, as discussed in the following. This results in 16 days, which span a broad spectrum of events and cover one order of magnitude in SSI (Table 1). Somewhat unexpectedly, five events occurred in summer and exhibit SSI comparable to weak winter storms. They involve convection in the first place but within different processes, ranging from a bow echo with extreme local gusts on June 9, 2014 (Pentecost storm; Barthlott *et al.*, 2017;

TABLE 1 Storms with SSI > 1 of the 2011–2016 period and earlier storms with SSI > 10 of the 1997–2010 period for comparison. Winter storms are named according to the Free University of Berlin "Adopt a Vortex" program, except for *Gonzalo*, which was a former tropical cyclone and thus named by the National Hurricane Center. Insured losses are taken from Gesamtverband der Deutschen Versicherungswirtschaft (2017)

Date	Name	SSI	Insured losses (M euro)
June 22, 2011	summer storm	1.0	<100
December 16, 2011	*Joachim*	1.7	<100
January 5, 2012	*Andrea*	2.1	180 (with 3 January)
June 30, 2012	summer storm	1.2	120
August 6, 2013	summer storm	1.8	220
October 28, 2013	*Christian*	5.1	400
December 5, 2013	*Xaver*	2.9	150 (with 6 December)
December 6, 2013	*Xaver*	1.1	*see above*
June 9, 2014	summer storm	1.0	450
October 21, 2014	*Gonzalo*	2.5	<100
January 9, 2015	*Elon*	1.9	150 (with 10 January)
January 10, 2015	*Felix*	2.6	*see above*
March 31, 2015	*Niklas*	12.0	590
July 7, 2015	summer storm	1.6	120
February 8, 2016	*Ruzica*	1.0	<100
February 9, 2016	*Susanna*	1.3	<100
December 26, 1999	*Lothar*	17.1	800
October 28, 2002	*Jeanett*	14.3	760
January 18, 2007	*Kyrill*	29.0	2060
March 1, 2008	*Emma*	13.2	390

Mathias *et al.*, 2017) to a cold front covering a broader area on June 22, 2011 (Weijenborg *et al.*, 2015). Although they differ from winter events and their detailed investigation is beyond the scope of the article, these summer events reveal the continuum between purely convective and large-scale dynamics (see also the footprints of all 16 days in Figure S2 in File S1). The range of processes among winter storms will be discussed further in Section 3.3.

According to the SSI, the most severe storm of the 2011–2016 period by far is *Niklas* on March 31, 2015 (Table 1), which caused widespread loss of forest cover in southern Germany (Einzmann *et al.*, 2017). *Niklas* is one of the most severe storms over Germany in records from the DWD surface network and reaches rank 14 since the 1970s in SSI corrected for the number of stations (see Figure S3 in File S1). Strong gusts were recorded during storms in the 1950s and 1960s, but available observations are sparse and mostly limited to former West Germany. Compared with the most severe storms of the past two decades (Table 1), *Niklas* approaches *Lothar* (1999), *Jeanett* (2002) and *Emma* (2008) but remains far behind the extreme storm *Kyrill* (2007).

Among the 16 selected storms of the 2011–2016 period, *Niklas* is followed in severity by *Christian* on October 28, 2013 (Table 1). These two storms were indeed responsible for the highest insurance losses over Germany, as expected from the rationale behind the definition of SSI (Klawa and Ulbrich, 2003). For weaker storms, insured losses scale less well with SSI and a precise estimate of damages would require a more sophisticated wind-loss relationship (Prahl *et al.*, 2015). Insured losses further depend on the population density and insured portfolio, but also on joined hazards such as snow in winter storms and hail in summer storms, as well as indirect impacts of wind, such as the storm surge caused by *Xaver* on December 5–6, 2013 (Dangendorf *et al.*, 2016). Finally, the impact of storms depends on the time of year they occur—for example, trees are more affected before they lose their leaves—but also on the coincidence with festivities such as *Ruzica* and *Susanna* during Carnival on February 8–9, 2016. Despite these limitations, the selection based on SSI > 1 captures all significant winter storms with insured losses above 100 million euros during the 2011–2016 period.[1] Excluding summer cases results in 10 winter storms, which are selected for investigation of the predictability of gusts.

3 | PREDICTABILITY OF WIND GUSTS

The predictability of wind gusts in COSMO-DE-EPS is first characterized using the whole dataset to assess the quality of the raw forecast and quantify the added value brought by global and local post-processing models in a statistical sense. The forecast quality becomes crucial in storm situations to issue precise warnings, thus the predictability of wind gusts is further investigated in raw and post-processed forecasts during 10 selected cases of winter storms. Outliers with uncharacteristic forecast error are identified and their dynamics are explored in detailed case studies.

3.1 | Whole dataset

In the following, the predictability of wind gusts is assessed in forecasts between June 2011 and December 2016, earlier months being used as training period. No visible effect of the initialization time was found, thus only model runs initialized at 0000 UTC are considered here.

Figure 1 shows verification rank and PIT histograms of raw and post-processed forecasts. The U-shaped verification rank histogram of the ensemble forecasts indicates a systematic lack of spread, as observations frequently fall outside the range of ensemble forecasts. The lack of spread is also reflected in the lack of reliability of raw forecasts for threshold exceedances, at high values in particular (see Figure S4 in File S1). The shape of the verification rank histogram further indicates that members tend to cluster into four groups driven by initial and boundary conditions from the four global models (Figure 1a). By contrast, the PIT histograms of the post-processed forecasts show much smaller deviations from the desired uniformity and thus are much better calibrated,

[1] Storms tend to cluster in series due to favorable large-scale conditions (Pinto *et al.*, 2014), which prevents unambiguously attributing both SSI and insured losses to single storms in some cases (for example, *Elon* and *Felix*; *Ruzica* and *Susanna*).

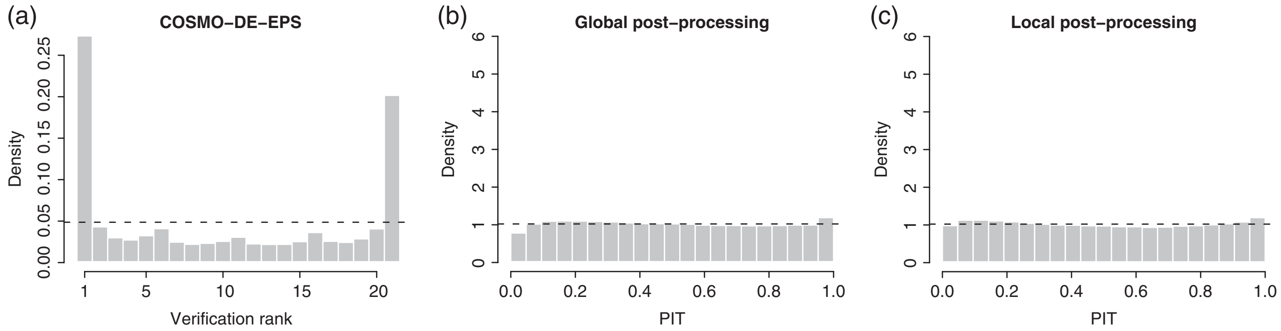

FIGURE 1 Verification rank of (a) raw forecasts and PIT histograms of post-processed forecasts using (b) global and (c) local models for wind gusts between June 2011 and December 2016 at 12 hr lead time

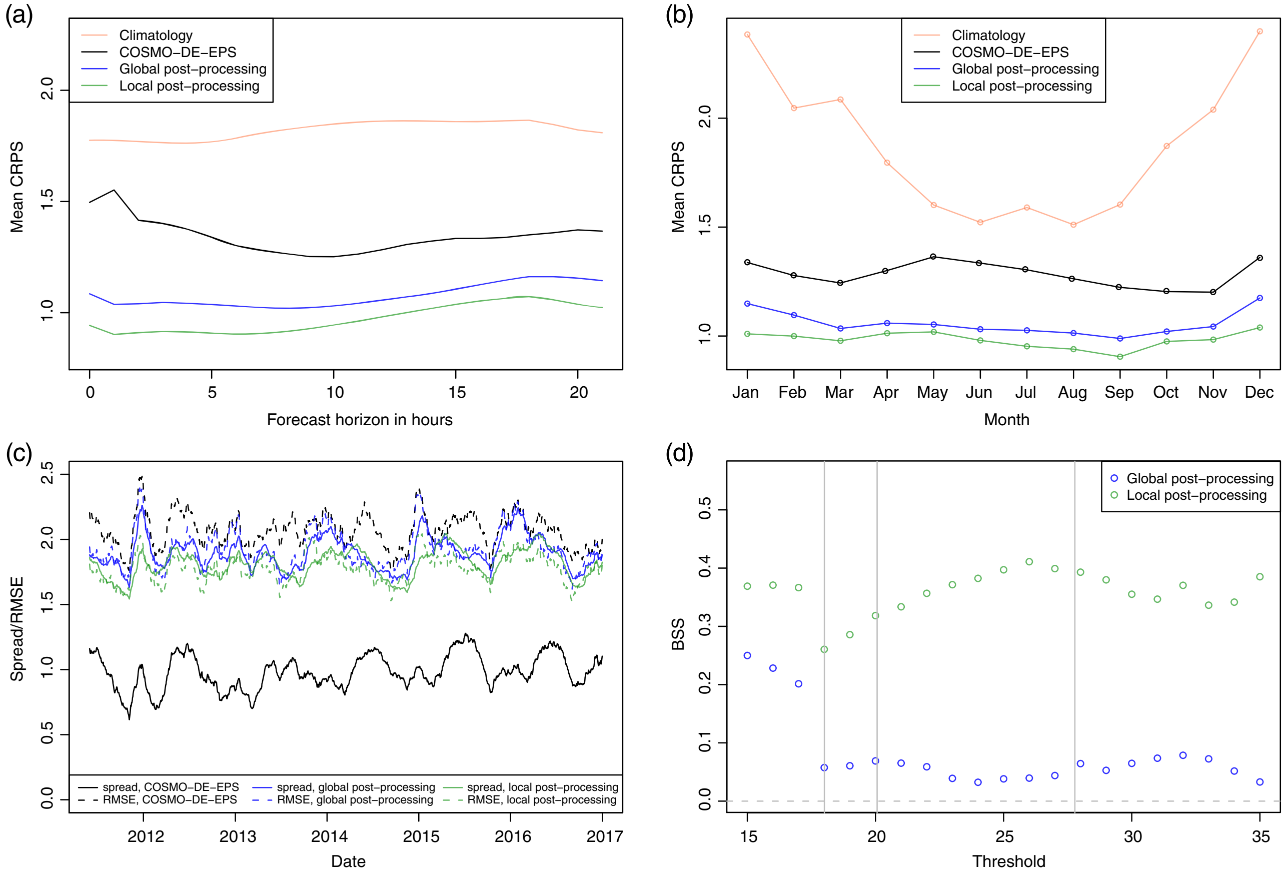

FIGURE 2 Comparison between raw and post-processed forecasts of wind gusts for the whole dataset: (a) mean CRPS as function of forecast lead time, (b) mean CRPS of 12 hr forecasts averaged by month, (c) daily averages of spread and RMSE for 12 hr forecasts, shown as averages over a 50 day rolling window, and (d) mean Brier skill score of 12 hr forecasts for the exceedance of different thresholds. The vertical gray lines in (d) indicate the 98th, 99th and 99.9th percentiles of all gust observations during the June 2011–December 2016 period [Colour figure can be viewed at wileyonlinelibrary.com]

particularly the local model (Figure 1b,c). The PIT histogram of the global model indicates slight underprediction of high values and overprediction of low values.

