Received: 27 May 2021 | Revised: 27 April 2022 | Accepted: 13 May 2022 | Published on: 30 June 2022

DOI: 10.1002/qj.4311

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Probability forecasts of ice accretion on wind turbines derived from multiphysics and neighbourhood ensembles

Lukas Strauss[1,2] | Stefano Serafin[1] | Manfred Dorninger[1]

[1]University of Vienna, Department of Meteorology and Geophysics, Vienna, Austria

[2]MeteoServe Wetterdienst GmbH, Vienna, Austria

Correspondence
Lukas Strauss and Manfred Dorninger, University of Vienna, Department of Meteorology and Geophysics, Josef-Holaubek Platz 2, 1090 Vienna, Austria.
Email: lukas.strauss@meteoserve.at; manfred.dorninger@univie.ac.at

Funding information
Klima- und Energiefonds, Grant/Award Number: 853575

Abstract

This study explores the potential of a multiphysics regional ensemble prediction system to improve forecasts of wind turbine icing, examining several error-representation schemes to capture the forecasting uncertainties of the icing process. An 11-member multiphysics ensemble based on the Weather Research and Forecasting (WRF) model is run for two winter periods over Europe. Regional verification of surface variables shows that parametrization diversity makes the multiphysics ensemble less underdispersive compared with the European Centre for Medium-Range Weather Forecasts (ECMWF) global ensemble, without deteriorating the overall forecast accuracy significantly, in particular at forecast ranges below 36 hr. Probability forecasts of active ice growth in the day-ahead time range (12–36 hr) are derived for two wind farms on hilly terrain in Central Europe and their skill is assessed in terms of relative operating characteristics, reliability, and potential economic value (PEV). Probability forecasts enhance the maximum PEV significantly, but the improvement of the multiphysics ensemble seems modest compared with a simple neighbourhood ensemble approach. Icing forecasts are affected by a considerable degree of overconfidence, meaning that forecast probabilities cannot be used at face value, but require calibration for users to draw benefit from them. The multiphysics ensemble fares slightly better in this regard; however, results point to persistent ensemble underdispersiveness and yet underrepresented forecast uncertainty. Overall, findings show that a large portion of the gain in skill through the use of probabilistic icing forecasts is obtained with a computationally cheap neighbourhood method, a technique easily accessible to forecast users without complex ensemble prediction systems.

KEYWORDS

ensemble prediction, icing, multiphysics, neighbourhood method, potential economic value, probability forecasts, verification, wind energy

1 | INTRODUCTION

Wind power plants are increasingly common at high latitudes and higher elevations. The total installed wind capacity in cold climates (Laakso *et al.*, 2009) across North America, Europe, and Asia is estimated at 156 GW, or 22% of global onshore capacity in 2020, and is projected to increase to 223 GW by the end of 2025 (Karlsson, 2021). While these regions have favourable wind climatologies, they are also prone to harsh weather conditions, such as icing events or periods with temperatures below the operational limits of wind turbines.

Three different atmospheric processes are known to be associated with wind turbine icing: (i) freezing of supercooled cloud droplets upon contact with rotor blades, (ii) freezing precipitation on surfaces below the freezing point, and (iii) wet-snow accretion. Turbine icing results in power production losses, due to degraded aerodynamic properties of turbine blades, increased downtimes of wind turbines for risk of ice fall and ice throw (Krenn *et al.*, 2018), or even damage to the turbine structure. Day-to-day icing forecasts can potentially inform wind-farm operators and traders and help them start anti-icing procedures such as blade heating, or alternatively account for production losses on day-ahead or intraday energy markets.

Research on icing models and forecasts has seen a sharp rise in the past 15 years. Early on, researchers tested various sets of boundary-layer and microphysics schemes to determine the sensitivity of icing forecasts or to select the best schemes to run operationally (Thompson *et al.*, 2009; Nygaard *et al.*, 2011; Davis *et al.*, 2014). From such studies, the Thompson microphysics scheme and the Mellor–Yamada–Nakanishi–Niino planetary boundary layer (PBL) scheme (Nakanishi and Niino, 2006; Thompson *et al.*, 2008) emerged as a promising combination, as the Thompson scheme was shown to model supercooled water accurately even at temperatures well below the freezing point. However, other combinations of physical schemes were shown to be only marginally worse, and considerable case-to-case variability of the skill of any model setup was observed. The latter is related to significant uncertainties as to the exact location, severity, timing, and duration of icing events. Uncertainty and skill variability are particularly pronounced because the turbine icing process is related to several atmospheric parameters, such as temperature, wind speed, and liquid water content, as well as to the properties of airflow around turbine blades, each contributing their uncertainty.

Recently, this aspect has been investigated systematically. Molinder *et al.* (2018) studied the improvement in the spread-error relationship of forecasts of icing-related power production losses for several events at wind farms in Sweden, including uncertainties in numerical weather prediction (NWP) model forecasts due to initial and boundary condition (IBC) error and spatial representation. Molinder *et al.* (2019) accounted for uncertainties in empirical icing models by perturbing parameters such as the median volume diameter of liquid droplets stochastically. Davis *et al.* (2016), Scher and Molinder (2019), and Molinder *et al.* (2021) used machine-learning techniques to incorporate both current observational values and NWP variables in the prediction of icing-related production losses. In all studies, encouraging gradual improvements were achieved in terms of, for instance, the mean absolute error of production loss and the ability to forecast losses greater than 25% of turbine installed capacity. However, forecasts were found to be underdispersive, that is, they still underrepresented forecast uncertainty, calling for further improvements in error representation. In the following section, error representation schemes are reviewed briefly in the context of icing and its predictability.

1.1 | Predictability of icing and error representation schemes

The predictability challenges involved in icing forecasts bear similarities to those related to deep moist convection. In midlatitude climates, both phenomena are infrequent, although they occur regularly during certain periods of the year. They both depend on the simultaneous occurrence of necessary but insufficient conditions (uplift beyond the level of free convection in one case, a liquid water cloud at subfreezing temperature in the other). Finally, their forecasts are critically dependent on microphysical parametrization schemes, and their occurrence or not may differ over very short distances, making site-specific forecasts challenging. For these reasons, progress in accounting for model uncertainty (e.g., by means of multiphysics ensembles) and location errors (e.g., by means of neighbourhood methods) in probability forecasts, largely achieved in the context of precipitation forecasting, is relevant to the icing problem.

Multiphysics ensembles (MPEs) have been suggested as a pragmatic approach to obtaining probablistic forecasts with large ensemble spread at short forecast ranges (Stensrud *et al.*, 2000), similar to poor-man ensemble prediction systems (Arribas *et al.*, 2005). MPEs are built by running several integrations of the same weather prediction model, each adopting a different combination of parametrization schemes. This strategy accounts for model error by using different approximations of the best possible model, as determined by different representations of subgrid-scale phenomena. Due to the different parametrization schemes, the members of MPEs differ in climatology and systematic errors. This conflicts with the principal aim of probabilistic forecasting, which should be the sampling of random forecast errors. However, the inclusion of physical configurations leading to opposite biases in the forecasts can cause error compensation, increasing ensemble spread without altering the ensemble mean error significantly (Berner *et al.*, 2015). In the context of deep moist convection forecasts, MPEs have been shown to create ensemble variance faster than a traditional IBC ensemble, and to have greater skill under weak synoptic forcing, where the ability of the models to trigger and maintain convection depends more strongly on parametrized small-scale dynamics (Stensrud *et al.*, 2000). Even in the MPE approach, the maintenance of large spread over long forecast ranges or in small spatial domains (where the propagation of lateral boundary conditions into the computational domain is most pronounced) requires using perturbed IBCs (Clark *et al.*, 2008), for example, forcing from different members of a global ensemble prediction system (EPS).

MPEs have been compared with other kinds of model error representation (e.g., stochastically perturbed parametrization tendency scheme (SPPT): Buizza *et al.*, 2007; stochastic kinetic energy backscatter scheme (SKEBS): Berner *et al.*, 2009; stochastically perturbed parameterization (SPP): Hacker *et al.*, 2011b) in a number of previous studies (Berner *et al.*, 2011; Hacker *et al.*, 2011a; Romine *et al.*, 2014; Berner *et al.*, 2015; Duda *et al.*, 2017; Jankov *et al.*, 2017). Findings generally demonstrate that all model error schemes improve the skill of probability forecasts in general, in particular in the boundary layer and when applied in concert (e.g., application of both SPP and SPPT in an MPE, or SKEBS in an MPE). These results hold true even after postprocessing forecasts with independent debiasing of each ensemble member and ensemble calibration with variance inflation, which constrains forecasts to have the same variance as the observations. It follows that forecast improvements achieved through model error representations are greater than those that could be obtained by mere postprocessing (Berner *et al.*, 2015). Skill differences between MPEs and stochastic model error representations (SPPT, SPP, SKEBS) are rarely statistically significant, and the latter yield spread/error ratios comparable with MPEs only when combined (Jankov *et al.*, 2017).

Besides complex error generation schemes such as the ones above, neighbourhood approaches have been put forward by various authors as a pragmatic, inexpensive method to account for basic spatial and temporal forecast uncertainties, in particular for convection-resolving (< 4 km grid spacing) models (see for instance Schwartz and Sobash 2017, for a review).

1.2 | Aims of this work

In the winters of 2016/2017 and 2017/2018, a field campaign to collect observational evidence of icing events was conducted at a wind farm in Germany (Ellern, hereafter EL) in the framework of the ICE CONTROL project (Strauss *et al.*, 2020, hereafter SSD20). Additionally, measurement data from another wind farm in the Czech Republic (Kryštofovy Hamry, hereafter KH) were obtained with courtesy of the turbine manufacturing company ENERCON GmbH.

Based on this dataset, SSD20 presented a two-winter verification study of the skill and potential economic value of several variants of deterministic icing forecasts, based on both global and regional NWP models. In line with the ICE CONTROL objectives, the focus was on phases of active ice accumulation, which have been associated with the strongest production losses (e.g., Bernstein *et al.*, 2012; Bergström *et al.*, 2013) and are the sensitive phases during which preventive turbine anti-icing systems can make a difference. Results by SSD20 showed that 2.5-km grid-spacing Weather Research and Forecasting (WRF) forecasts are superior at predicting active ice growth relative to coarse-resolution models, measured in terms of potential economic value. However, in absolute terms, the skill of deterministic models in icing predictions is limited, suggesting ample potential for forecast improvements by means of an ensemble prediction system.

Therefore, the aims of this work are the following:

1. to build a regional ensemble model for the short (day-ahead) forecast range and test its performance with respect to the global European Centre for Medium-Range Weather Forecasts (ECMWF) EPS model by means of regional ensemble verification for surface variables;
2. to investigate whether the related improvements in skill can be ported to site-specific probability forecasts of ice growth and prerequisite conditions, such as freezing and humid conditions; and
3. to assess if skill improvements are significant in a statistical sense, compared with less expensive ensemble generation methods.

To this end, we consider both multiphysics ensembles (drawing from a range of setups tested by Thompson *et al.*, 2009; Nygaard *et al.*, 2011; Davis *et al.*, 2014) and neighbourhood ensembles (following the neighbourhood ensemble probability (NEP) method by Schwartz and Sobash, 2017). The choice of these error representation schemes is inspired by the studies presented in Section 1.1 and seeks to extend previous attempts in the icing context (Molinder *et al.*, 2018; 2019; Strauss *et al.*, 2020, hereafter SSD20). Neighbourhood ensembles serve as a term of comparison, which multiphysics ensembles need to beat within statistical significance for them to be attributed added value.

The above research aims are also of practical relevance. Probabilistic forecasting with ensembles is a consolidated practice in the meteorology community, but end users take advantage of ensemble forecast products relatively rarely. From a wind-power perspective, it is highly valuable to prove or disprove, on the basis of sound scientific methodology, the added value of probabilistic icing forecasts for decision making by wind-farm operators and energy traders.

The remainder of this article is structured as follows: In Section 2, the numerical weather prediction and icing models, as well as the icing observations, are described. Section 3 presents methods for verification of regional ensemble forecasts and probability forecasts at wind-farm sites. In Section 4, first the results of verification of ensemble forecasts of surface parameters are discussed, corroborating the choice of model setups. Then two case studies of icing events at a wind farm are presented, followed by the verification of site-specific probability forecasts of icing and related conditions. Conclusions are drawn and routes for future work are outlined in Section 5.

2 | MODELS AND OBSERVATIONS

2.1 | Numerical weather prediction models

The limited-area ensemble used in this study is based on the Advanced Research Weather Research and Forecasting Model (ARW-WRF) model, version 3.9 (Skamarock and Klemp, 2008). Ensemble size is limited to 11 members, which is large enough for the ensemble mean to be approximately independent of the ensemble size (Leith, 1974) and at the same time computationally affordable. Two one-way nested simulation domains covering large parts of western Europe (D01, $\Delta x = 12.5$ km, $\Delta t = 30$ s, 216×180 grid points) and Germany (D02, $\Delta x = 2.5$ km, $\Delta t = 15$ s, 316×256 grid points) are defined (Figure 1). Both domains contain 51 terrain-following vertical levels, with grid spacing ranging from ~20 m below 200 m above ground level (AGL) to ~680 m for the layer between 200 hPa and the model top (100 hPa). Simulations are always initialized at 0000 UTC.