The predictive performance of the ensemble forecasts increases with lead time up to around 10 hr, as shown by the decrease in CRPS in Figure 2a. This is due to a severe lack of spread in the ensemble forecast for short lead times (see Figure S5 in File S1). In contrast, the predictive performance after post-processing is substantially better and remains almost constant during this time window (Figure 2a). The forecast quality of both the raw ensemble and the post-processed forecasts decreases at longer lead times, as indicated by the increase in CRPS after 10 hr. However, post-processing substantially improves the ensemble predictions over the entire forecast period. In the following, only 12 hr forecasts are used to illustrate the forecast quality, but the results are also valid for other lead times.

Figure 2b shows monthly mean CRPS values of the raw and post-processed COSMO-DE-EPS forecasts. Post-processing consistently improves the ensemble predictions, the corresponding skill scores ranging between 13 and 23% for the global model, and between 18 and 27% for the local model (not shown). The raw ensemble shows higher CRPS values in winter and spring (Figure 2b), when wind gusts are stronger on average. The relative improvement through global post-processing shows a stronger seasonal cycle and is generally higher in spring. The local model shows further improvement compared with the global model, particularly in winter. This might indicate that errors of the ensemble forecasts are more systematic over the whole domain in spring and more station-specific in winter. For comparison, a simple climatology computed locally over the previous 100 days shows a strong seasonal cycle in CRPS, but remains substantially worse than the raw ensemble during the whole year and at all lead times (Figure 2a,b).

The local improvement via post-processing can partially be explained by the local variability of the bias of the raw forecast. The mean bias of the ensemble mean is positive for 120 of the 175 stations, and depends strongly on the observation station location (see Figure S6 in File S1). Due to the single set of model coefficients for all stations, global post-processing is unable to account for the station-specific variability. While it decreases the overall mean bias, leading to improvements at stations with positive bias, it worsens mean forecasts at stations with negative bias. By contrast, local post-processing is able to remove the station-specific biases on average. The bias depends further on the magnitude of the ensemble mean, which underestimates weak gusts below 7 m/s but overestimates moderate to strong gusts above 10 m/s by around 1 m/s (see Figure S7 in File S1). Note that the mean bias is not a proper measure of forecast accuracy and seasonal effects may average out over the entire period. To assess the accuracy of mean forecasts, the root-mean-square error (RMSE) provides a suitable alternative.

Apart from the bias, a main reason for the higher CRPS values of the raw ensemble is the lack of spread. Figure 2c shows that the spread is about twice too small compared with the RMSE in the raw ensemble, whereas after post-processing the spread is increased and matches the slightly reduced RMSE much better. While both spread and RMSE of raw and post-processed ensemble show a seasonal cycle similar to the CRPS (Figure 2b), the spread of the raw ensemble increases further slowly over time (Figure 2c). This likely results from the frequent updates of COSMO-DE-EPS discussed in Section 2.1, although none of the described upgrades corresponds to a clearly identifiable change in forecast performance during the time period under consideration.

To ensure that the post-processing models improve the forecast of damaging gusts, mean Brier skill scores relative to the raw ensemble forecasts are computed for the exceedance of high thresholds between 15 and 35 m/s (Figure 2d). All BSS values are positive, which clearly indicates that post-processing actually improves forecasts of strong gusts compared with the raw ensemble. However, the improvement obtained by global post-processing decreases towards BSS values of only around 5% for higher thresholds. By contrast, local post-processing yields consistently high relative improvements between 25 and 40% for all threshold values considered.[2] Accounting for station-specific error characteristics thus appears to be of importance for skilful probabilistic forecasts of damaging gusts.

In summary, post-processing improves various aspects of forecast quality and predictability of wind gusts consistently and substantially compared with the raw ensemble. The larger relative improvements obtained through local post-processing indicate strongly station-specific error characteristics, particularly in winter. The predictability assessment for the selected storms and case studies presented in the following thus focuses on comparisons with the local post-processing model only.

3.2 | Ten selected storms

The predictability of wind gusts is investigated in raw and post-processed ensemble forecasts for the 10 most severe winter storms of the dataset based on the SSI (Table 1). *Niklas*—the most severe storm in terms of SSI—is also the most intense storm in terms of observed wind gusts averaged over all stations, while other storms with high SSI, such as *Christian* and *Gonzalo*, show comparatively weaker gusts (Figure 3a). For each storm, the initialization time is chosen such that the maximum intensity is reached after 12–15 hr lead time. This allows storms to develop in the forecasts and also appears as a relevant range for issuing warning in an operational framework. Quantitatively comparing forecasts of storms with different dynamics—fast or slow-moving, tracking across or at the edge of the model domain, and with widespread or concentrated wind fields—is challenging based on hourly wind gusts only. However, the results succeed in highlighting outliers and are generally consistent with earlier or later initialization times.

The predictability is first measured with the CRPS of the raw ensemble. Surprisingly, while it exhibits high CRPS in an early phase, *Niklas* shows relatively low CRPS during its period of maximum intensity (Figure 3b). In contrast, higher CRPS is reached by *Andrea* and *Christian* at 14–17 hr lead time, that is, shortly after and during the maximum intensity, respectively. Beyond these intense storms, weaker storms *Gonzalo* and *Susanna* also exhibit peaks of relatively high CRPS. This emphasizes that the predictability is not related to the intensity of wind gusts only. For both *Andrea* and *Christian*, the ensemble mean is strongly biased compared with observations, indicating systematic over- and underestimation of gusts, respectively (Figure 3c). Following *Andrea*, *Gonzalo* also stands out with positive bias during its period of maximum intensity, while other storms generally exhibit negative

[2] Note that the BSS values show a discontinuity between 17 and 18 m/s. This is likely due to the conversion of recorded wind gust observations from kt to m/s, leading to only one possible wind gust value between 17 and 18 m/s, but two possible values between all other pairs of adjacent thresholds under consideration, causing a discontinuity in the climatological event frequencies.

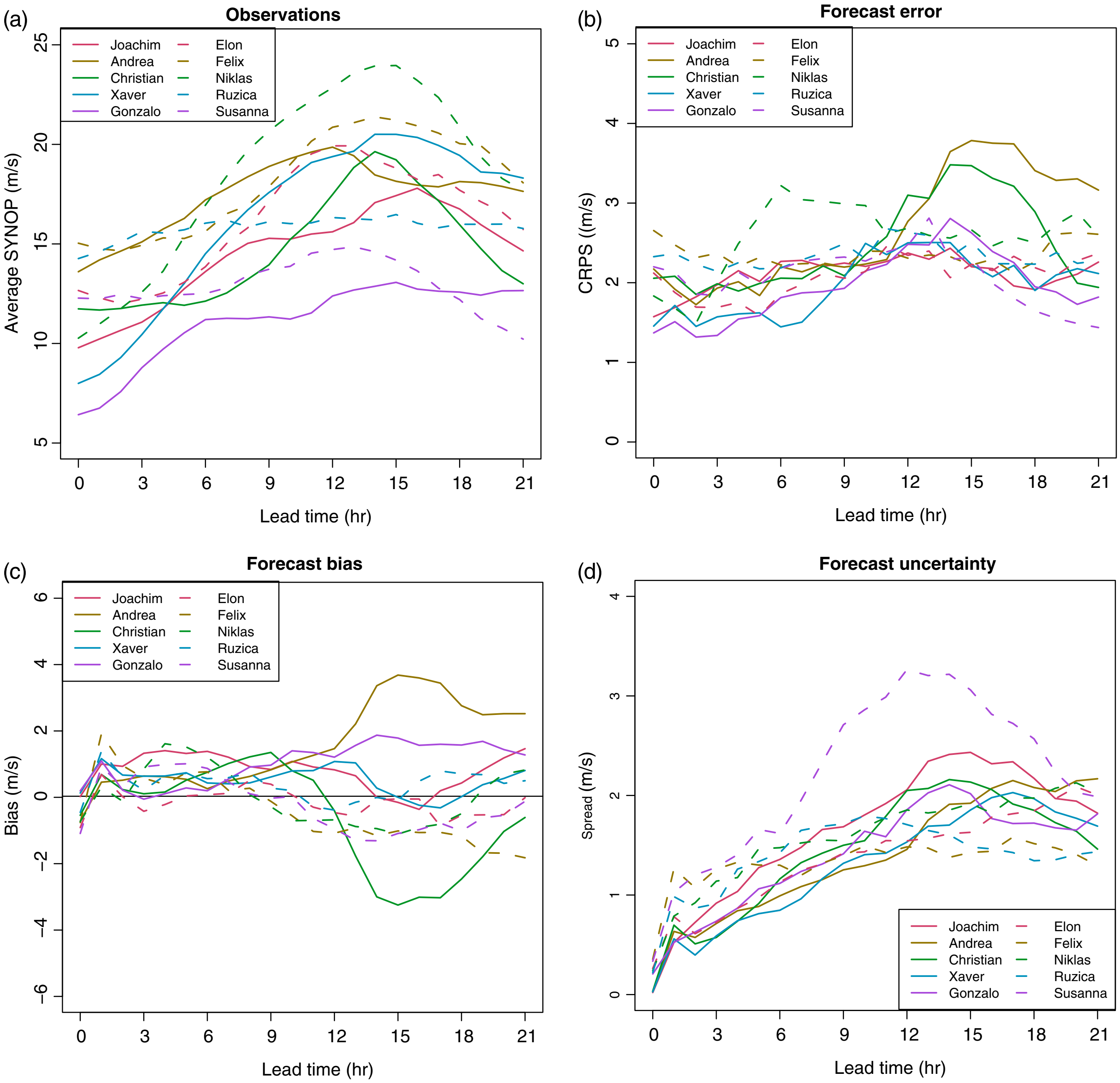

FIGURE 3 Wind gusts during the 10 selected winter storms (see Table 1): (a) observations averaged over all available stations and evaluation of raw forecasts with (b) CRPS, (c) mean bias, and (d) ensemble spread. The initialization time is chosen for each storm such that the maximum occurs after 12–15 hr lead time in (a) [Colour figure can be viewed at wileyonlinelibrary.com]

biases at this stage. Biases remain small compared with the average wind gusts—below 10% in most cases and about 20% for outliers—but can reach large values for some storms locally, as will be discussed in Section 3.3.