Two instances of the ARW-WRF ensemble are considered. A dynamical downscaling ensemble, WRFDY, is run only on domain 1 for the period November 2016–March 2017 (151 days) over a forecast range of 48 hr. In WRFDY, ensemble members differ only by their IBCs. WRFDY is run over one winter only, merely as a benchmark to evaluate the skill improvements brought by a competing MPE, WRFMP. This is run on both domains for the periods November 2016–March 2017 and November 2017–March 2018 (302 days total) over a forecast range of 60 hr. In WRFMP, ensemble members differ both by

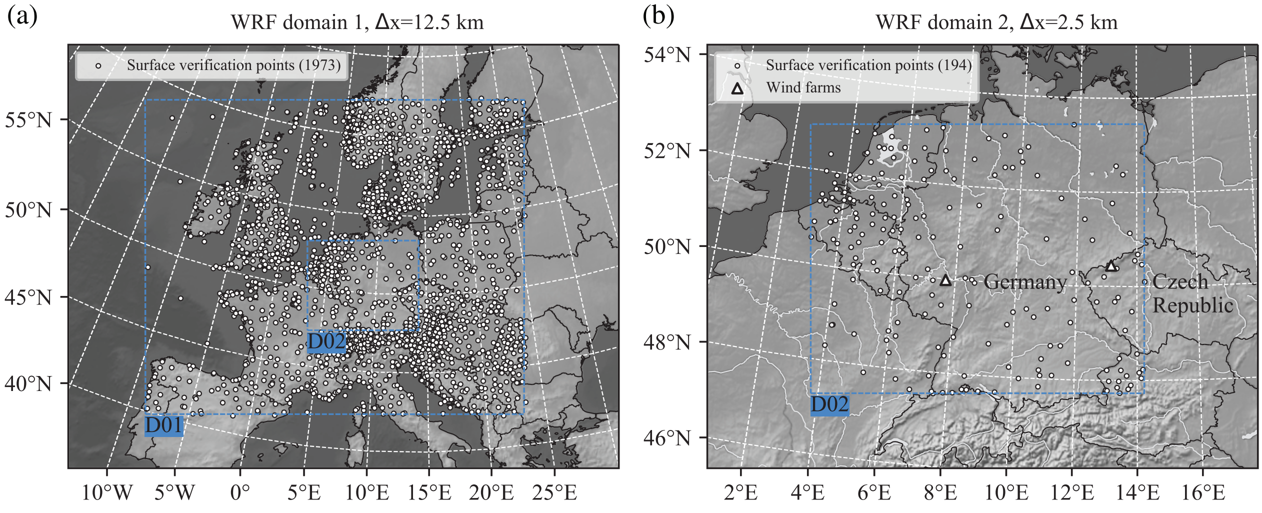

FIGURE 1 (a) Map of the simulation domains D01 (WRFDY and WRFMP ensembles) and D02 (WRFMP only). (b) Detailed map of the WRFMP domain D02 and location of the wind-farm sites Ellern (Germany) and Kryštofovy Hamry (Czech Republic). White bullets represent the surface measurement stations that were used for the verification of temperature, dew-point, and wind-speed forecasts (Section 4.1) [Colour figure can be viewed at wileyonlinelibrary.com]

IBCs and by the configuration of parametrization schemes. The comparison of the verification scores of WRFDY and WRFMP for two-day forecasts in winter 2016–2017 permits assessment of whether or not the multiphysics approach improves the quality of ensemble forecasts significantly (Section 4.1).

IBCs for the WRFDY and WRFMP integrations are interpolated from operational runs of the ECMWF EPS, based on the IFS model (Molteni *et al.*, 1996). The different ensemble sizes (51 members for the ECMWF EPS, 11 members each for the two WRF ensembles) imply that the WRF IBCs must be derived from a subset of ECMWF EPS members. Here, ensemble reduction is performed by randomly sampling 10 out of the 50 perturbed members of the ECMWF EPS. As outlined by Serafin *et al.* (2019), random sampling avoids the ensemble spread reduction associated with the clustering of ensemble members and the selection of a representative member in each cluster. The control member of the WRF ensembles is always initialized with the control member of the ECMWF EPS. WRFMP and WRFDY simulations on the same day are initialized with identical subsets of the ECMWF EPS.

In addition to the WRFDY and WRFMP ensembles, a deterministic run (WRF-IFS) is considered as well, the verification of which was conducted by SSD20. WRF-IFS serves as a term of comparison to quantify the added value of probability forecasts derived from WRFMP (Sections 3.2 and 4.4). It is configured in exactly the same way as the control member of WRFMP and only differs in the IBCs, which are interpolated from the high-resolution deterministic IFS run instead of the EPS control member.

The selection of physical schemes in WRF-IFS, WRFMP and WRFDY is detailed in Table 1, which also includes information on the respective acronyms and literature references. The configuration of parametrization schemes in WRF-IFS and WRFDY is identical and based on SSD20. WRFMP is instead designed to estimate the model error determined by the physical schemes relevant to the simulation of the wind turbine icing process. Because the icing rate is affected primarily by wind speed, temperature, and supercooled liquid water content (Equation (1)), ensemble members are defined by permutations of the parametrization schemes for the surface layer (four options), the boundary layer (four options), and cloud microphysics (three options), as well as the land-use database (two different options) and land-surface model (two options). Besides the physics configuration of the WRFMP control member, which is identical to that

TABLE 1 Physics configurations of WRF deterministic (WRF-IFS), WRF dynamical downscaling (WRFDY), and WRF multiphysics (WRFMP) ensemble members

WRF configuration	ICs & BCs	Land use	Land surface	SL	PBL	Microphysics
WRF-IFS	EC-IFS	MODIS	RUC LSM	MYNN	MYNN-3	T&E
WRFDY	EC-EPS M0 and r.m.	MODIS	RUC LSM	MYNN	MYNN-3	T&E
WRFMP M00	EC-EPS M0	MODIS	RUC LSM	MYNN	MYNN-3	T&E
WRFMP M01	EC-EPS r. m.	CORINE	RUC LSM	MOJ	BL	T
WRFMP M02	EC-EPS r. m.	MODIS	Noah	MO	YSU	M
WRFMP M03	EC-EPS r. m.	CORINE	Noah	MYNN	MYNN-3	M
WRFMP M04	EC-EPS r. m.	MODIS	RUC LSM	MO	YSU	T
WRFMP M05	EC-EPS r. m.	MODIS	Noah	MO	YSU	T&E
WRFMP M06	EC-EPS r. m.	MODIS	Noah	MOJ	BL	M
WRFMP M07	EC-EPS r. m.	MODIS	Noah	QNSE	QNSE	T&E
WRFMP M08	EC-EPS r. m.	CORINE	RUC LSM	MYNN	MYNN-3	T
WRFMP M09	EC-EPS r. m.	CORINE	Noah	MOJ	BL	T&E
WRFMP M10	EC-EPS r. m.	CORINE	Noah	QNSE	QNSE	T

Key to abbreviations: r. m.: *random member* (see Section 2.1), ICs: initial conditions, BCs: boundary conditions, SL: surface layer, PBL: planetary boundary layer. MODIS: 30-s NOAH-modified IGBP-MODIS 20-category database with lakes; CORINE: CORINE land cover inventory 2012 (Copernicus Land Monitoring Service, 2012); RUC: Rapid-update cycle scheme (Smirnova *et al.*, 2016); Noah: Unified NOAH scheme (NCAR Research Applications Laboratory, 2020); MYNN-3: Mellor–Yamada–Nakanishi–Niino third-level TKE scheme (Nakanishi and Niino, 2006); MOJ: Monin–Obukhov–Janjic scheme (Janjić, 1994); MO: Revised Monin–Obukhov scheme (Jiménez *et al.*, 2012); QNSE: Quasi-Normal Scale Elimination scheme (Sukoriansky *et al.*, 2005); BL: Bougeault–Lacarrère scheme (Bougeault and Lacarrere, 1989); YSU: Yonsei University scheme (Hong *et al.*, 2006); T&E: Thompson–Eidhammer “aerosol-aware” scheme (Thompson and Eidhammer, 2014), using GOCART model monthly global climatologies of number concentrations of water- and ice-friendly aerosols (Ginoux *et al.*, 2001; Colarco *et al.*, 2010); T: Thompson scheme (Thompson *et al.*, 2008); M: Morrison scheme (Morrison *et al.*, 2009).

of WRF-IFS and WRFDY, 10 other configurations are defined, whereby most WRFMP members differ from the control run in at least three parametrization options.

The boundary-layer models in WRFMP are chosen so as to represent different turbulence closure approaches, from simple one-dimensional models based on eddy-diffusivity profiles (e.g., YSU) to more complex ones based on a prognostic equation for the turbulent kinetic energy (e.g., MYNN: cf. Cohen *et al.*, 2015, for a review). Surface-layer schemes are coupled pairwise to boundary-layer schemes. Microphysical schemes are all of the bulk type, consider five hydrometeor classes in addition to water vapour, and account for ice-phase cloud processes. They differ in the number of hydrometeor classes modelled with a two-moment approach and whether or not they account for climatological aerosol concentrations. Specifically, only the Thompson and Eidhammer (2014) scheme predicts explicitly the number concentration of cloud droplets and aerosol particles acting as cloud condensation nuclei and ice nuclei.

The configurations of the 10 WRFMP perturbed members were selected from a larger pool of 20 options, which were evaluated over the first winter period (not shown). The root-mean-square deviation (RMSD) between ensemble members was determined for several forecast fields (geopotential height, temperature, wind components, mixing ratios of water vapour, cloud water, and rain water) at pressure levels of 850 and 925 hPa, which are most relevant for wind-resource forecasting at the wind-farm sites considered. Pairs or groups of physics configurations that systematically displayed low RMSD for most forecast fields were considered redundant, and only one element of these pairs/groups was included in WRFMP. This approach is in line with the principal advantage of MPEs of furthering ensemble spread (Section 1.1).

2.2 | Icing model

The Makkonen model (Makkonen, 2000), along with recent modifications (e.g., Nygaard *et al.*, 2013; Davis *et al.*, 2014), is coupled to the output of WRF simulations to produce icing forecasts. The model is a standard tool for the modelling of atmospheric icing on built structures. It assumes the growth rate of ice mass M on a vertically oriented structure with cross-sectional area A to be proportional to the incoming mass flux of liquid water ρVA, where ρ is the liquid water density and V the horizontal wind-speed:

$$\frac{dM}{dt} = \alpha_1 \alpha_2 \alpha_3 \rho V A. \tag{1}$$

The correction factors α_1, α_2, and α_3 are dimensionless and represent, respectively, the collision, sticking, and accretion efficiencies for water droplets. The formulation of the efficiency factors is empirically based and depends on the aerodynamics of the flow around the reference structure. In the Makkonen model, the latter is a vertically oriented cylinder, 3 cm in diameter and 1 m in length. In our implementation, Equation (1) is integrated independently for two of the hydrometeor classes modelled in WRF, namely cloud droplets and rain water. The ice-mass growth rates associated with these two classes are then added. The collision efficiency α_1 is assumed to depend on the median volume diameter of the droplets. For cloud droplets, this is diagnosed from their mixing ratio and number concentration (Thompson *et al.*, 2009, equations (2)–(4)). For rain drops it is set to 100 μm, which is large enough to yield $\alpha_1 \approx 1$ (i.e., all rain drops are assumed to collide with the structure, in accordance with Makkonen, 2000).

The production of icing forecasts occurs in two steps: (i) run the WRF simulation and (ii) compute M by integrating Equation (1) for the available forecast range, using the hourly output of WRF. This two-step procedure allows for the postprocessing of WRF forecasts (e.g., correction of a systematic temperature bias) prior to running the icing model. Ice load and thickness are initialized with their values at +24 hr of the previous icing forecast. Similarly to Davis *et al.* (2014), icing episodes are assumed to end whenever temperature exceeds 1°C for more than 1 hr, leading to complete removal of the accreted ice mass. This criterion might seem crude, but it represents ice shedding from rotor blades reasonably well. In fact, as soon as temperature rises past zero, loss of adhesion to the structure causes the ice to fall off very quickly, even if it is not completely melted. A more complex model is required only if the dominant ice removal process is sublimation, as is typical at higher latitudes (Davis *et al.*, 2014).

2.3 | Observations

The forecasts of near-surface parameters by the ECMWF, WRFDY, and WRFMP ensembles are verified against hourly observations from SYNOP stations in the areas corresponding to D01 (1973 verification points) and D02 (194 verification points) of the WRF simulations. Surface measurements were retrieved from the Integrated Surface Database of the U.S. National Centers for Environmental Information (Smith *et al.*, 2011).[1] No spatial analysis or interpolation of the measured data is

[1] https://www.ncdc.noaa.gov/isd

performed. At each valid forecast time, grid-point values of the numerical forecast are compared with the respective nearest-neighbour observations.

During the ICE CONTROL project, a two-winter field campaign was conducted in 2016/2017 and 2017/2018 at wind farm Ellern (hereafter EL), situated in the hilly Hunsrück Range in Rhineland–Palatinate, Germany (Figure 1b). Instruments were located on a hilltop on the nacelle of an ENERCON E-126 turbine at 780 m above mean sea level (AMSL). A second dataset, for a wind farm located in the Ore Mountains at the border between the Czech Republic and Germany, close to the village of Kryštofovy Hamry (hereafter KH: Figure 1b), was made available by ENERCON GmbH. Measurements at KH were taken on a turbine with hub height of 930 m AMSL and on a nearby mast reaching 900 m AMSL. Table 2 gives the site coordinates.