As a measure of forecast uncertainty, the ensemble spread is an additional, important property of the EPS. For all storms, it quickly increases with lead time and peaks during the period of maximum intensity (Figure 3d). The weak storm *Susanna* exhibits high spread and thus appears as an outlier with large forecast uncertainty. By contrast, the spread barely reaches half of the RMSE for all other storms (not shown) and thus indicates the underdispersiveness of the ensemble. In the case of *Susanna*, the high ensemble spread is due to the perturbation of lateral boundary conditions. This is illustrated by grouping all members according to corresponding global models, which reveals four diverging scenarios, ranging from a clear peak to a decrease in wind gust intensity (Figure 4a). The four scenarios are in turn related to four different tracks and depth of the associated low-pressure system, which crossed the model domain within one day (Figure 4b,c). This emphasizes that high synoptic-scale uncertainty can be found in global forecasts, even at short range. For the other storms, the ensemble spread is lower but it is also mostly inherited from the four global models (not shown). Only one member of each group—corresponding to a specific perturbation of the boundary-layer parametrization—systematically

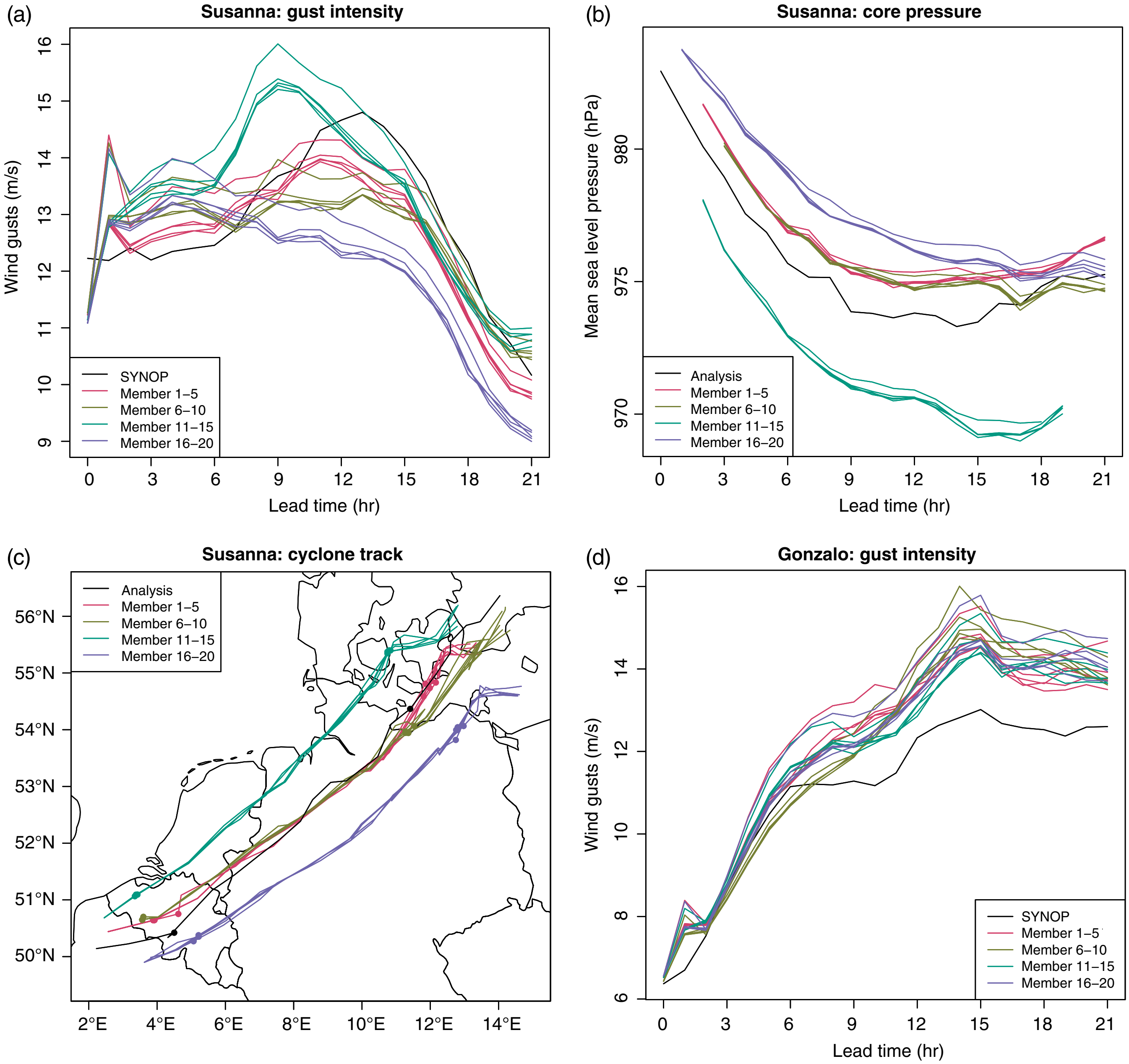

FIGURE 4 Time series of average gusts in observations (black curves) and ensemble members (curves coloured by group of same forcing global model) initialized (a) at 0900 UTC on February 9, 2016 for storm *Susanna* and (d) at 0600 UTC on October 21, 2014 for storm *Gonzalo*. Also shown are time series of (b) mean sea-level pressure and (c) track of the center of storm *Susanna* as in (a) but with black curves denoting the COSMO-DE analysis. Dots in (c) indicate the position at 1200 UTC ($t+3$) and 0000 UTC ($t+15$) [Colour figure can be viewed at wileyonlinelibrary.com]

stands out by stronger wind gusts, as illustrated by the case of *Susanna* (Figure 4a). In the case of *Gonzalo*, physical perturbations are more efficient to increase the ensemble spread and their contribution is similar to that of the four global models (Figure 4d). However, the ensemble clearly remains underdispersive in this case, as none of the ensemble members captures the observed peak in wind gusts.

Applying statistical post-processing improves the raw forecast during the first 12 hr, as measured by the CRPSS, but large variability is found between storms at longer lead times (Figure 5a). In particular, storms *Andrea* and *Christian* are again extreme cases, with the strongest improvement and worsening, respectively, compared with the raw forecast. This dramatically increases the CRPS for *Christian*, which becomes by far the case with the poorest predictability of the sample, while the other storms stay close together (Figure 5b). The increase in CRPS can be explained through the impact of post-processing on the mean bias, which systematically decreases from positive to negative values on average (Figure 5c). This partially compensates the positive bias in the case of *Andrea* but adds to the negative bias in the case of *Christian* and thus strengthens the absolute forecast error. Post-processing also strongly increases spread, at short

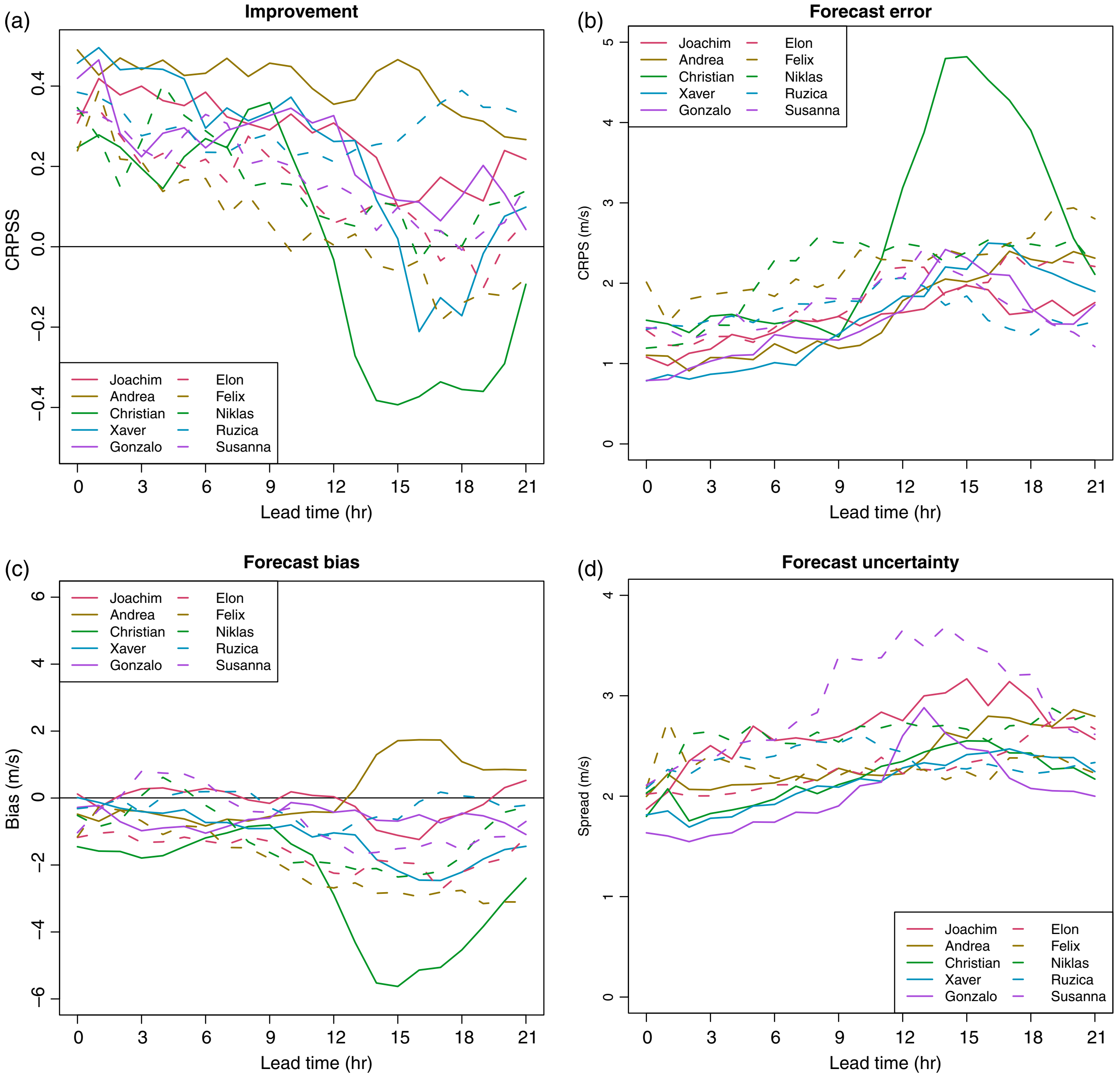

FIGURE 5 Evaluation of post-processed forecasts using the local model for wind gusts during the 10 selected storms: (a) CRPSS compared with the raw ensemble and (b) CRPS, (c) mean bias, and (d) ensemble spread, as in Figure 3b–d [Colour figure can be viewed at wileyonlinelibrary.com]

lead times in particular (Figure 5d). The forecasts become better calibrated, but *Susanna* remains an outlier with high spread.

These results show that applying statistical post-processing generally improves the predictability of storms, although outliers with high CRPS, bias and spread in the raw forecast still stand out after calibration. Furthermore, the characteristics of hourly wind gusts for these few cases are also found in the hourly average wind speed and in maximum wind gusts over the whole forecast time range, which shows that they are not due to approximations in the gust parametrization or to timing errors only (see Figures S8 and S9 in File S1). Altogether, this suggests that the predictability of outliers is related to specific physical processes that cannot be completely corrected with a statistical approach and thus motivates detailed case studies in the following.

3.3 | Case studies

Based on the results above, the dynamics of storms showing uncharacteristic forecast errors are investigated here in detail. The strong negative bias during storm *Christian* is mainly due to a few stations located in northern Germany over or near the North and Baltic Seas. These stations recorded extreme gusts at that time (Figure 6a), which were strongly underestimated in the ensemble forecast (Figure 6b). The COSMO-DE analysis shows that they are located directly south of the cyclone centre and correspond to the region of strongest winds above the boundary layer (Figure 6c). The intensity of wind gusts

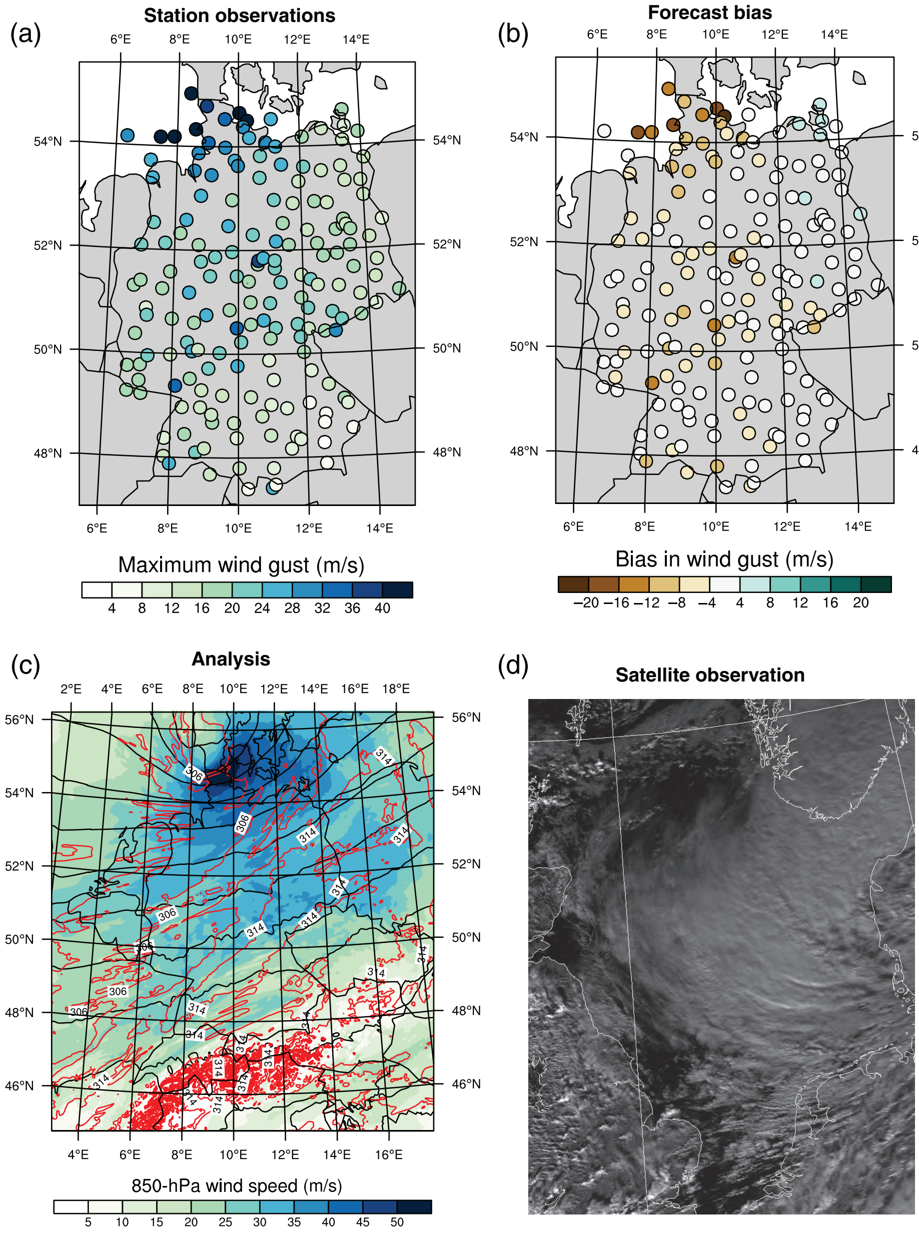

FIGURE 6 Storm *Christian* at 1400 UTC on October 28, 2013 ($t + 14$): wind gusts in (a) station observations and (b) the mean bias of the raw forecast; 850 hPa wind speed (shading, in m/s), geopotential (black contours every 20 gpdam), and equivalent potential temperature (labeled contours every 4 K) in (c) the COSMO-DE analysis and (d) AVHRR channel 9 satellite observation at 1300:3900 UTC over the North Sea [Colour figure can be viewed at wileyonlinelibrary.com]

and their location suggests that they may originate from a sting jet (Browning, 2004), which is supported by the presence of mesoscale bands in the cloud head (Figure 6d). Indeed, Browning *et al.* (2015) identified a sting jet during the earlier passage of *Christian* over southern England using observations from a high-resolution Doppler radar and a network of high-frequency surface stations. Such observations are not available for northern Germany and extreme gusts may alternatively be due to a cold jet, which was also identified during the passage of *Christian* (Browning *et al.*, 2015). Distinguishing the two low-level jets is challenging and often requires trajectory calculations within the air streams (Coronel *et al.*, 2016). This is beyond the scope of the study, but it is likely that both sting and cold jet contributed to the extreme gusts recorded at the North and Baltic Seas.

In addition to surface wind gusts, the wind speed above the boundary layer is also underestimated in the ensemble mean compared with the COSMO analysis close to the center of *Christian* (not shown). This suggests that the negative bias is related to the representation of synoptic- and mesoscale features of the storm and not of wind gusts only. One representative ensemble member issued from each of the four global models is further displayed at the time of landfall on the North Sea coast, which occurs between 1100 and 1500 UTC in forecasts (Figure 7). Other members are either almost identical or show differences in intensity but not in pattern. All four representative members predict the strongest winds to affect