Turbines at both wind farms and the mast at KH were instrumented with standard meteorological sensors to measure temperature T, wind speed V, and relative humidity RH. In addition, heated webcam systems were installed at the turbine hubs, pointing towards the instrumentation on the nacelles. A camera image analysis by Meteotest provided a categorical assessment of five classes of instrumental icing severity (“light", “light-moderate”, “moderate”, “moderate-heavy", and “heavy”) and three classes of ice growth (“light”, “moderate”, and “strong”). Details of the sensors and the camera image analysis are found in SSD20.

Due to their slightly different elevations and locations in Central Europe, EL and KH exhibit differences in their site climates. Periods of active icing (ice growth class ≥ “light”) occur three times more often at KH (two-winter climatological rate of $s_{\mathrm{KH}} = 15\%$) than at EL ($s_{\mathrm{EL}} = 6\%$). “Light” severities of ice growth dominate the samples at both sites ($> 85\%$ at EL, $> 66\%$ at KH). The predominant icing type at both wind farms is in-cloud icing. The two sites belong to ice classes 2–3 (EL) and 4 (KH), according to the International Energy Agency (IEA) ice classification for wind energy sites (Bredesen *et al.*, 2017; SSD20).

3 | VERIFICATION METHODS

3.1 | Verification scores for ensemble forecasts of continuous variables

To quantify the overall accuracy of a set of probability forecasts of a variable ϕ, we use the average value of the continuous rank probability score (CRPS). For a single forecast–observation pair,

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} [F_{\mathrm{f}}(\phi) - F_{\mathrm{o}}(\phi)]^2 \, d\phi, \tag{2}$$

where F_{f} is the cumulative probability density function (CDF) of the ensemble forecasts and F_{o} that of the observation. Herein, CRPS is computed following Hersbach (2000). CRPS values averaged over a set of forecast–observation pairs are denoted by $\overline{\mathrm{CRPS}}$. To favour comparability, $\overline{\mathrm{CRPS}}$ values are scaled to similar numerical ranges and made nondimensional by normalisation with the standard deviation of observations, σ_{o}.

The dispersion properties of the ensemble forecasts are described using the ratio between the ensemble spread,

$$\mathrm{Var}_{\bar{f}} = \frac{1}{M}\frac{1}{N-1}\sum_{m=1}^{M}\sum_{n=1}^{N}(f_{mn} - \bar{f}_m)^2, \tag{3}$$

and the mean-square error of the ensemble mean,

$$\mathrm{MSE}_{\bar{f}} = \frac{1}{M}\frac{N}{N+1}\sum_{m=1}^{M}(\bar{f}_m - o_m)^2. \tag{4}$$

Here, f denotes the forecasts by individual ensemble members, $\bar{f}$ the ensemble mean forecast, o the observations, N the ensemble size, and M the number of forecast–observation pairs. The factors $1/(N-1)$ and $N/(N+1)$ in Equations (3) and (4) ensure that the consistency between spread and error is evaluated properly for finite-size ensembles (Eckel and Mass, 2005). For a reliable ensemble, the spread-error ratio is equal to 1 (Weigel, 2012, p. 144 ff.). The spread/error ratio is smaller than one when

TABLE 2 Coordinates of site locations, types of wind turbine, and names of instruments used at icing measurement sites during the winter periods November 2016–March 2017 and November 2017–March 2018

Site names	Wind turbines	Coordinates	Hub heights (m AMSL (m AGL))
(EL) Ellern, DE	ENERCON E-126	49.9785° N, 7.6822° E	780 (135)
(KH) Kryštofovy Hamry, CZ	ENERCON E-82	50.4449° N, 13.1446° E	930 (78)
	Measurement mast	50.4433° N, 13.1417° E	900 (61)

AMSL, above mean sea level.

the verifying observations often lie outside the ensemble range, indicating underdispersiveness.

For a visual demonstration of the degree of dispersiveness of ensemble forecasts, we use rank histograms (Hamill, 2001). These represent the frequency distribution of the observation rank, that is, the position of the verifying observation in the ordered set it forms together with the ensemble forecasts. The optimal rank histogram corresponds to a uniform distribution, while U-shaped histograms indicate ensemble underdispersiveness. Following Weigel (2012, p. 146 ff.), we assess the deviation of rank histograms from uniformity using the quantity

$$\tau = \sum_{n=1}^{N+1} \frac{(m_n - e)^2}{e}, \tag{5}$$

where m_n is the number of times the observation has rank n and $e = M/(N+1)$. For a perfectly uniform rank histogram, $m_n = e$ for all values of n, hence $\tau = 0$.

It has been shown that conclusions about spread-error characteristics based on rank histograms can change drastically if observation uncertainty is accounted for. Hacker *et al.* (2011a) demonstrate that U-shaped rank histograms can become nearly flat if the verifying observations are perturbed stochastically according to a predetermined observation-error variance. Because the observation-error variance accounts partially for representativeness error, it is to some extent model-dependent and can be specified in a consistent manner only in the framework of a cycling convective-scale DA system. This is not available in our case. Therefore, in analogy with Hacker *et al.* (2011a), we ignore the effects of observation errors on the verification, because we lack rigorous estimates of the observation-error variance. Our considerations on ensemble (under)dispersiveness are based on spread and error as defined in Equations (3) and (4).

3.2 | Probability forecasts of binary weather events at measurement sites

Three specific weather conditions related to icing serve to define binary weather events at icing measurement sites: (i) freezing temperatures ($T \leq 0°$C), (ii) freezing and humid conditions ($T \leq 0°$C and RH $\geq$ 85%), and (iii) active ice growth (ice growth class ≥ "light"). For the latter type of binary event, ice growth is modelled using the Makkonen model, and the ice growth rate threshold above which a "yes" forecast occurs is set to 2.5 $\times 10^{-4}$ kg$\cdot$hr^{-1}, in line with SSD20. Probability forecasts of these events are derived from WRF forecasts (both WRF-IFS and individual WRFMP members) by applying the respective thresholds and counting the number of ensemble members predicting a given weather condition. All ensemble members are assumed to be equally likely ("frequentist" ensemble interpretation, cf. Weigel, 2012).

Four variants of probability forecasts are considered: (i) WRF-IFS (one member, giving only probabilities of 0 and 1); (ii) WRF-IFS NN (using a neighbourhood of 27 grid points around a verification site); (iii) WRFMP (11 members); (iv) WRFMP NN (11 × 27 members). Neighbourhood forecasts are extracted from a cubed volume of $3 \times 3 \times 3 = 27$ grid points around the observation sites. The vertical extent of the neighbourhood roughly corresponds to the span of the rotor blades (126 m). Forecasts at each of the 27 points are corrected for their biases, based on their two-winter forecast-error statistics (cf. Section 4.3).

The impact of neighbourhood size (e.g., $5 \times 5 \times 5$ grid points or larger) is not addressed here, for two reasons: (i) WRF-IFS NN serves here only as the simplest possible reference for WRFMP—to optimize it was not the goal of this study, and (ii) the terrain surrounding the wind farms is relatively small-scale and a larger neighbourhood would be expected to introduce spatial inhomogeneity in the forecasts. For instance, cloud water content would be affected by orographic lifting in the centre of the neighbourhood, but not at its edges.

3.3 | Verification scores for probability forecasts

Probability forecasts are verified against site measurements and assessed using 2 × 2 contingency tables for each probability threshold $p_{\mathrm{t}} > 0$, 10, 25, 50, 75, and 90%. In the following, the verification scores and diagrams derived from these are introduced briefly (for a detailed discussion, cf. Wilks, 2011; Jolliffe and Stephenson, 2012).

Hit rate H quantifies the fraction of correctly forecast events with respect to the total of observed events. False-alarm rate F represents the fraction of observed non-events for which an incorrect "yes" forecast was issued. The Peirce skill score (PSS) is defined as $H - F$. PSS > 0 indicates positive model skill. The frequency bias B is the ratio of "yes" forecasts to observed events and measures the degree of overforecasting ($B > 1$) or underforecasting ($B < 1$). H, F, PSS, and B all are standard metrics for verification of binary events and are nicely related graphically using the relative operating characteristic (ROC) diagram, with F and H on the x and y axes and PSS and B drawn as isolines. For probability forecasts, $H(p_{\mathrm{t}})$ and $F(p_{\mathrm{t}})$ vary with probability threshold p_{t}, thereby spanning the ROC curve (Wilks, 2011, section 8.4.7).

The reliability diagram is a graphical means of studying the calibration function of a probability forecast, which expresses the conditional probability of the observed event as a function of the forecast probability. On the diagram, a perfectly reliable forecast follows the diagonal closely (i.e., within sampling uncertainty). Overconfident forecasts, issuing "extreme" probabilities towards 0 and 1 too frequently, are characterized by a calibration function with slopes smaller than the diagonal. Further aid in interpreting the calibration function is offered by reference lines on the reliability diagram, namely the horizontal "no-resolution" and inclined "no-skill" lines, which relate to the algebraic decomposition of the Brier (Skill) Score into *reliability*, *resolution*, and *uncertainty* terms (Wilks, 2011, section 8.4.4).

The Brier score (BS) is an integral measure of the accuracy of probability forecasts. It is defined as the mean of the squared differences of forecast probabilities (in the range $[0, 1]$) and observations (either 0 or 1). Here, we make use of the Brier skill score (BSS), the skill-score form of the BS, where the reference BS is given by the event climatology. For BSS to be positive, a probability forecast must achieve improvement upon the climatological forecast (i.e., a constant forecast probability identical to the climatological relative frequency s). A perfect forecast, instead, receives a value of $\mathrm{BSS} = 1$ (Wilks, 2011, section 8.4.3).

The potential economic value (PEV) is a measure of the added value of forecasts with respect to purely climatological information across a range of user cost–loss ratios:

$$\mathrm{PEV}(C/L) = \frac{\min(C/L, s) - F(1-s)\,C/L + Hs(1-C/L) - s}{\min(C/L, s) - s\,C/L}. \qquad (6)$$

Here, L is the loss incurred when an adverse weather event occurs without the user taking action and C is the cost of taking protective measures against the event. It is assumed that taking protective measures excludes the possibility of incurring any losses ($L = 0$ if $C > 0$). For a given C/L in the range $0 < C/L < 1$, PEV is a function of H, F, and s, and can be computed readily from contingency table elements (Richardson, 2012).

For probability forecasts, one PEV curve exists per each p_t. The optimal PEV curve $\mathrm{PEV}^{\mathrm{opt}}$ is the outer envelope of all PEV curves and describes the maximum value of the forecasting system. This value is obtained when individual users choose the optimal probability threshold according to their cost–loss ratio $p_\mathrm{t}^{\mathrm{opt}} = C/L$. By virtue of Equation (6), $\mathrm{PEV}^{\mathrm{opt}}$ peaks at the climatological frequency $C/L = s$, where it takes the value $\mathrm{PEV}^{\mathrm{opt}}_{\max} = \mathrm{PSS}_{\max} = \max(H(p_\mathrm{t}) - F(p_\mathrm{t}))$.

The area under the $\mathrm{PEV}^{\mathrm{opt}}$ curve,

$$\mathrm{AUC}_{\mathrm{PEV}} = \int_0^1 \mathrm{PEV}^{\mathrm{opt}}(C/L)\; d(C/L), \qquad (7)$$

is a scalar measure of the added value of probability forecasts. For lack of specific user information, Equation (7) assumes that all user cost–loss ratios are equally likely, resulting in $\mathrm{AUC}_{\mathrm{PEV}} \equiv \overline{\mathrm{PEV}^{\mathrm{opt}}}$. Albeit not a very common verification measure, $\mathrm{AUC}_{\mathrm{PEV}}$ has a thorough theoretical underpinning (Roebber and Bosart, 1996; Richardson, 2001; Wilks, 2001) and is employed here to afford an integral comparison of pairs of forecast variants (Section 4.5).

The statistical uncertainty of verification measures, a result of the limited size of the verification sample, is quantified in terms of 80% and 95% confidence intervals using block bootstrapping (Wilks, 1997; 2011; SSD20).

4 | VERIFICATION OF ENSEMBLE FORECASTS

4.1 | Skill of multiphysics ensemble forecasts

The forecast skill of WRFDY and WRFMP for D01 ($\Delta x = 12.5$ km), for forecast ranges up to 48 hr and during the period November 2016–March 2017, is demonstrated in Figure 2 in terms of normalized $\overline{\mathrm{CRPS}}$ and spread/error ratio. The verification scores for all three near-surface parameters (2-m temperature and dewpoint, 10-m wind) display a marked diurnal cycle. This is similar to what was observed in SSD20 concerning the temperature and wind-speed biases of deterministic forecasts. Similarly to SSD20, we hypothesize that such diurnal variability in model error depends on the imperfect representation of the land surface, surface energy balance, cloud cover, vertical mixing in the planetary boundary layer, or a combination of these factors. $\overline{\mathrm{CRPS}}$ increases (deteriorates) slowly throughout the forecast, and skill differences between the two ensembles are not always statistically significant. However, dew-point forecasts at very short ranges (< 12 hr) and wind-speed forecasts until 48 hr are significantly better (lower $\overline{\mathrm{CRPS}}$ with 95% confidence) for WRFMP. As expected, the spread-/error ratio is better (closer to 1, 95% confidence) for WRFMP for all parameters. Hence, parametrization diversity makes the ensemble forecasts less underdispersive (Figure 2b), without deteriorating the overall forecast accuracy (Figure 2a), proving the usefulness of the multiphysics approach.

The properties of WRFMP in relation to ensemble spread are clarified further in Figure 3, which compares the dispersiveness of ensemble forecasts over the D02 area from the two WRFMP domains and from the 11-member subset of the ECMWF ensemble that provides the WRFMP

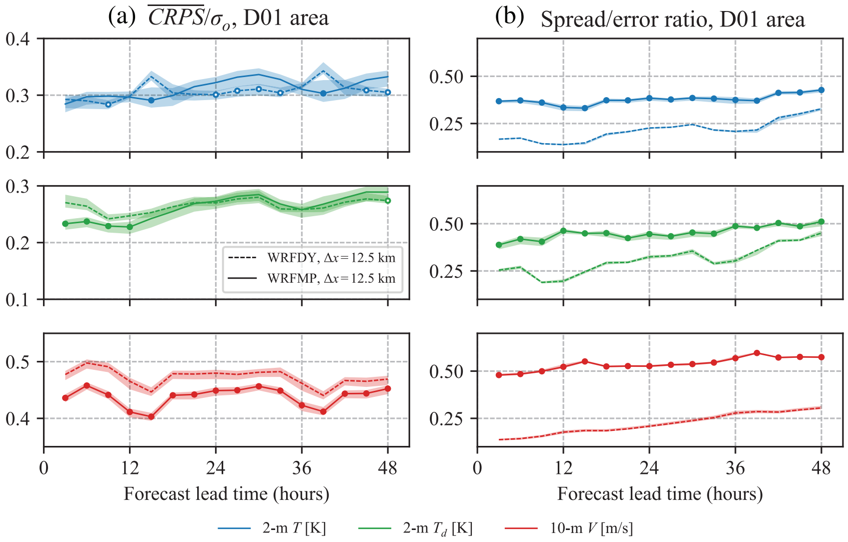

FIGURE 2 Variation with forecast range of (a) normalized continuous rank probability score and (b) ratio between ensemble spread and error of the ensemble mean, for forecasts of 2-m temperature (blue), 2-m dew point (green), and 10-m wind speed (red) by the WRFMP (solid line, full circles) and WRFDY (dashed line, open circles) ensembles with grid spacing $\Delta x = 12.5$ km. The uncertainty of verification scores is computed with a block bootstrap method, where the block size is 60 days (cf. last paragraph of Section 3.3). Shading indicates 95% confidence intervals, while bullets denote significant differences (95% confidence) between the two ensembles. Bullets are only plotted for the model with better scores (CRPS closer to zero, spread/error ratio closer to 1) [Colour figure can be viewed at wileyonlinelibrary.com]

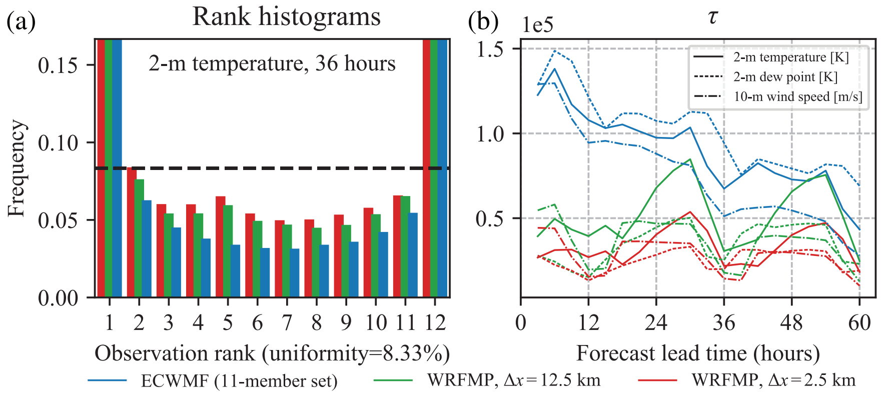

FIGURE 3 (a) Rank histograms of 36-hr forecasts of 2-m temperature and (b) τ for the rank histograms of forecasts of 2-m temperature (solid line), 2-m dew point (dashed), and 10-m wind speed (dot–dashed) at lead times between 3 and 60 hr, by the ECMWF (blue) and WRFMP (green and red, respectively, for $\Delta x = 12.5$ km and $\Delta x = 2.5$ km) ensembles [Colour figure can be viewed at wileyonlinelibrary.com]

initial and boundary conditions. A sample rank histogram (Figure 3a, referring to 36-hr forecasts of 2-m temperature) shows that WRFMP forecasts are less underdispersive than ECMWF forecasts, and that higher-resolution WRFMP forecasts ($\Delta x = 2.5$ km) are less underdispersive than coarse-resolution ones ($\Delta x = 12.5$ km). Time series of τ (Equation (5)), which quantifies the deviation of rank histograms from uniformity, are given in Figure 3b for all three verification parameters. They demonstrate that all ensembles are markedly underdispersive throughout the forecast horizon. In fact, despite a large degree of diurnal variability, the T statistic is always large enough to reject the null hypothesis of rank histograms being uniform. Ensemble forecasts from the higher-resolution WRFMP domain consistently have the lowest values of the T statistic (least underdispersiveness) at forecast ranges shorter than 36 hr. At longer ranges, the benefits of the multi-physics approach are less obvious. This is due to the degree of underdispersiveness decreasing much faster in time for the ECMWF ensemble forecasts.

Figures 4 and 5 demonstrate the effects of enhanced grid resolution and diversity in model physics on the overall accuracy of the ensemble forecasts, quantified by $\overline{\text{CRPS}}$. As expected, forecasts of all surface parameters are more accurate (lower $\overline{\text{CRPS}}$) in the range [12,36) hr than in the range [36,60) hr (Figures 4a, 5a). The spread/error ratio increases with the forecast range (Figures 4b, 5b). For WRFMP, high-resolution forecasts consistently have lower $\overline{\text{CRPS}}$ (for all parameters, at all lead times; Figure 4a), even if the result is not statistically significant because of the small dataset (194 verification points in D02). Despite the small sample size, differences in the spread/error ratio, which is closer to unity for the 2.5-km integrations, are statistically significant at 95% confidence for temperature and wind speed (Figure 4b). Ensemble forecasts from WRFMP have significantly better spread/error properties than those from ECMWF (Figure 5a), and this is accomplished by raising the ensemble spread without changing the forecast accuracy significantly in terms of $\overline{\text{CRPS}}$ (Figure 5b). Wind-speed forecasts, for which $\overline{\text{CRPS}}$

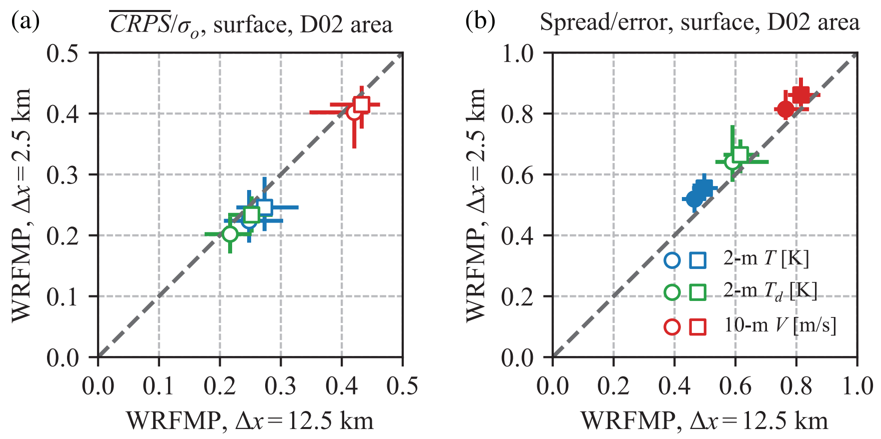

FIGURE 4 Differences in (a) CRPS and (b) spread/error ratio of temperature (blue), dew-point (green), and wind speed (red) surface-level forecasts by the low- and high-resolution WRFMP ensembles. Forecasts in the [12,36) hr and [36,60) hr intervals are denoted by circles and squares, respectively. The calculation of the uncertainty of verification scores follows the same approach used in Figure 2. Significant differences (95% confidence) between the two ensembles are represented by filled markers [Colour figure can be viewed at wileyonlinelibrary.com]

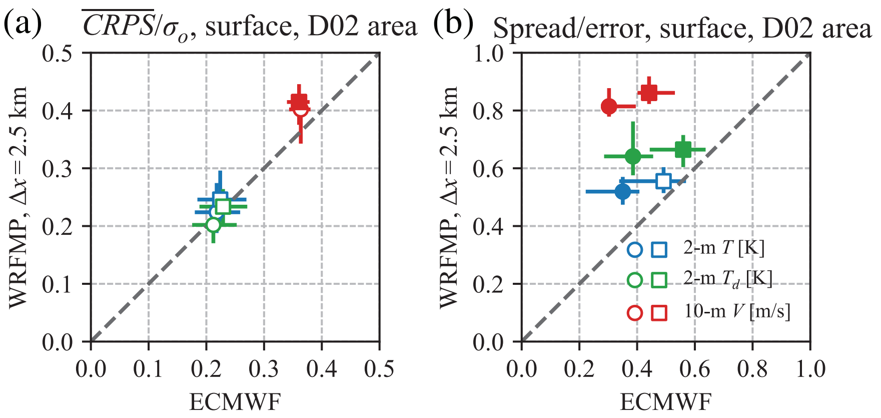

FIGURE 5 Same as Figure 4, but for the the ECMWF and high-resolution WRFMP ensembles. Unlike Figure 3, verification scores for the full 51-member ECMWF ensemble are displayed [Colour figure can be viewed at wileyonlinelibrary.com]

is significantly worse for the WRFMP example (95% confidence), are the only exception.

To summarize, WRFMP forecasts of basic weather parameters are roughly as accurate as ECMWF ones, despite the smaller ensemble size. Furthermore, short-term WRFMP forecasts are significantly less underdispersive than ECMWF ones for all three variables relevant to icing forecasting. Given the importance of short-range forecasts to make decisions on the placement of wind power on the energy market or on operating anti-icing systems on individual turbines, we expect WRFMP to enhance the value of icing forecasts to end users.

4.2 | Case studies of icing events

In SSD20, two case studies were presented to highlight the difficulty of predicting icing at a given wind-farm site accurately. The first case, January 3–4, 2017, was associated with a low-pressure system bringing moist air with a northwesterly flow from the North Sea to Germany, including to the area around EL. The second case, January 24–25, 2017, was characterized instead by a stable high-pressure system over large parts of Europe with weak pressure gradients, low temperatures, and widespread low stratus clouds. Both events were associated with significant ice accretion on the turbine blades and turbine shutdowns. Deterministic icing forecasts (WRF-IFS) predicted the ice load accreted on the turbine hub accurately in the first case, but significantly underestimated the load for the second. The differing forecast performance in the two cases seemed related to sensitivities in the modelled cloud droplet concentrations, calling for a probabilistic approach to account better for the associated forecast uncertainties.

In Figure 6, the two icing cases are revisited in the context of probabilistic forecasts. The large spreads between 5th and 95th percentiles show that forecasts of all parameters differ markedly between both WRFMP members and WRF-IFS NN members (i.e., between neighbouring grid points in the same run). In other words, parametric uncertainty (related to model formulation) and location uncertainty (related to the spatial variability of the forecast skill) are both significant. WRFMP NN forecasts, however, seem to cover a larger range of scenarios and are more dispersive. Differences in forecast spread are especially visible in Figure 6g,h, showing the accumulated ice load. In case study 1, only the 95th percentile of WRFMP NN