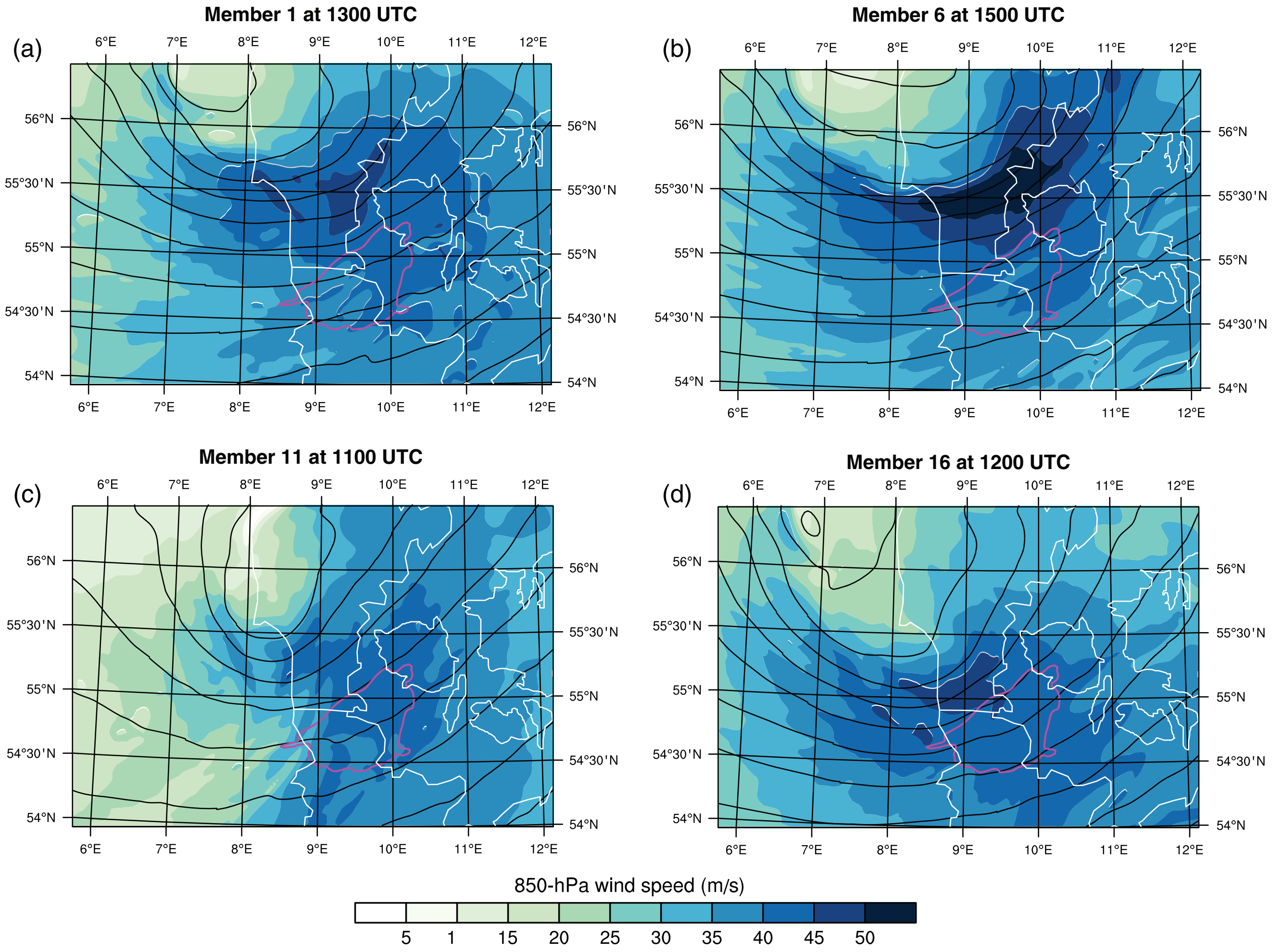

FIGURE 7 Predicted 850 hPa wind speed (shading, in m/s) and geopotential (black contours every 20 gpdam) of storm *Christian* on October 28, 2013 in the ensemble: (a) member 1 at 1300 UTC, (b) member 6 at 1500 UTC, (c) member 11 at 1100 UTC, and (d) member 16 at 1200 UTC. The thick contour shows the 850 hPa wind speed exceeding 50 m/s in the COSMO-DE analysis at 1400 UTC (see Figure 6c) [Colour figure can be viewed at wileyonlinelibrary.com]

Denmark, while they occur over Germany in the analysis (thick contour in Figure 7). The northward shift appears as the main cause of underestimation of wind gusts at DWD stations. However, the wind speed is generally high south of the cyclone center and exceeds (c) 40 m/s, (a, d) 45 m/s, and (b) even 50 m/s, depending on the ensemble member. Some model forecasts may thus actually be able to develop a sting jet, but they do not predict the synoptic scale correctly. This deficiency in turn appears inherited from the driving global models, although the spin-up of forecasts may also contribute, due to the track of the storm at the edge of the domain. The performance of the gust parametrization in this case is still unclear and may be a further limitation that cannot be investigated here. Finally, statistical post-processing amplifies the error by correcting for the systematic overestimation of gusts in the region. All these factors restrain the predictability of Christian.

The positive bias in wind gusts during storm *Andrea*, in contrast, is related to continuous strengthening predicted by all ensemble members, while observed gusts reach a peak and start weakening (Figure 3a). The positive bias occurs in a region of relatively weak gusts in observations (Figure 8a,b). This region is located behind a zonally oriented convective line embedded in the cold front of the cyclone and crossing central Germany southward (Figure 8c,d). Strong gusts are widespread in the warm sector and do not appear to be enhanced by the convective line, which denotes a classical warm jet situation. This means that the overestimation of gusts in the ensemble members is due to their lack of ability to capture the drop of intensity after the passage of the cold front. As in the case of *Christian*, the bias is already present in the wind speed above the boundary layer compared with the COSMO analysis and thus cannot be attributed to the gust parametrization only (not shown). Furthermore, the intensity of convective precipitation is systematically underestimated in all COSMO-DE-EPS members (see Figure S10 in File S1). This suggests that the convective dynamics of the cold front and the resulting stabilization of lower levels are not correctly represented in the model forecasts, despite the convection-permitting resolution allowed by the 2.8 km grid spacing. A lack of convective organization was found during summer cases over Germany using COSMO ensemble simulations with the same grid spacing, which points to issues related to the boundary-layer parametrization (Rasp *et al.*,

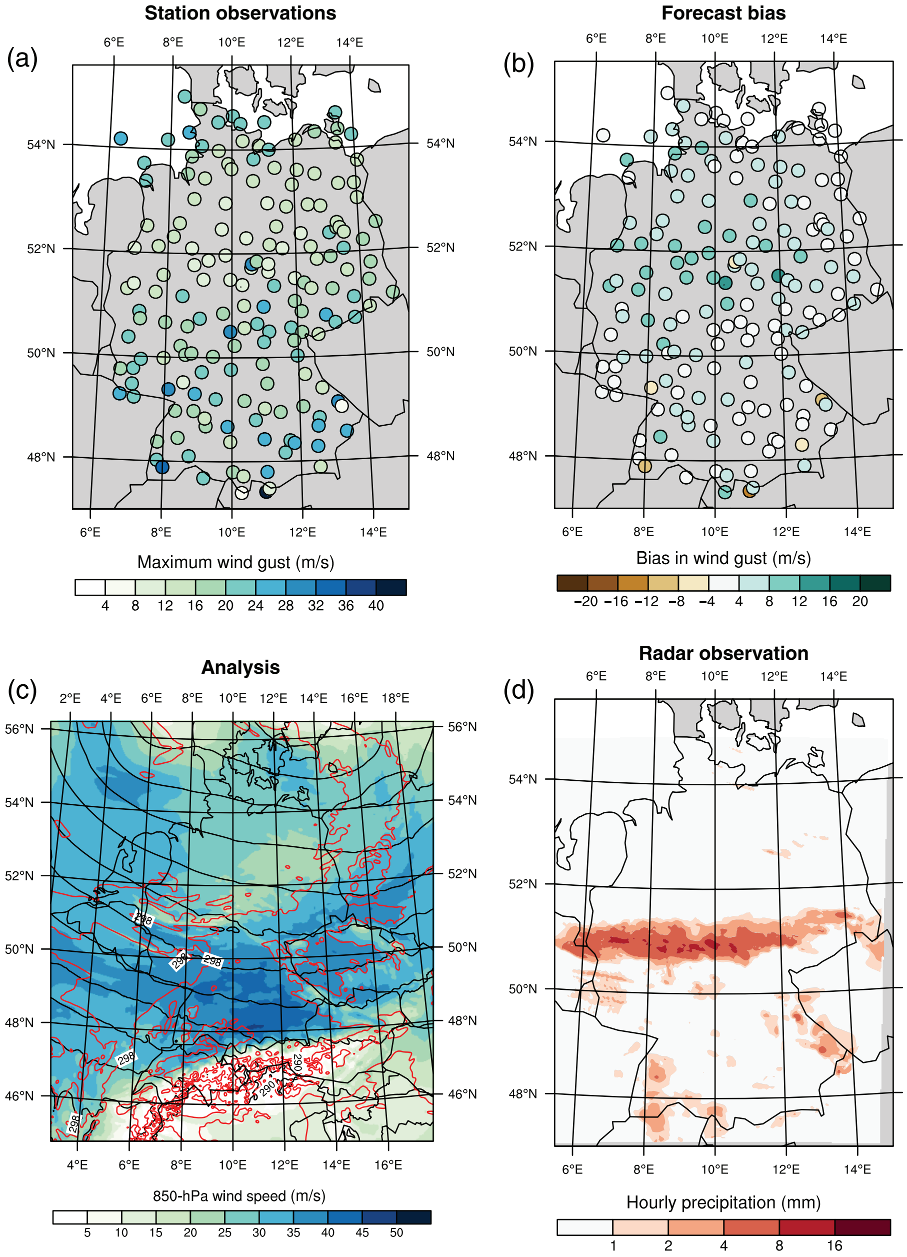

FIGURE 8 As Figure 6 but for storm *Andrea* at 0900 UTC on January 5, 2012 ($t + 15$) and with hourly precipitation derived from the DWD radar network (in mm; d) [Colour figure can be viewed at wileyonlinelibrary.com]

2018). In the case of *Andrea*, the total bias is reduced by statistical post-processing but the positive bias related to convective dynamics persists after calibration and *Andrea* remains an outlier (Figure 5c).

Convective dynamics are also involved during the passage of storm *Gonzalo* and again result in positive bias, albeit smaller (Figure 3c). Intense convection is embedded in the active cold front of the cyclone over southeastern Germany, which involves large gradients of equivalent potential temperature θ_E (Figure 9c,d). In this case, strong gusts are recorded locally along the convective line and are thus mainly produced by the downward transport of momentum from higher levels (Figure 9a). The small spatial extent of strong gusts contrasts with *Andrea*, where they are widespread and thus the contribution of the convective line is not clearly discernible (Figure 8). This results in a more scattered bias for *Gonzalo*, with both over- and underestimation locally in the ensemble mean compared with observations (Figure 9b). No homogeneous bias is visible in the wind speed above the boundary layer either (not shown), and ensemble members exhibit some variability related to stochastic physics within each group issued from the same global model (Figure 4d). The error and uncertainty in this case highlight the challenging prediction of the precise location and intensity of gusts driven by convection. The subgrid-scale parametrization may contribute further to the overestimation by adding a turbulent contribution to gusts that are already explicitly represented by model dynamics.