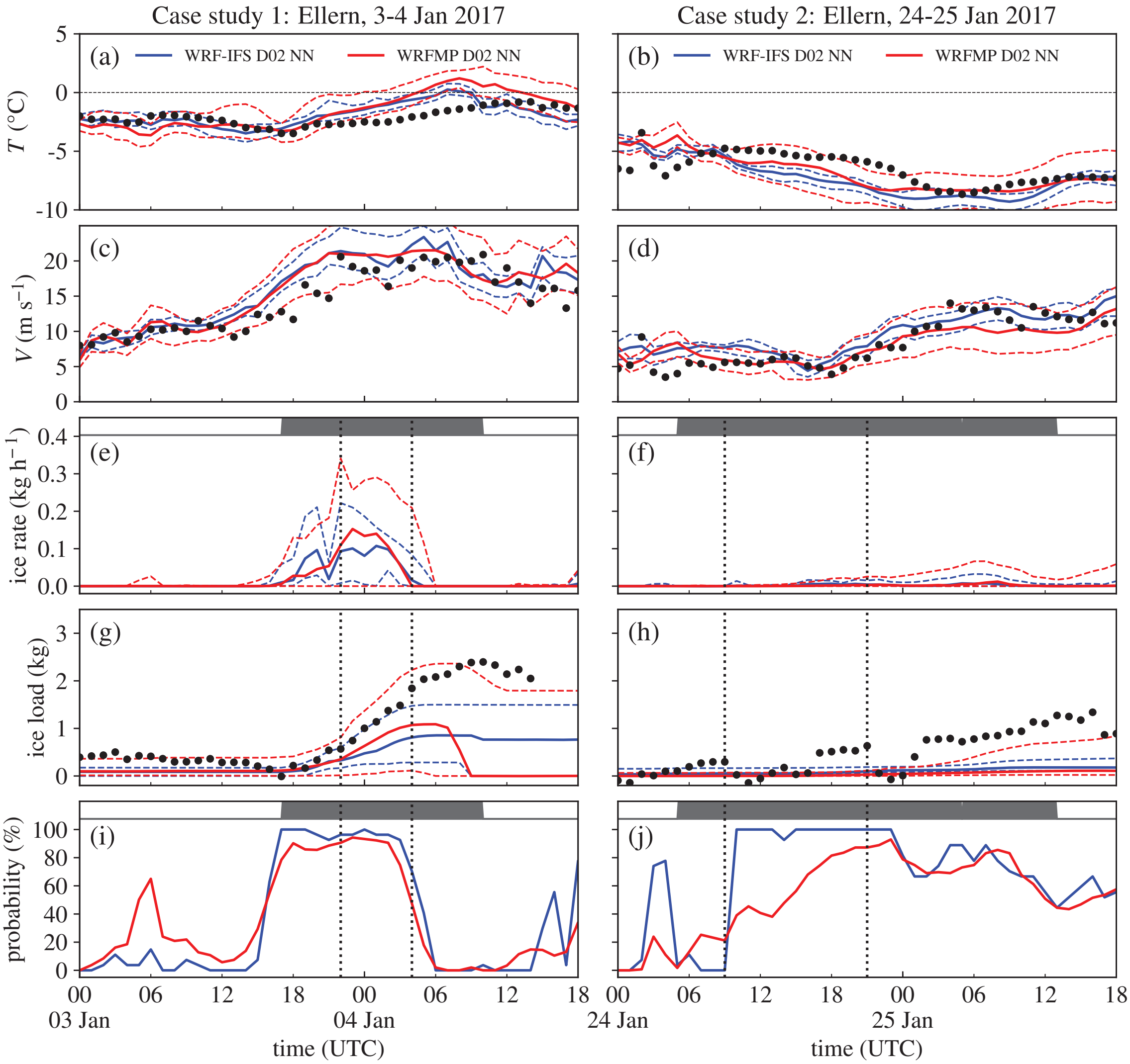

FIGURE 6 Observations and WRF ensemble forecasts for two case studies at wind farm Ellern. Left column: January 3–4, 2017; right column: January 24–25, 2017. (a,b) Temperature, (c,d) wind speed, (e,f) ice growth rate, (g,h) ice load on the Makkonen reference cylinder, and (i,j) probability of ice growth rate $> 2.5 \times 10^{-4}$ kg·hr^{-1}. Measurements are drawn as black bullets, the median of ensemble forecasts as solid lines, the 5th and 95th percentiles as dashed lines; WRF-IFS D02 NN forecasts are drawn in blue, WRFMP D02 NN forecasts in red. In panels (e,f) and (i,j), grey horizontal bars indicate the period with active ice growth as derived from camera image analyses and Makkonen cylinder ice load. Short-dashed vertical lines indicate times of icing-related turbine shutdown reported for these cases [Colour figure can be viewed at wileyonlinelibrary.com]

includes the maximum measured ice load. In case study 2, featuring a slow accumulation of ice over a two-day period, the 95th percentile of WRFMP NN doubles that of WRF-IFS NN but still underestimates the maximum observed load.

Considering probability forecasts of active ice *growth* paints a slightly different picture. These forecasts are produced by applying a relatively low threshold of $> 2.5 \times 10^{-4}$ kg·hr^{-1} to the modelled ice growth rate in each ensemble member (cf. Section 3.2). The threshold is the result of a procedure optimizing the user-oriented $\text{PEV}_{\text{max}}^{\text{opt}}$ (cf. SSD20). Using probability forecasts, both icing events are captured with reasonably good accuracy by both ensembles. The greater dispersiveness of WRFMP NN translates into reduced sharpness compared with WRF-IFS NN, evident, for instance, for case study 2 in the "smoother", more gradual increase of probability values during the event. Beyond that, both forecast variants exhibit some obvious imperfections, i.e., they produce false alarms (e.g., WRFMP NN around 0600 UTC on January 3, 2017 and WRF-IFS NN around 0300 UTC on January 24, 2017) and misses (e.g., event termination predicted too early around 0600 UTC on January 4, 2017 by both ensembles).

Overall, the case studies illustrate the character of ensemble icing forecasts on a case-by-case basis. They also give an idea of some of the forecast properties in terms of dispersiveness, sharpness, hit rates, and false-alarm rates. However, the examples do not represent these in a statistical sense, that is, they do not allow a statistically meaningful assessment of the comparison or ranking of forecast variants, which would inform users about what forecast to prefer, and at what additional expense. This comparison is achieved only by computing probabilistic verification scores, including their uncertainties, over a reasonably long verification period, which is the focus of the remainder of this article.

4.3 | Bias correction of site-specific forecasts

Since different models are in general subject to different systematic errors, a fair model comparison for site-specific forecasts requires at least the removal of mean biases, determined on the basis of measurement records. Here, we adopt the same approach as SSD20, consisting of the bin-wise removal of the mean error of forecast temperature T, dew-point temperature T_d, and wind speed V as a function of the parameter value. Bin widths of 1 K for T and T_d and $1\,\mathrm{m{\cdot}s^{-1}}$ for V are used. Since each of the physics schemes of WRFMP members (Table 1) is expected to exhibit characteristic model errors, the debiasing is carried out individually for each ensemble member.

The impact of debiasing on H and F is evident in ROC diagrams (Figure 7a–c) and reliability diagrams (Figure 7d–f). In the ROC diagrams, the effect is most evident with the deterministic WRF-IFS D02 forecasts, which were found to suffer from a cold and dry bias around and below the freezing point (SSD20). Debiasing visibly alleviates this shortcoming by a reduction of F (shift to the left in the ROC diagram) and concomitant improvement of B, including for active ice growth. A similar impact is observed for ROC curves of probability forecasts derived from the WRFMP D01 and WRFMP D02 ensembles, expressed by the reduction of maximum F for the lowest probability threshold ($p_\mathrm{t} = 10\%$, upper-rightmost points along the ROC curves).

The effect on the raw calibration function is more subtle, but most clearly evident for WRFMP D02 and T (Figure 7d), with overforecasting and overconfidence in the midrange of probabilities being removed after debiasing. The overconfidence seems to be the result of the cold bias (approx. 1 K) of most ensemble members, through which WRFMP tends to issue too high probabilities just when temperatures linger around or transition through the freezing point with oncoming frontal systems. Apart from the T condition, however, calibration is only marginally improved (Figure 7e,f), in particular for ice growth. This is attributed to the fact that T, T_d, and V are only three of several factors affecting icing (Equation (1)), liquid water content (LWC) being a crucial one, for which bias removal is not possible due to the lack of reference measurements. Nevertheless, the overall positive effects of the debiasing procedure in ROC diagrams prevail, and we follow it in the rest of this work.

4.4 | Skill and potential economic value of deterministic and probabilistic icing forecasts

SSD20 accounted for the uncertainties of icing forecasts using grid-point neighbourhood ensembles and relaxing the forecast timing. Results were deemed encouraging (judging, for instance, from increases in PSS and PEV). In this section, the added value of WRFMP with respect to previous results for deterministic forecasts is analysed. All of the following considerations regard forecasts derived from WRF D02, which were shown to be superior to D01 forecasts by SSD20. Also, we lay emphasis on the forecast range 12–35 hr, since it is the typical “day-ahead” time horizon relevant in users’ decision making.

Figure 8 offers a comparison of the performance of the four forecast variants introduced in Section 3.2 at EL. After bias correction, ROC curves (Figure 8a–c) largely overlap and differences between forecasts are hard to discern. However, the WRFMP and WRFMP NN ensembles stick out for the lowest/highest probability thresholds (their upper-rightmost/lower-leftmost points along ROC curves), where they are characterized by higher H/lower F compared with the WRF-IFS variants. For ice growth, Figure 8c, the lowest probability threshold also achieves the highest PSS.

These features are also reflected in PEV$^\mathrm{opt}$ curves (Figure 8g–i). PEV curves for condition $T \leq 0\,°\mathrm{C}$ (Figure 8g) all share approximately the same maximum value $\mathrm{PEV} = \mathrm{PSS}_\mathrm{max}$ at $C/L = s$, but their flanks differ depending on F and H. WRFMP and WRFMP NN reach comparatively lower F for the highest probability threshold, resulting in a broadening of the curves for high C/L. In other words, forecast users with only limited losses compared with their protection costs at $T \leq 0\,°\mathrm{C}$ (i.e., users with large C/L) are benefitting from smaller F, even at limited H. The same is true for users with small C/L, who can tolerate poorer F at the advantage of high H, achieved at low probability thresholds. A similar picture is obtained for the condition $T \leq 0\,°\mathrm{C}$ and $\mathrm{RH} \geq 85\%$.

For active ice growth (Figure 8i), the starkest improvement is obtained at and around the PEV maximum, with

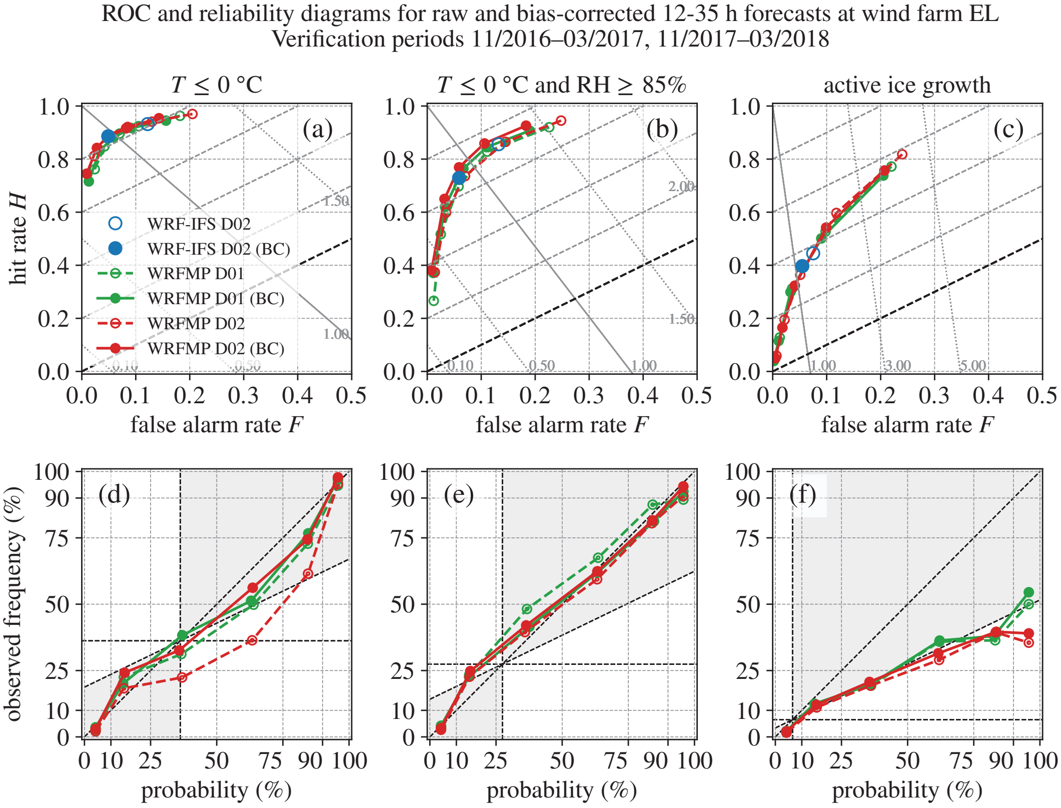

FIGURE 7 Relative operating characteristic (ROC) curves (upper panels) and reliability diagrams (lower panels) of raw and bias-corrected (BC) forecasts for range 12–35 hr in the two-winter verification period at wind farm Ellern. Left: weather condition $T \leq 0$ °C; middle: $T \leq 0$°C and RH ≥ 85%; right: active ice growth, all measured and forecast at turbine hub height. Results for the deterministic WRF D02 and ensemble WRFMP in D01 and D02 are shown (cf. legend in panel a). In ROC diagrams, F is limited to the range of 0–0.5 for better visual separation of nearby points, in contrast to the usual display of ROC diagrams showing equal ranges 0–1.0 on both axes. Diagonal grey dashed lines represent isolines of the Peirce Skill Score (PSS = $H - F$) with values increasing in steps of 0.2 towards the upper-left corner, the dashed black line corresponding to *no skill* (PSS = 0). Diagonal dotted grey lines represent isolines of forecast frequency bias B, with values given at each isoline. The line denoting $B = 1$ is in solid grey. Because each panel refers to a different type of binary event (with different climatological frequency of occurrence), frequency bias isolines have different slope. In reliablity diagrams, the diagonal black dashed 1:1 line indicates perfect reliability. The grey-coloured area marks the region with positive contribution of points along the reliability curve to the Brier Skill Score (BSS, cf.Wilks, 2011, section 8.4.4) [Colour figure can be viewed at wileyonlinelibrary.com]

the WRFMP variants achieving an additional ~0.1 or ~20% of relative increase on top of WRF-IFS NN. Overall, however, differences are slim and their statistical significance cannot be demonstrated, as indicated by the broad confidence intervals around the mean PEV curves.

Reliability diagrams generally show decent calibration for the T and T+RH conditions (Figure 8d,e), although the removal of systematic errors in T and T_d (Section 4.3) yields still imperfect results for WRF-IFS NN. The shape of the calibration curves for ice growth (Figure 8f) shows a strong degree of overconfidence of the forecasts. Also, compared with WRF-IFS NN, WRFMP variants do not improve this characteristic significantly, except maybe for WRFMP NN in the 25–50% and 90–100% probability bins. The signature of overconfidence in reliablity diagrams is in part a result of sampling the probability distribution with an ensemble of finite size, an aspect especially pronounced when considering rare events (Richardson, 2001). It is also a sign that the uncertainty of icing forecasts is still underrepresented by the multiphysics approach, similar to what was observed for other error-generation schemes (e.g., Molinder *et al.*, 2018; 2019). The consequence of overconfidence is that probabilities for ice growth cannot be used at face value by users but require calibration. Calibration, on the other hand, will effectively constrain forecast probabilities to the ranges 0–50% for WRF-IFS NN and WRFMP and 0–75% for WRFMP NN. This is also reflected in PEV curves (Figure 8i) showing no added value for $C/L > 0.5$.

Lastly, in Figure 9, we reconsider these findings for the second wind-farm site KH, as introduced in Section 2.3. Climatologically, active icing is three times more frequent there than at EL (cf. Section 2.3, and figure 3 and table 2 in SSD20). The impact of this climatological difference on verification results is not obvious, since most standard

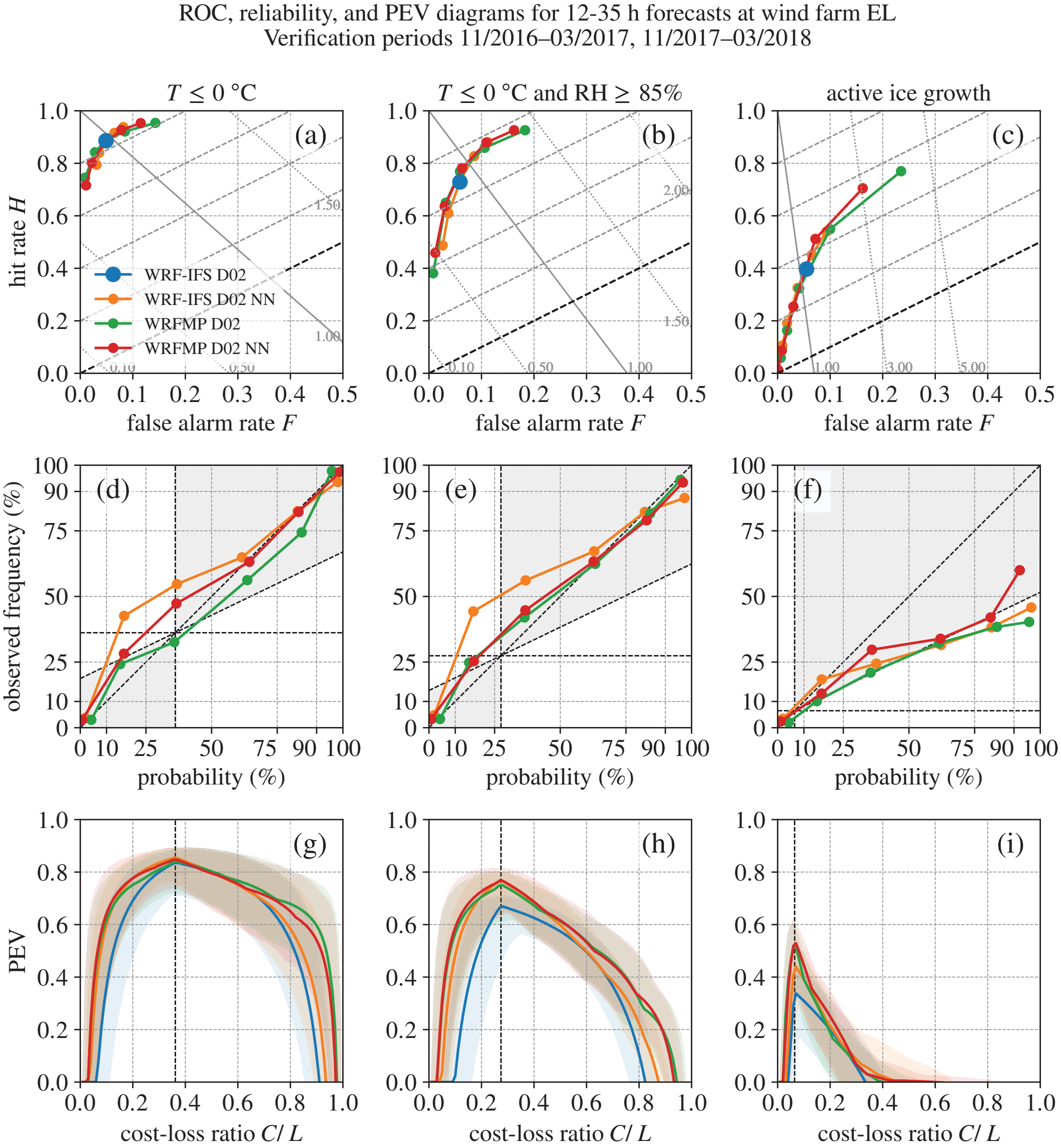

FIGURE 8 Similar to Figure 7, but with a third row of panels (g,h,i) showing optimal potential economic value curves. Results for four different variants of bias-corrected forecasts derived from WRF D02 are shown (cf. legend in panel a). On PEV diagrams, 95% confidence intervals as determined from block bootstrapping (Section 3.3) are indicated as transparent areas. The vertical dashed lines in reliablity and PEV diagrams mark the sample climatological rates of the respective weather conditions [Colour figure can be viewed at wileyonlinelibrary.com]

verification diagrams and scores for binary variables have a nonlinear dependence on the climatological rate s (cf. Section 3 and Equation (6)).

Due to the differences in climatologies, ROC and PEV curves are located in slightly different regions of the diagrams, but otherwise behave very similarly, for instance, regarding the gradual broadening of PEV curves going from WRF-IFS to WRF-IFS NN to WRFMP. Differences are most pronounced in the reliability diagram, with WRFMP and WRFMP NN forecasts being more reliable (less overcondident) even for high probabilities. Calibrated probabilities would fall in the ranges up to 75 and 90% for WRFMP and WRFMP NN, respectively. As a result, PEV^{opt} curves gain in width (Figure 9i), with WRFMP NN delivering positive PEV up to $C/L = 0.8$. The higher calibrated probabilities, as well as the broader range of nonzero PEV at KH compared with EL, is a consequence of icing being more frequent and therefore easier to predict correctly ("hit").

In view of the differences in climatology, it is remarkable that the results for KH overall confirm those for EL. It is also reassuring, in that findings for these central European wind farms could be ported to other sites: for instance, those in more Northern latitudes.

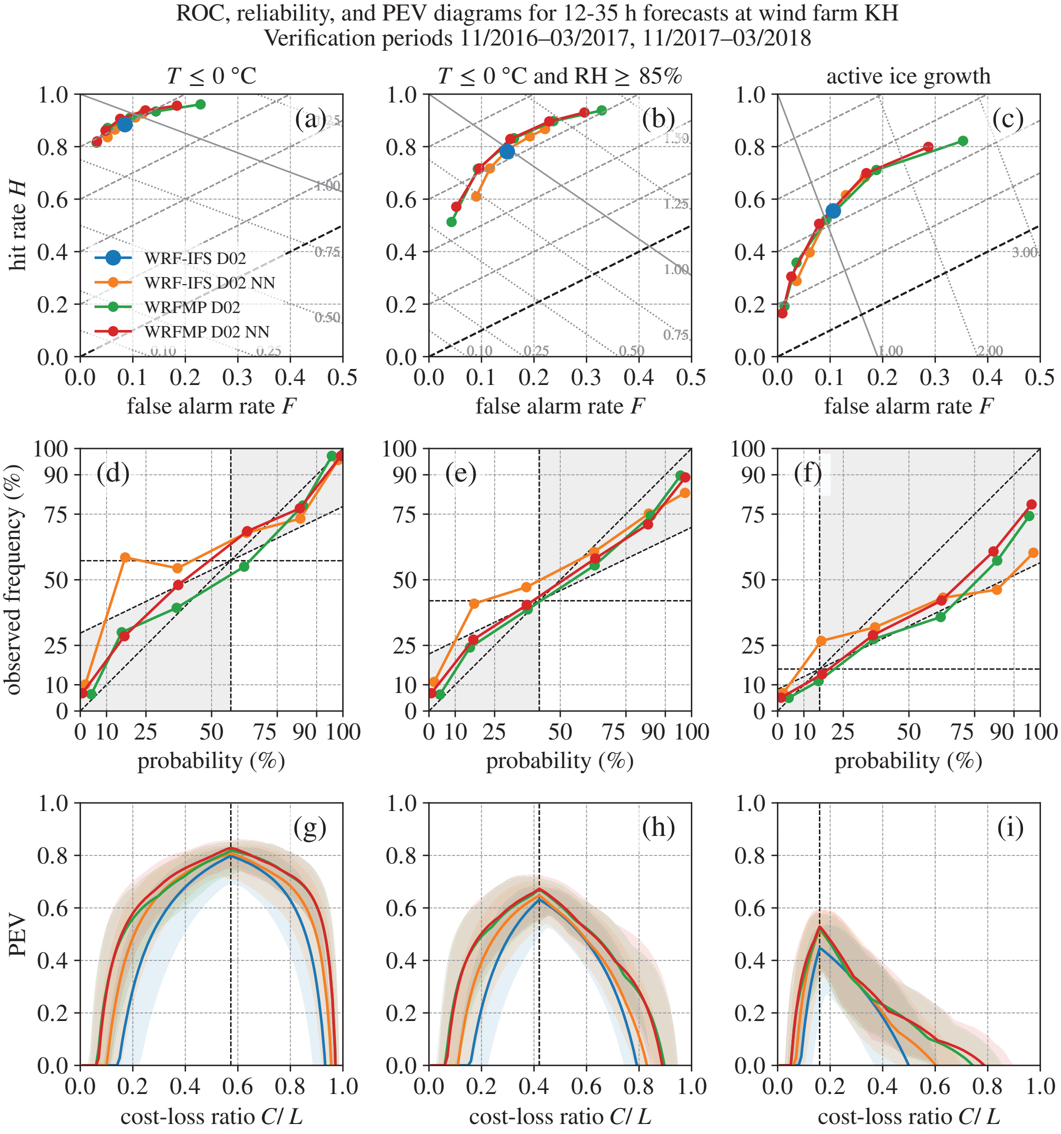

FIGURE 9 Same as Figure 8, but for wind farm Kryštofovy Hamry (KH) [Colour figure can be viewed at wileyonlinelibrary.com]

4.5 | Summary of verification scores and significance of results

Results at both EL and KH paint a picture of positive but modest improvements of WRFMP and WRFMP NN relative to WRF-IFS NN. Large confidence intervals around PEV curves (Figures 8g–i and 9g–i), however, call into question the statistical significance of improvements. To address this aspect, Figure 10 offers a comparison of verification scores for pairs of forecast variants, similar to Figures 4 and 5 for regional ensemble forecasts. To achieve such a comparison, results on ROC, reliability, and PEV diagrams (Figures 8 and 9) have been condensed to scalar verification scores: BSS (Figure 10a–c), $\mathrm{PEV}^{\mathrm{opt}}_{\mathrm{max}}$ (Figure 10d–f), and $\mathrm{AUC}_{\mathrm{PEV}}$ (Figure 10g–i). In addition, a fourth weather condition is considered: the forecast of positive ice growth within a *period* of 6 hr (±3hr), as discussed by SSD20. The pairs of forecast variants are chosen so as to allow the assessment of whether or not WRFMP and WRFMP NN offer a significant improvement of forecasts compared with the relatively “cheaper” forecasts derived from a single deterministic run. Here, cheaper refers to the effort required to set up and test a multiphysics ensemble and the computational expense of running it.