Finally, large errors during the early phase of storm *Niklas* occurred in the morning, at a time when strong gusts were

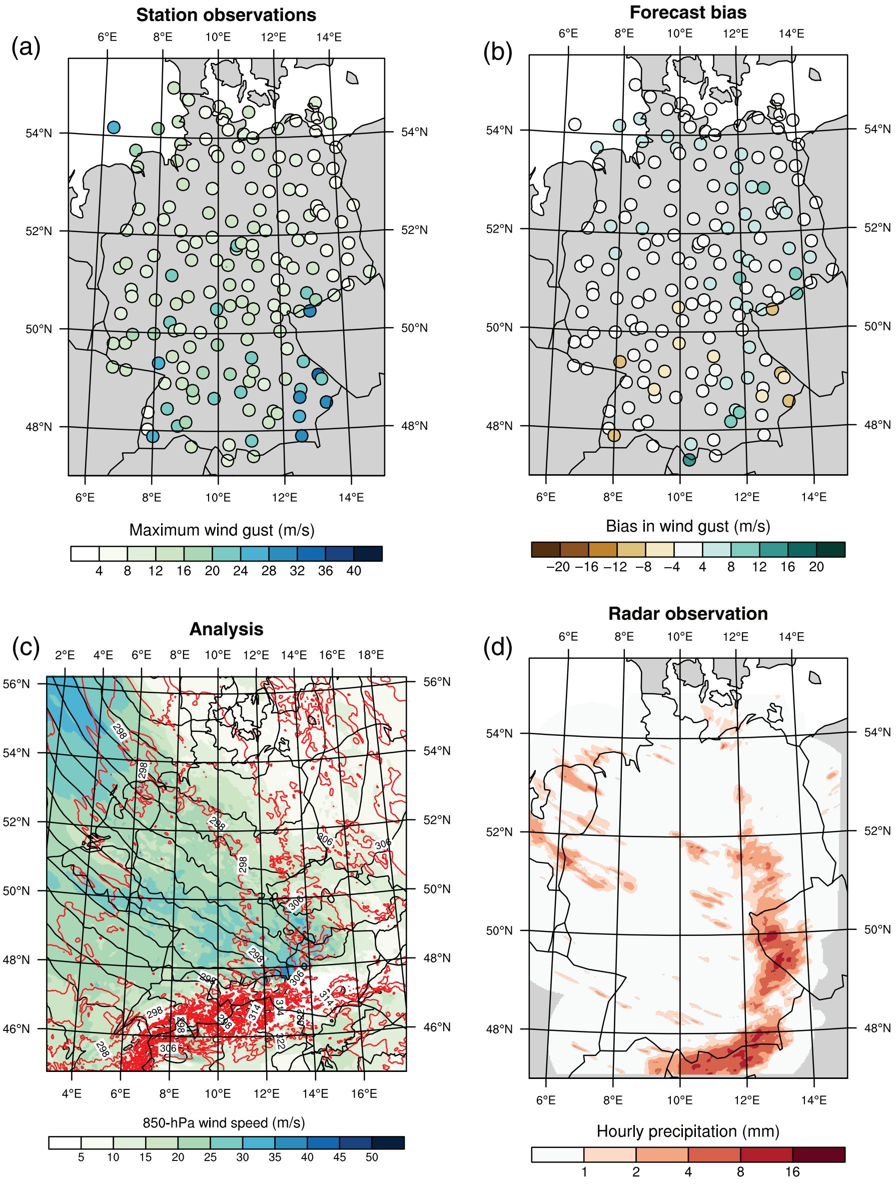

FIGURE 9 As Figure 8 but for storm *Gonzalo* at 2100 UTC on October 21, 2014 ($t + 15$) [Colour figure can be viewed at wileyonlinelibrary.com]

confined to the warm sector of the cyclone over western Germany (Figure 10a,c). As in the case of *Andrea*, the location matches the concept of warm jet. Convection was also present in the cold front, but did not impact the gusts clearly (Figure 10d). The ensemble mean mainly exhibits negative bias in the warm sector at that time, that is, it underestimates strong gusts related to the warm jet (Figure 10b). However, the overestimation of gusts in other regions—in general and at a few specific stations located behind the cold front and over southeastern Germany in particular—results in positive bias overall. Large positive and negative bias is further found at mountain and coastal stations. Statistical post-processing succeeds in reducing such local biases, which are systematic in the dataset, and decreases the CRPS overall (Figure 5b). The maximum intensity of storm *Niklas* occurred later in the afternoon and strong gusts were widespread, related to the warm jet that was still present over southern Germany, a cold jet that was arriving over northern Germany, and, in between, convective showers behind the cold front (see Figure S11 in File S1). In contrast to the previous situation, the bias of the ensemble mean does not exhibit a clear pattern and the CRPS remains relatively low at that time. However, statistical post-processing does not improve the forecast (Figure 5a). This illustrates that, on one hand, strong gusts are not necessarily affected by systematic biases, but, on the other hand, a certain level of random errors remains inherent to strong gusts.

Beyond the four cases detailed here, all other six selected storms involve a warm jet and most of them involve a cold jet in the formation of wind gusts. Most also show frontal or post-frontal convection and storm *Felix* further includes a convection line embedded in the cold front but without clear impact on predictability. These features thus appear typical of severe storms over Germany. In contrast, none of them exhibits signs of a possible sting jet and *Christian* remains an exception. Furthermore, among the other six storms, only

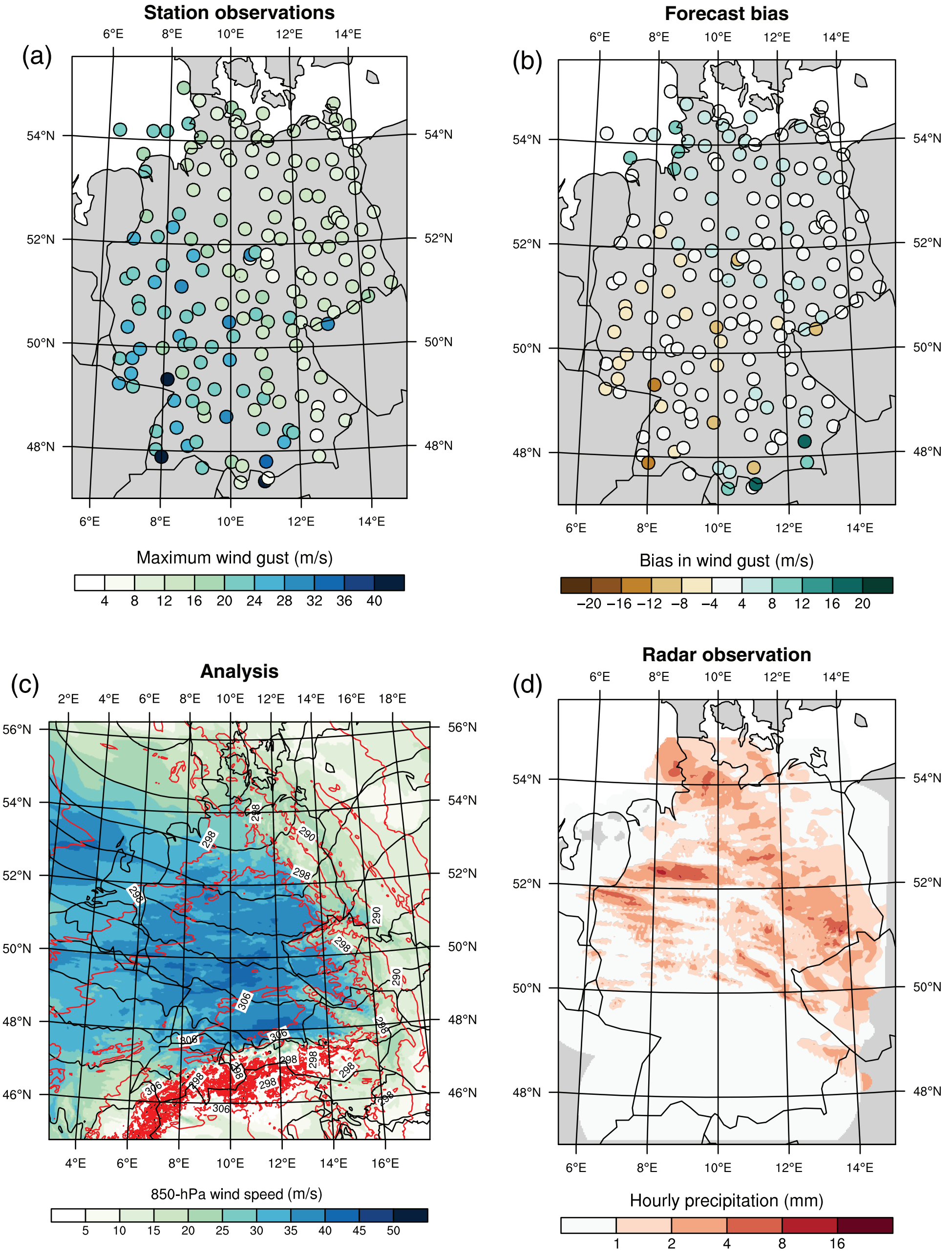

FIGURE 10 As Figure 8 but for storm *Niklas* at 0600 UTC on March 31, 2015 ($t + 6$) [Colour figure can be viewed at wileyonlinelibrary.com]

Joachim tracks across Germany, which makes *Susanna* an unusual case. This suggests that the storms detailed above are rather rare and that their uncharacteristic forecast errors are due to specific dynamics.

4 | CONCLUSIONS AND PERSPECTIVES

A novel six-year dataset of convection-permitting ensemble forecasts is exploited to investigate the predictability of wind gusts in winter storms over Germany. The dataset presents multiple advantages: the high resolution captures mesoscale features that are not resolved by global ensemble forecasts, while the long period both contains several cases of intense storms and sufficient data for robust statistics.

Statistical post-processing substantially improves ensemble forecasts of wind gusts in the whole dataset for all years, all seasons, and all lead times. While the raw ensemble is clearly underdispersive, especially at short lead times, it becomes much better calibrated after post-processing and the ensemble spread matches the magnitude of the RMSE. Compared with a global post-processing model encompassing all stations, a local model trained at each station individually improves the forecasts further by reducing systematic local errors, in winter in particular. However, improvements relative to the raw ensemble are generally smaller during 10 selected winter storms. For instance, wind gusts are relatively well predicted during the time of maximum intensity of the most severe storm of the dataset—*Niklas*, on March 31, 2015—but are not improved by post-processing.