As demonstrated by SSD20, a simple grid-point neighbourhood ensemble forecast clearly outperforms a deterministic forecast for all weather conditions (Figure 10d,g). Significant improvements within 95% confidence (within 80% for freezing temperature at EL) are achieved for all weather conditions, including the icing criteria. In particular, at EL, $\mathrm{PEV}_{\mathrm{max}}$ for active ice growth increases by

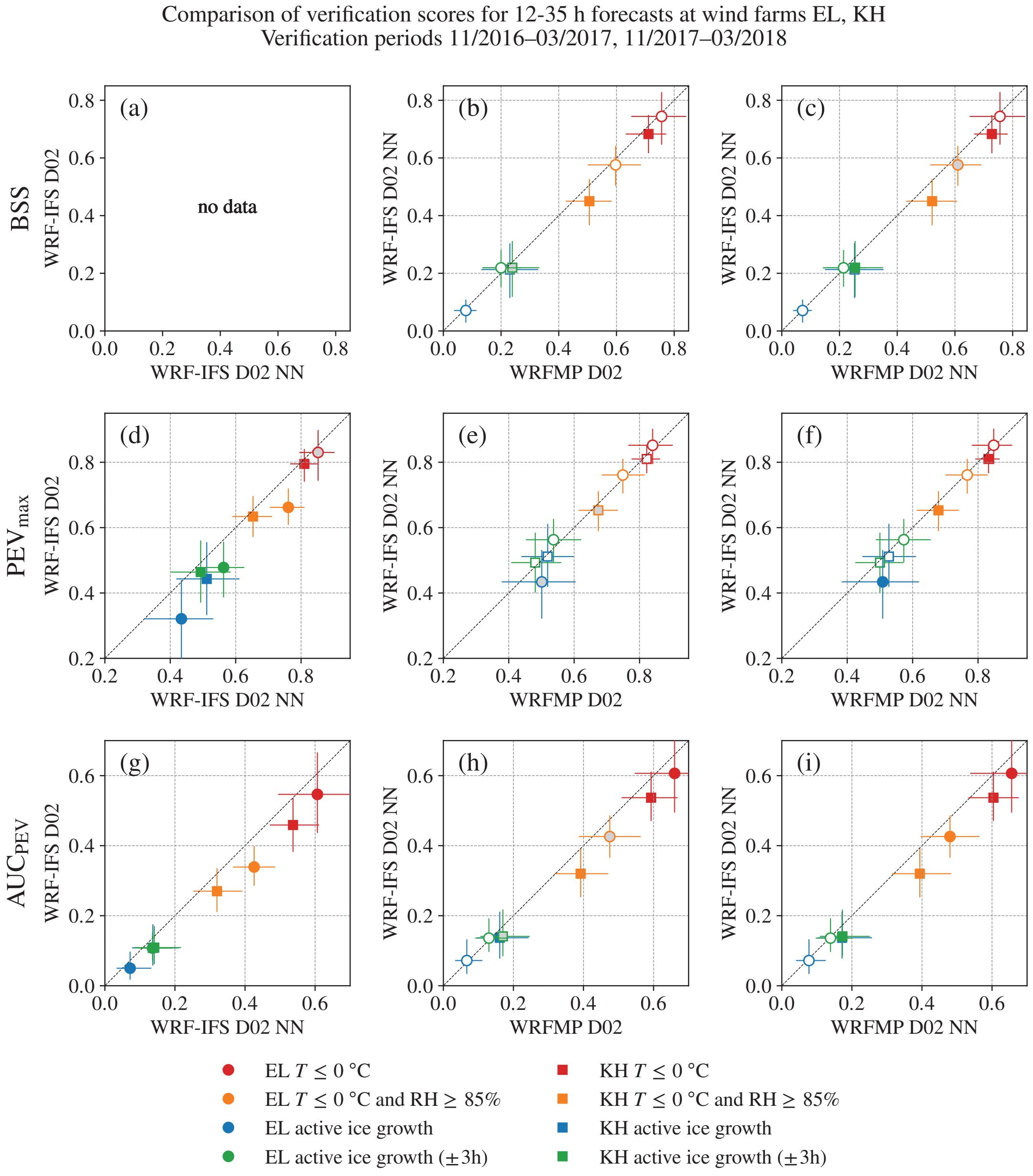

FIGURE 10 Pairwise differences of verification scores of 12–35 hr forecasts for wind farms Ellern and Kryštofovy Hamry. Upper row: Brier Skill Score BSS; middle row: maximum of optimal PEV curves PEV$_{max}$; lower row: area under the PEV curve AUC$_{PEV}$. Left column: WRF-IFS D02 versus WRF-IFS D02 NN forecasts; middle column: WRF-IFS D02 NN versus WRFMP D02 forecasts; right column: WRF D02 NN versus WRFMP D02 NN forecasts. Horizontal and vertical bars on all panels mark the 95% confidence intervals of the corresponding score and forecast variant. Colour-filled symbols indicate differences in verification scores deemed significant at the 95% confidence level, grey-filled symbols at 80% confidence level, empty circles at ≤80% confidence [Colour figure can be viewed at wileyonlinelibrary.com]

~35%, from ~0.32 (WRF-IFS) to ~0.43 (WRF-IFS NN). Forecasts of ice growth for a period of 6 hr at EL are considerably better than their hourly counterparts and are competitive with those at KH, a pervasive result across all scores. For practical purposes, such temporally coarsened forecasts can be preferable for users with inexpensive anti-icing solutions at hand, who can afford extended heating periods.

The positive changes in PEV$_{max}$, going from WRF-IFS to WRF-IFS NN, do not translate directly into AUC$_{PEV}$ (Figure 10g). Considerable gains are seen for freezing and humid conditions, as expected from the PEV curves in Figures 8g,h and 9g,h. On the other hand, gains for icing forecasts are nominally deemed significant at 95% confidence level, but seem marginal in absolute terms. The diverging findings for PEV$_{max}$ and AUC$_{PEV}$

suggest that, using grid-point neighbourhood ensembles, added value is obtained only for a small range of user cost–loss ratios, mostly around the icing climatological base rates.

Figure 10b,e,h shows the comparison between WRFMP and WRF-IFS NN, including BSS. BSS is obtained after a simple linear calibration using the respective calibration functions from Figures 8d–f and 9d–f. BSS values are almost identical at EL, but slightly better for WRFMP at KH, even if the result for icing is found to be insignificant only at 80% confidence level. Similarly, $\mathrm{PEV}_{\max}$ of WRFMP forecasts is not improved significantly. The exception is icing on an hourly basis for EL. For $\mathrm{AUC}_{\mathrm{PEV}}$, instead, the broadening of $\mathrm{PEV}^{\mathrm{opt}}$ curves (Figures 8g–i and 9g–i) translates into significant gains (except for icing at EL). Generally speaking, this broadening constitutes the greatest advantage of probabilistic over deterministic forecasting systems, bringing added value over a broader range of user cost–loss ratios (cf. Richardson, 2012).

Finally, Figure 10c,f,i presents the differences between WRFMP NN and WRF-IFS NN, both taking into account the grid-point neighbourhood. Previous gains by WRFMP are enhanced further, the significance of most results reaching the 80% or 95% confidence level. The gradual improvement in all measures going from left to right in Figure 10 is in line with previous findings by Molinder *et al.* (2018), showing that the inclusion of both boundary and initial condition errors (through a single-physics ensemble) *and* spatial representativeness error (through a neighbourhood ensemble) scores best.

5 | SUMMARY AND CONCLUSIONS

In this study, a regional multiphysics ensemble forecasting system based on WRF is set up and evaluated for the application of icing on wind turbines. The multiphysics approach, involving combinations of surface and boundary-layer and microphysics parametrizations, represents an extension to previous works that accounted for initial and boundary-condition uncertainties in NWP models or model parameter uncertainties in icing models.

An ensemble verification of the regional 11-member WRFMP ensemble is conducted to determine its added value with respect to the ECMWF EPS and a dynamical downscaling ensemble WRFDY, which accounts only for IC and BC uncertainties. The accuracy and dispersiveness of the ensembles are studied by means of $\overline{\mathrm{CRPS}}$, spread/error ratio, and rank histograms for surface parameters T, T_d, and wind speed. WRFMP forecasts have significantly better (i.e., closer to unity) spread/error properties than those from ECMWF and WRFDY, and this is accomplished by raising the ensemble spread without changing the $\overline{\mathrm{CRPS}}$ significantly. That is, parametrization diversity makes the ensemble forecasts less underdispersive without deteriorating the overall forecast accuracy. The benefits of WRFMP are most obvious at forecast ranges below 36 hr, most relevant for short-term decision making by forecast users. Nevertheless, WRFMP forecasts remain underdispersive (spread/error ratio around 0.5 for temperature, dewpoint and wind-speed forecasts). We speculate that the WRFMP spread/error properties could be improved further by drawing its ICs and BCs from previous-day runs of the parent ECMWF EPS ensemble. This is based on the assumption that, for the parent ensemble, the 24–60 hr forecast range should have a negligible loss in deterministic forecast quality compared with the 0–36 hr range, but considerably larger spread.

Data from two wind parks in Central Europe, near Ellern (Germany) and Kryštofovy Hamry (Czech Republic), are used to verify site-specific forecasts of ice accretion on turbine hubs. The reference icing measurements stem from a categorical analysis of ice growth based on camera images. Icing events are defined as ice growth conditions of the “light” or greater category. Several variants of forecasts are compared: a deterministic forecast (WRF-IFS), the probability forecasts derived from a neighbourhood ensemble only (WRF-IFS NN), the multiphysics ensemble WRFMP, and finally WRFMP combined with the neighbourhood method (WRFMP NN). WRF-IFS NN serves a computationally cheap, minimal probabilistic model that other model variants need to beat for them to be attributed added value.

The properties of the site-specific probability forecasts are studied using relative operating characteristic, reliability, and potential economic value diagrams. The main findings are the following.

1. Probability forecasts of ice growth increase the maximum potential economic value significantly, especially so at the Ellern site (from ∼0.32 for WRF-IFS to ∼0.50 for the WRFMP NN, a relative improvement of ∼55%). On the other hand, only slim gains in the width of PEV curves are obtained, meaning that the added value is constrained to forecast users whose application is characterized by cost–loss ratios around the climatological base rate at the site.
2. Icing forecasts are characterized by a strong degree of overconfidence, similar to what was found in other studies of limited-area ensembles for convective precipitation (cf. references in Section 1.1). For icing, the WRFMP variants fare slightly better in this regard than neighbourhood ensembles. Generally, however, forecast probabilities cannot be used at face value, but

require calibration for users to draw benefit from them in a cost/loss framework (Richardson, 2012).

3. While the relative improvements of the probability forecasts with increasing model complexity are found to be positive and statistically significant, gains offered by the WRFMP variants seem modest in absolute terms, in particular when compared with the neighbourhood method. The exception appear to be sites with a rare occurrence of icing, such as Ellern, for which the icing prediction is especially challenging.

We are left to speculate about why the benefits of the multiphysics approach are limited. Results from regional verification of surface parameters point to physics model error being an important component to uncertainty, which is well represented choosing a set of SL and BL schemes in the ensemble. In the literature about icing forecasts, some studies (Nygaard *et al.*, 2011; Davis *et al.*, 2014) alluded to the potential of using multiple physics parametrizations in an ensemble sense, without performing a quantitative evaluation of the impact. Our work clearly shows that, for this specific application and with the current state of the art in model physics, ensemble forecasts have only marginally larger value for end users than the deterministic ones. Accounting for location and representativeness errors with a relatively cheap neighbourhood approach built on those deterministic forecasts may be considered as the best compromise between model complexity and actual forecast value for end users.

For the future, advances could be made along two routes. This and previous works have explored error-representation schemes for icing forecasts: (i) initial and boundary condition sensitivity (Molinder *et al.*, 2018), (ii) spatial representativeness (Molinder *et al.*, 2018; SSD20), (iii) icing model parameter uncertainties (Molinder *et al.*, 2019), and (iv) model physics sensitivity (Davis *et al.*, 2014, and the present study). In addition, stochastic perturbations schemes (SPP, SPPT, SKEBS, cf. Section 1.1) should be considered. Indeed, very recent research introducing SPP to the Thompson & Eidhammer microphysics scheme seems promising, although still too little spread is reported for cold-season case studies when applying SPP only to cloud physics (Thompson *et al.*, 2021). Therefore, ideally, all these approaches should be applied in concert, testing their impact on the icing uncertainty forecast.

The practicality of setting up, running, and maintaining an ensemble prediction system of such complexity, however, may seem questionable from an operational forecasting perspective. Efforts could therefore refocus on the nowcasting of icing conditions in the shortest ($\lesssim$8 hr) forecast ranges, helping users to compensate for losses in the short term on the intraday market or prevent them at short notice with proactive anti-icing measures. Novel postprocessing techniques have been tested recently, incorporating predictor variables from both current observations of weather and turbine-specific parameters and NWP model output (Kreutz *et al.*, 2019; Scher and Molinder, 2019; Molinder *et al.*, 2021). These techniques, too, are not without challenges. For example, evidence is yet to be produced that they can outperform persistence forecasts for the shortest lead times (cf. Scher and Molinder, 2019, figures 2, 5, 6) and define the onset, duration, and end of icing episodes better.

AUTHOR CONTRIBUTIONS

Lukas Strauss: conceptualization; data curation; formal analysis; investigation; methodology; validation; visualization; writing – original draft; writing – review and editing. **Stefano Serafin:** conceptualization; data curation; formal analysis; investigation; methodology; validation; visualization; writing – original draft; writing – review and editing. **Manfred Dorninger:** conceptualization; data curation; funding acquisition; investigation; project administration; resources; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

This research was funded by the Austrian Research Promotion Agency (FFG) within the second call of the “Energieforschungsprogramm des Klima- und Energiefonds”, project ICE CONTROL (853575). A visit to NCAR by L.S. was supported through a stipend by the Ministry of Education, Science and Research of Austria. The computational results presented have been achieved using the Vienna Scientific Cluster (VSC). Access to ECMWF data was provided through the Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Austria. Data for wind farm Kryštofovy Hamry were kindly made available by ENERCON GmbH through a cooperation agreement. The authors thank all members of the ICE CONTROL project team for their individual contributions and are grateful to William Cheng, Gregory Thompson, Neil Davis, and Jennie Molinder for useful discussions. The constructive criticism and suggestions for improvement by two anonymous reviewers are gratefully acknowledged.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

ORCID

Lukas Strauss https://orcid.org/0000-0001-8275-0806
Stefano Serafin https://orcid.org/0000-0002-5838-7514
Manfred Dorninger https://orcid.org/0000-0002-7389-2236

REFERENCES

Arribas, A., Robertson, K.B. and Mylne, K.R. (2005) Test of a poor man's ensemble prediction system for short-range probability forecasting. *Monthly Weather Review*, 133(7), 1825–1839. https://doi.org/10.1175/MWR2911.1.

Bergström, H., Olsson, E., Söderberg, S., Thorsson, P. and Undén, P. (2013) *Wind power in cold climates: Ice mapping methods*. Technical report. Available online at http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A704372&amp;dswid=4541 [Accessed July 27, 2020].