Case studies reveal that, for a few storms with uncharacteristic forecast errors, post-processing can even worsen the ensemble forecasts. The results presented indicate that the ensemble forecast errors—and thereby the appropriateness of specific post-processing models—depend strongly on mesoscale structures and corresponding wind-gust generation mechanisms. The results thus call for the development of physically based post-processing approaches that account for the dependence of misrepresentations of wind gusts on weather regimes. Analog- and similarity-based model estimation approaches proposed by Junk *et al.* (2015) and Lerch and Baran (2017) provide natural first steps in this direction.

In particular, two storms involving frontal convection exhibit systematic overestimation of gusts. In the case of storm *Andrea* on January 5, 2012, the observed drop in gust intensity behind the cold front is missed by forecasts, which suggests deficiencies in the representation of vertical stabilization due to the passage of convection. In the case of storm *Gonzalo* on October 21, 2014, strong gusts formed by the downward mixing of momentum from higher levels are not well captured, which points to the difficult representation of convective gusts that are represented partly by explicit dynamics and partly by the gust parametrization. Although the 2.8 km grid spacing of COSMO-DE-EPS allows representation of convection lines that would not be captured by coarser model forecasts (Ludwig *et al.*, 2015), finer resolution still may be required to resolve convective dynamics fully. This argues for extending pioneering large-eddy simulations over large domains (such as Heinze *et al.*, 2017) to case studies of winter storms in order to understand better the contributions of turbulence and convection to the formation of wind gusts. Model studies can be assessed further and complemented by high-resolution, high-frequency wind observations from Doppler lidars, which have become available in the past years (Pantillon *et al.*, 2018).

However, the ability of a model to predict turbulent and mesoscale dynamics is controlled in the first place by the representation of the synoptic scale. In COSMO-DE-EPS, the ensemble spread is largely inherited from the four driving global models and leads to high forecast uncertainty in the case of the small, fast-moving cyclone *Susanna* on February 9, 2016. This may appear surprising at short lead times of less than one day, but it emphasizes the difficult forecast of the track and intensity of certain storms (Pantillon *et al.*, 2017). Similarly, the representation of the synoptic scale appears responsible for the underestimation of extreme gusts during the passage of the rare storm *Christian* on October 28, 2013 involving a possible sting jet, although the northward shift in the location of strong winds may be due to the problematic track of the storm at the edge of the model domain. These issues may be solved in the current operational version of COSMO-DE-EPS, which is now downscaled from the global ICON-EPS only and the domain size of which has just been increased. However, careful investigation of case studies will be necessary to investigate whether this accounts correctly for the synoptic-scale uncertainty. The multimodel approach has been proved useful in regions where convection-permitting EPSs overlap (Beck *et al.*, 2016) and, with the ongoing increase in domain size of operational models run by national weather services, it may present increased potential for forecasts of extreme events such as winter storms in the future.

ACKNOWLEDGEMENTS

The authors thank Reinhold Hess, David Hansmeier, Susanne Theis, and the Deutscher Wetterdienst (DWD) for providing model data and observations, as well as Robert Redl for pre-processing data. The authors also acknowledge two anonymous reviewers for their comments, which helped to improve the manuscript. Daily maximum wind gusts were downloaded from the DWD Climate Data Center at ftp://ftp-cdc.dwd.de/pub/CDC/, while the AVHRR satellite observation was taken from the NERC Satellite Receiving Station, Dundee University, Scotland at http://www.sat.dundee.ac.uk/. The research leading to these results has been done within the subprojects C5 “Forecast uncertainty for peak surface gusts associated with European cold-season cyclones” and C7 “Statistical postprocessing and stochastic physics for ensemble predictions” of the Transregional Collaborative Research Center SFB / TRR 165 “Waves to Weather”, funded by the German Research Foundation (DFG).

REFERENCES

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., Reinhardt, T., Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M. and Reinhardt, T. (2011) Operational convective-scale numerical weather prediction with the COSMO model: description and Sensitivities. *Monthly Weather Review*, 139(12), 3887–3905. https://doi.org/10.1175/MWR-D-10-05013.1.

Baran, S. and Lerch, S. (2015) Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141, 2289–2299. https://doi.org/10.1002/qj.2521.

Barthlott, C., Mühr, B. and Hoose, C. (2017) Sensitivity of the 2014 Pentecost storms over Germany to different model grids and microphysics schemes. *Quarterly Journal of the Royal Meteorological Society*, 143(704), 1485–1503. https://doi.org/10.1002/qj.3019.

Beck, J., Bouttier, F., Wiegand, L., Gebhardt, C., Eagle, C. and Roberts, N. (2016) Development and verification of two convection-allowing multi-model ensembles over Western Europe. *Quarterly Journal of the Royal Meteorological Society*, 142, 2808–2826. https://doi.org/10.1002/qj.2870.

Born, K., Ludwig, P. and Pinto, J.G. (2012) Wind gust estimation for Mid-European winter storms: towards a probabilistic view. *Tellus A*, 64, 1. https://doi.org/10.3402/tellusa.v64i0.17471.

Brasseur, O. (2001) Development and application of a physical approach to estimating wind gusts. *Monthly Weather Review*, 129(1), 5–25. https://doi.org/10.1175/1520-0493(2001)129<0005:DAAOAP>2.0.CO;2.

Brier, G.W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1–3. https://doi.org/1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Browning, K.A. (2004) The sting at the end of the tail: damaging winds associated with extratropical cyclones. *Quarterly Journal of the Royal Meteorological Society*, 130(597), 375–399. https://doi.org/10.1256/qj.02.143.

Browning, K.A., Panagi, P. and Vaughan, G. (1998) Analysis of an ex-tropical cyclone after its reintensification as a warm-core extratropical cyclone.

Quarterly Journal of the Royal Meteorological Society, 124, 2329–2356. https://doi.org/10.1002/qj.49712455108.

Browning, K.A., Smart, D.J., Clark, M.R. and Illingworth, A.J. (2015) The role of evaporating showers in the transfer of sting-jet momentum to the surface. *Quarterly Journal of the Royal Meteorological Society*, 141, 2956–2971. https://doi.org/10.1002/qj.2581.

Buizza, R. and Hollingsworth, A. (2002) Storm prediction over Europe using the ECMWF ensemble prediction system. *Meteorological Applications*, 9, 289–305. https://doi.org/10.1017/S1350482702003031.

Catto, J.L. (2016) Extratropical cyclone classification and its use in climate studies. *Reviews of Geophysics*, 54, 486–520. https://doi.org/10.1002/2016RG000519.

Coronel, B., Ricard, D., Rivière, G. and Arbogast, P. (2016) Cold-conveyor-belt jet, sting jet and slantwise circulations in idealized simulations of extratropical cyclones. *Quarterly Journal of the Royal Meteorological Society*, 142, 1781–1796. https://doi.org/10.1002/qj.2775.

Dabernig, M., Mayr, G.J., Messner, J.W. and Zeileis, A. (2017) Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*, 143, 909–916. https://doi.org/10.1002/qj.2975.

Dangendorf, S., Arns, A., Pinto, J.G., Ludwig, P. and Jensen, J. (2016) The exceptional influence of storm ‘Xaver’ on design water levels in the German Bight. *Environmental Research Letters*, 11(5), 054 001. https://doi.org/10.1088/1748-9326/11/5/054001.

Earl, N., Dorling, S., Starks, M. and Finch, R. (2017) Subsynoptic-scale features associated with extreme surface gusts in UK extratropical cyclone events. *Geophysical Research Letters*, 44, 3932–3940. https://doi.org/10.1002/2017GL073124.

Einzmann, K., Immitzer, M., Böck, S., Bauer, O., Schmitt, A. and Atzberger, C. (2017) Windthrow detection in European forests with very high-resolution optical data. *Forests*, 8, 21. https://doi.org/10.3390/f8010021.

Fink, A.H., Brücher, T., Ermert, V., Krüger, A. and Pinto, J.G. (2009) The European storm *Kyrill* in January 2007: synoptic evolution, meteorological impacts and some considerations with respect to climate change. *Natural Hazards and Earth System Sciences*, 9, 405–423. https://doi.org/10.5194/nhess-9-405-2009.

Friederichs, P., Wahl, S. and Buschow, S. (2018) Post-processing for extreme events. In: Vannitsem, S., Wilks, D.S. and Messner, J. (Eds.) *Statistical Postprocessing of Ensemble Forecasts*. Amsterdam, Netherlands: Elsevier, pp. 127–154.

Gatzen, C., Púčik, T. and Ryva, D. (2011) Two cold-season derechoes in Europe. *Atmospheric Research*, 100(4), 740–748. https://doi.org/10.1016/j.atmosres.2010.11.015.

Gebhardt, C., Theis, S., Krahe, P. and Renner, V. (2008) Experimental ensemble forecasts of precipitation based on a convection-resolving model. *Atmospheric Science Letters*, 9, 67–72. https://doi.org/10.1002/asl.177.

Gesamtverband der Deutschen Versicherungswirtschaft (2017) *Sachversicherung: Die stärksten Sturm- und Hagelereignisse in Zahlen 1997–2016*. Serviceteil zum Naturgefahrenreport 18. Available at: https://www.gdv.de/resource/blob/11664/e45cf20992e55f221adfcc2b3ef2723b/online-serviceteil-zum-naturgefahrenreport-2017-data.pdf [Accessed 20th April 2018].

Gneiting, T. and Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102, 359–378. https://doi.org/10.1198/016214506000001437.

Gneiting, T., Raftery, A.E., Westveld, A.H.I. and Goldman, T. (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133, 1098–1118. 10.1175/MWR2904.1.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B*, 69, 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Greybush, S.J., Saslo, S., Grumm, R., Greybush, S.J., Saslo, S. and Grumm, R. (2017) Assessing the ensemble predictability of precipitation forecasts for the January 2015 and 2016 east coast winter storms. *Weather and Forecasting*, 32, 1057–1078. https://doi.org/10.1175/WAF-D-16-0153.1.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W. (2017) The Met Office convective-scale ensemble, MOGREPSUK. *Quarterly Journal of the Royal Meteorological Society*, 143, 2846–2861. https://doi.org/10.1002/qj.3135.

Hamill, T.M. (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129, 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hart, N.C.G., Gray, S.L., Clark, P.A., Hart, N.C.G., Gray, S.L. and Clark, P.A. (2017) Sting-jet windstorms over the North Atlantic: climatology and contribution to extreme wind risk. *Journal of Climate*, 30, 5455–5471. https://doi.org/10.1175/JCLI-D-16-0791.1.

Heinze, R., Dipankar, A., Henken, C., Moseley, C., Sourdeval, O., Trömel, S., Xie, X., Adamidis, P., Ament, F., Baars, H., Barthlott, C., Behrendt, A., Blahak, U., Bley, S., Brdar, S., Brueck, M., Crewell, S., Deneke, H., Di Girolamo, P., Evaristo, R., Fischer, J., Frank, C., Friederichs, P., Göcke, T., Gorges, K., Hande, L., Hanke, M., Hansen, A., Hege, H.C., Hoose, C., Jahns, T., Kalthoff, N., Klocke, D., Kneifel, S., Knippertz, P., Kuhn, A., van Laar, T., Macke, A., Maurer, V., Mayer, B., Meyer, C., Muppa, S., Neggers, R., Orlandi, E., Pantillon, F., Pospichal, B., Röber, N., Scheck, L., Seifert, A., Seifert, P., Senf, F., Siligam, P., Simmer, C., Steinke, S., Stevens, B., Wapler, K., Weniger, M., Wulfmeyer, V., Zängl, G., Zhang, D. and Quaas, J. (2017) Large-eddy simulations over Germany using ICON: a comprehensive evaluation. *Quarterly Journal of the Royal Meteorological Society*, 143, 69–100. https://doi.org/10.1002/qj.2947.

Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hewson, T.D. and Neu, U. (2015) Cyclones, windstorms and the IMILAST project. *Tellus A*, 6, 1–33. https://doi.org/10.3402/tellusa.v67.27128.

Jordan, A., Krüger, F. and Lerch, S. (2017) *Evaluating probabilistic forecasts with the R package scoringRules*. Preprint. https://arxiv.org/abs/1709.04743.

Junk, C., Delle Monache, L. and Alessandrini, S. (2015) Analog-based ensemble model output statistics. *Monthly Weather Review*, 143, 2909–2917. https://doi.org/10.1175/MWR-D-15-0095.1.

Keil, C., Heinlein, F. and Craig, G.C. (2014) The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quarterly Journal of the Royal Meteorological Society*, 140, 480–490. https://doi.org/10.1002/qj.2143.

Klawa, M. and Ulbrich, U. (2003) A model for the estimation of storm losses and the identification of severe winter storms in Germany. *Natural Hazards and Earth System Sciences*, 3, 725–732. https://doi.org/10.5194/nhess-3-725-2003.

Layer, M. and Colle, B.A. (2015) Climatology and ensemble predictions of nonconvective high wind events in the New York City metropolitan region. *Weather and Forecasting*, 30, 270–294. https://doi.org/10.1175/WAF-D-14-00057.1.

Lerch, S. and Baran, S. (2017) Similarity-based semilocal estimation of post-processing models. *The Journal of the Royal Statistical Society, Series C*, 66, 29–51. https://doi.org/10.1111/rssc.12153.

Lerch, S. and Thorarinsdottir, T.L. (2013) Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A*, 65, 206. https://doi.org/10.3402/tellusa.v65i0.21206.

Ludwig, P., Pinto, J.G., Hoepp, S.A., Fink, A.H. and Gray, S.L. (2015) Secondary cyclogenesis along an occluded front leading to damaging wind gusts: windstorm Kyrill, January 2007. *Monthly Weather Review*, 143, 1417–1437. https://doi.org/10.1175/MWR-D-14-00304.1.

Martínez-Alvarado, O., Baker, L.H., Gray, S.L., Methven, J. and Plant, R.S. (2014) Distinguishing the cold conveyor belt and sting jet airstreams in an intense extratropical cyclone. *Monthly Weather Review*, 142, 2571–2595. https://doi.org/10.1175/MWR-D-13-00348.1.

Mass, C. and Dotson, B. (2010) Major extratropical cyclones of the Northwest United States: historical review, climatology, and synoptic environment. *Monthly Weather Review*, 138, 2499–2527. https://doi.org/10.1175/2010MWR3213.1.

Matheson, J.E. and Winkler, R.L. (1976) Scoring rules for continuous probability distributions. *Management Science*, 22, 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087.

Mathias, L., Ermert, V., Kelemen, F.D., Ludwig, P. and Pinto, J.G. (2017) Synoptic analysis and hindcast of an intense bow echo in Western Europe: the 9 June 2014 storm. *Weather and Forecasting*, 32(3), 1121–1141. https://doi.org/10.1175/WAF-D-16-0192.1.

Messner, J.W., Mayr, G.J., Zeileis, A. and Wilks, D.S. (2014) Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, 142, 448–456. https://doi.org/10.1175/MWR-D-13-00271.1.

Oesting, M., Schlather, M. and Friederichs, P. (2017) Statistical post-processing of forecasts for extremes using bivariate Brown–Resnick processes with an application to wind gusts. *Extremes*, 20, 309–332. https://doi.org/10.1007/s10687-016-0277-x.

Panofsky, H.A., Tennekes, H., Lenschow, D.H. and Wyngaard, J.C. (1977) The characteristics of turbulent velocity components in the surface layer under convective conditions. *Boundary-Layer Meteorology*, 11, 355–361. https://doi.org/10.1007/BF02186086.

Pantillon, F., Knippertz, P. and Corsmeier, U. (2017) Revisiting the synoptic-scale predictability of severe European winter storms using ECMWF ensemble reforecasts. *Natural Hazards and Earth System Sciences*, 17, 1795–1810. https://doi.org/10.5194/nhess-17-1795-2017.

Pantillon, F., Wieser, A., Adler, B., Corsmeier, U. and Knippertz, P. (2018) Overview and first results of the Wind and Storms Experiment (WASTEX): a field campaign to observe the formation of gusts using a Doppler lidar. *Advances in Science and Research*, 15, 91–97. https://doi.org/10.5194/asr-15-91-2018.

Peralta, C., Ben Bouallègue, Z., Theis, S.E., Gebhardt, C. and Buchhold, M. (2012) Accounting for initial condition uncertainties in COSMO-DE-EPS. *Journal of Geophysical Research, Atmospheres*, 117(D7), D07108. https://doi.org/10.1029/2011JD016581.

Pinto, J.G., Gómara, I., Masato, G., Dacre, H.F., Woollings, T. and Caballero, R. (2014) Large-scale dynamics associated with clustering of extratropical cyclones affecting Western Europe. *Journal of Geophysical Research, Atmospheres*, 119(24), 13704–13719. https://doi.org/10.1002/2014JD022305.

Prahl, B.F., Rybski, D., Burghoff, O. and Kropp, J.P. (2015) Comparison of storm damage functions and their performance. *Natural Hazards and Earth System Sciences*, 15, 769–788. https://doi.org/10.5194/nhess-15-769-2015.

Rasp, S., Selz, T. and Craig, G.C. (2018) Variability and clustering of mid-latitude summertime convection: testing the Craig and Cohen theory in a convection-permitting ensemble with stochastic boundary layer perturbations. *Journal of the Atmospheric Sciences*, 75, 691–706. https://doi.org/10.1175/JAS-D-17-0258.1.

Raveh-Rubin, S. and Wernli, H. (2016) Large-scale wind and precipitation extremes in the Mediterranean: dynamical aspects of five selected cyclone events. *Quarterly Journal of the Royal Meteorological Society*, 142, 3097–3114. https://doi.org/10.1002/qj.2891.

Raynaud, L. and Bouttier, F. (2017) The impact of horizontal resolution and ensemble size for convectivescale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143, 3037–3047. https://doi.org/10.1002/qj.3159.

Scheuerer, M. and Möller, D. (2015) Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *The Annals of Applied Statistics*, 9, 1328–1349. https://doi.org/https://doi.org/10.1214/15-AOAS843.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A. and Potthast, R. (2016) Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, 1453–1472. https://doi.org/10.1002/qj.2748.

Schulz, J.P. (2008) Revision of the turbulent gust diagnostics in the COSMO model. *COSMO Newsletters*, 8, 17–22. http://www2.cosmo-model.org/content/model/documentation/newsLetters/newsLetter08/cnl8_schulz.pdf [Accessed 11 September 2018].

Schwartz, C.S., Romine, G.S., Sobash, R.A., Fossell, K.R. and Weisman, M.L. (2015) NCAR's experimental real-time convection-allowing ensemble prediction system. *Weather and Forecasting*, 30, 1645–1654. https://doi.org/10.1175/WAF-D-15-0103.1.

Staid, A., Pinson, P. and Guikema, S.D. (2015) Probabilistic maximum-value wind prediction for offshore environments. *Wind Energy*, 18, 1725–1738. https://doi.org/https://doi.org/10.1002/we.1787.

Thorarinsdottir, T.L. and Johnson, M.S. (2012) Probabilistic wind gust forecasting using nonhomogeneous Gaussian regression. *Monthly Weather Review*, 140, 889–897. https://doi.org/10.1175/MWR-D-11-00075.1.

Weijenborg, C., Friederichs, P. and Hense, A. (2015) Organisation of potential vorticity on the mesoscale during deep moist convection. *Tellus A*, 67, 25–705. https://doi.org/10.3402/tellusa.v67.25705.

Wernli, H., Dirren, S., Liniger, M.A. and Zillig, M. (2002) Dynamical aspects of the life cycle of the winter storm 'Lothar' (24–26 December 1999). *Quarterly Journal of the Royal Meteorological Society*, 128, 405–429. https://doi.org/10.1256/003590002321042036.

Wilks, D.S. (2011) *Statistical Methods in the Atmospheric Sciences*, 3rd edition. Oxford, UK: Elsevier Academic Press.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Pantillon F, Lerch S, Knippertz P, Corsmeier U. Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble. *Q J R Meteorol Soc.* 2018;144:1864–1881. https://doi.org/10.1002/qj.3380