Berner, J., Shutts, G.J., Leutbecher, M. and Palmer, T.N. (2009) A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, 66(3), 603–626. https://doi.org/10.1175/2008JAS2677.1.

Berner, J., Ha, S.Y., Hacker, J.P., Fournier, A. and Snyder, C. (2011) Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Monthly Weather Review*, 139, 1972–1995. https://doi.org/10.1175/2010MWR3595.1.

Berner, J., Fossell, K.R., Ha, S.-Y., Hacker, J.P. and Snyder, C. (2015) Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Monthly Weather Review*, 143(4), 1295–1320. https://doi.org/10.1175/MWR-D-14-00091.1.

Bernstein, B.C., Hirvonen, J., Gregow, E. and Wittmeyer, I. (2012). Experiences from real-time LAPS–LOWICE runs over Sweden: 2011–2012 icing season. In: *Winterwind International Wind Energy Conference 2012, Skellefteåa, Sweden*. Available online at url=http://www.slideshare.net/WinterwindConference/3a-bernstein-lapslowice [Accessed July 27, 2020].

Bougeault, P. and Lacarrere, P. (1989) Parameterization of orography-induced turbulence in a mesobeta-scale model. *Monthly Weather Review*, 117, 1872–1890. https://doi.org/10.1175/1520-0493(1989)117<1872:POOITI>2.0.CO;2.

Bredesen, R., Cattin, R., Clausen, N.-E., Davis, N., Jordaens, P.J., Khadiri-Yazami, Z., Klintström, R., Krenn, A., Lehtomäki, V., Ronsten, G., Wadham-Gagnon, M. and Wickman, H. (2017) *IEA Wind TCP Task 19 Report: Recommended practices* (2nd Edition). International Energy Agency. Available online at https://iea-wind.org/wp-content/uploads/2021/09/2017-IEA-Wind-TCP-Recommended-Practice-13-2nd-Edition-Wind-Energy-in-Cold-Climates.pdf [Accessed July 27, 2020].

Buizza, R., Miller, M. and Palmer, T.N. (2007) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560), 2887–2908. https://doi.org/10.1002/qj.49712556006.

Clark, A.J., Gallus, W.A. and Chen, T.C. (2008) Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Monthly Weather Review*, 136(6), 2140–2156. https://doi.org/10.1175/2007MWR2029.1.

Cohen, A.E., Cavallo, S.M., Coniglio, M.C. and Brooks, H.E. (2015) A review of planetary boundary layer parametrization schemes and their sensitivity in simulating southeastern U.S. cold season severe weather environments. *Weather and Forecasting*, 30(3), 591–612. https://doi.org/10.1175/WAF-D-14-00105.1.

Colarco, P., da Silva, A., Chin, M. and Diehl, T. (2010) Online simulations of global aerosol distributions in the NASA GEOS-4 model and comparisons to satellite and ground-based aerosol optical depth. *Journal of Geophysical Research–Atmospheres*, 115, D14207. https://doi.org/10.1029/2009JD012820.

Copernicus Land Monitoring Service (2012) *CORINE land cover inventory*. Available online at https://land.copernicus.eu/pan-european/corine-land-cover [Accessed August 25, 2020].

Davis, N., Hahmann, A.N., Clausen, N.-E. and Žagar, M. (2014) Forecast of icing events at a wind farm in Sweden. *Journal of Applied Meteorology and Climatology*, 53, 262–281. https://doi.org/10.1175/JAMC-D-13-09.1.

Davis, N., Pinson, P., Hahmann, A.N., Clausen, N.-E. and Žagar, M. (2016) Identifying and characterizing the impact of turbine icing on wind farm power generation. *Wind Energy*, 19, 1503–1518. https://doi.org/10.1002/we.1933.

Duda, J.D., Wang, X., Kong, F., Xue, M. and Berner, J. (2017) Impact of a stochastic kinetic energy backscatter scheme on warm season convection-allowing ensemble forecasts. *Monthly Weather Review*, 144, 1887–1908. https://doi.org/10.1175/MWR-D-15-0092.1.

Eckel, F.A. and Mass, C.F. (2005) Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, 20, 328–350. https://doi.org/10.1175/WAF843.1.

Ginoux, P., Chin, M., Tegen, I., Prospero, J.M., Holben, B., Dubovik, O. and Lin, S.J. (2001) Sources and distributions of dust aerosols simulated with the GOCART model. *Journal of Geophysical Research–Atmospheres*, 106, 20255–20273. https://doi.org/10.1029/2000JD000053.

Hacker, J.P., Ha, S.-Y., Snyder, C., Berner, J., Eckel, F.A., Kuchera, E., Pocernich, M., Rugg, S., Schramm, J. and Wang, X. (2011a) The U.S. Air Force Weather Agency's mesoscale ensemble: scientific description and performance results. *Tellus A: Dynamic Meteorology and Oceanography*, 63(3), 625–641. https://doi.org/10.1111/j.1600-0870.2010.00497.x.

Hacker, J.P., Snyder, C., Ha, S.-Y. and Pocernich, M. (2011b) Linear and non-linear response to parameter variations in a mesoscale model. *Tellus A: Dynamic Meteorology and Oceanography*, 63(3), 429–444. https://doi.org/10.1111/j.1600-0870.2010.00505.x.

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.

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

Hong, S.-Y., Noh, Y. and Dudhia, J. (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, 134, 2318–2341. https://doi.org/10.1175/MWR3199.1.

Janjić, Z.I. (1994) The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Monthly Weather Review*, 122, 927–945. https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

Jankov, I., Berner, J., Beck, J., Jiang, H., Olson, J.B., Grell, G., Smirnova, T.G., Benjamin, S.G. and Brown, J.M. (2017) A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Monthly Weather Review*, 145, 1161–1179. https://doi.org/10.1175/MWR-D-16-0160.1.

Jiménez, P.A., Dudhia, J., Fidel González-Rouco, J., Navarro, J., Montávez, J.P. and García-Bustamante, E. (2012) A revised scheme for the WRF surface layer formulation. *Monthly Weather Review*, 140, 898–918. https://doi.org/10.1175/MWR-D-11-00056.1.

Jolliffe, I.T. and Stephenson, D.B. (2012) *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (Second edition). Chichester, UK: John Wiley & Sons, Ltd.

Karlsson, T. (2021). Cold climate wind market study 2020–2025. In: *Winterwind International Wind Energy Conference 2021*. Available online at https://windren.se/WW2021/14_2_21_Karlsson_IEA_Wind_Task_19_Cold_climate_wind_market_study_Public.pdf [accessed June 9, 2022].

Krenn, A., Stökl, A., Weber, N., Barup, S., Weidl, T., Hoffmann, A., Bredesen, R.E., Lannic, M., Müller, S., Stoffels, N., Hahm, T., Storck, F. and Lautenschlager, F. (2018) *IEA Wind TCP Task 19: International recommendations for ice fall and ice throw risk assessments*. Technical report. Available online at https://iea-wind.org/wp-content/uploads/2021/09/Task19_Recommendations_ice_throw_2018.pdf [Accessed July 11, 2020].

Kreutz, M., Ait-Alla, A., Varasteh, K., Oelker, S., Greulich, A., Freitag, M. and Thoben, K.-D. (2019) Machine learning-based icing prediction on wind turbines. *Proc. CIRP*, 81, 423–428. https://doi.org/10.1016/j.procir.2019.03.073.

Laakso, T., Talhaug, L., Ronsten, G., Horbaty, R., Baring-Gould, I., Lacroix, A., Peltola, E., Wallenius, T. and Durstewitz, M. (2009) *IEA Wind TCP Task 19: Wind energy in cold climates*. Technical Report. [Accessed: February 24, 2021].

Leith, C.E. (1974) Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, 102, 409–418. https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Makkonen, L. (2000) Models for the growth of rime, glaze, icicles and wet snow on structures. *Philosophical Transactions of the Royal Society A*, 358, 2913–2939. https://doi.org/10.1098/rsta.2000.0690.

Molinder, J., Körnich, H., Olsson, E., Bergström, H. and Sjöblom, A. (2018) Probabilistic forecasting of wind power production losses in cold climates: a case study. *Wind Energy Science*, 3, 667–680. https://doi.org/10.5194/wes-3-667-2018.

Molinder, J., Körnich, H., Olsson, E. and Hessling, P. (2019) The use of uncertainty quantification for the empirical modeling of wind turbine icing. *Journal of Applied Meteorology and Climatology*, 58, 2019–2032. https://doi.org/10.1175/JAMC-D-18-0160.1.

Molinder, J., Scher, S., Nilsson, E., Körnich, H., Bergström, H. and Sjöblom, A. (2021) Probabilistic forecasting of wind turbine icing related production losses using quantile regression forests. *Energies*, 14, 158. https://doi.org/10.3390/en14010158.

Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T. (1996) The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529), 73–119. https://doi.org/10.1256/smsqj.52904.

Morrison, H., Thompson, G. and Tatarskii, V. (2009) Impact of cloud microphysics on the development of trailing stratiform precipitation in a simulated squall line: Comparison of one- and two-moment schemes. *Monthly Weather Review*, 137, 991–1007. https://doi.org/10.1175/2008MWR2556.1.

Nakanishi, M. and Niino, H. (2006) An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Boundary-Layer Meteorology*, 119, 397–407. https://doi.org/10.1007/s10546-005-9030-8.

NCAR Research Applications Laboratory (2020) *Unified NOAH Land-Surface Model*. Available online at https://ral.ucar.edu/solutions/products/unified-noah-lsm [Accessed August 25, 2020].

Nygaard, B.E.K., Kristjánsson, J.E. and Makkonen, L. (2011) Prediction of in-cloud icing conditions at ground level using the WRF model. *Journal of Applied Meteorology and Climatology*, 50, 2445–2459. https://doi.org/10.1175/JAMC-D-11-054.1.

Nygaard, B.E.K., Ágústsson, H. and Somfalvi-Toth, K. (2013) Modeling wet snow accretion on power lines: Improvements to previous methods using 50 years of observations. *Journal of Applied Meteorology and Climatology*, 52, 2189–2203. https://doi.org/10.1175/JAMC-D-12-0332.1.

Richardson, D.S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127, 2473–2489. https://doi.org/10.1002/qj.49712757715.

Richardson, D.S. (2012). Economic value and skill. In I.T. Jolliffe and D.B. Stephenson (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, (2nd edition), pp. 167–184: Chichester, UK: John Wiley & Sons, Ltd, DOI 10.1002/9781119960003.

Roebber, P.J. and Bosart, L.F. (1996) The complex relationship between forecast skill and forecast value: A real-world analysis. *Weather and Forecasting*, 11, 544–559. https://doi.org/10.1175/1520-0434(1996)011<0544:TCRBFS>2.0.CO;2.

Romine, G.S., Schwartz, C.S., Berner, J., Fossell, K.R., Snyder, C., Anderson, J.L. and Weisman, M.L. (2014) Representing forecast error in a convection-permitting ensemble system. *Monthly Weather Review*, 142, 4519–4541. https://doi.org/10.1175/MWR-D-14-00100.1.

Scher, S. and Molinder, J. (2019) Machine learning-based prediction of icing-related wind power production loss. *IEEE Access*, 7, 129421–129429. https://doi.org/10.1109/ACCESS.2019.2939657.

Schwartz, C.S. and Sobash, R.A. (2017) Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Monthly Weather Review*, 145, 3397–3418. https://doi.org/10.1175/MWR-D-16-0400.1.

Serafin, S., Strauss, L. and Dorninger, M. (2019) Ensemble reduction using cluster analysis. *Quarterly Journal of the Royal Meteorological Society*, 145, 659–674. https://doi.org/10.1002/qj.3458.

Skamarock, W.C. and Klemp, J.B. (2008) A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics*, 227, 3465–3485. https://doi.org/10.1016/j.jcp.2007.01.037.

Smirnova, T.G., Brown, J.M., Benjamin, S.G. and Kenyon, J.S. (2016) Modifications to the Rapid Update Cycle Land Surface Model (RUC LSM) available in the Weather Research and Forecasting (WRF) model. *Monthly Weather Review*, 144, 1851–1865. https://doi.org/10.1175/MWR-D-15-0198.1.

Smith, A., Lott, N. and Vose, R. (2011) The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, 92, 704–708. https://doi.org/10.1175/2011BAMS3015.1.

Stensrud, D.J., Bao, J.-W. and Warner, T.T. (2000) Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Monthly Weather Review*, 128, 2077–2107. https://doi.org/10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

Strauss, L., Serafin, S. and Dorninger, M. (2020) Skill and potential economic value of forecasts of ice accretion on wind turbines. *Journal of Applied Meteorology and Climatology*, 59, 1845–1864. https://doi.org/10.1175/JAMC-D-20-0025.1.

Sukoriansky, S., Galperin, B. and Perov, V. (2005) Application of a new spectral theory of stably stratified turbulence to the atmospheric boundary layer over sea ice. *Boundary-Layer Meteorology*, 117, 231–257. https://doi.org/10.1007/s10546-004-6848-4.

Thompson, G. and Eidhammer, T. (2014) A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *Journal of the Atmospheric Sciences*, 71, 3636–3658. https://doi.org/10.1175/JAS-D-13-0305.1.

Thompson, G., Field, P.R., Rasmussen, R.M. and Hall, W.D. (2008) Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parametrization. *Monthly Weather Review*, 136, 5095–5115. https://doi.org/10.1175/2008MWR2387.1.

Thompson, G., Nygaard, B.E., Makkonen, L. and Dierer, S. (2009). Using the Weather Research and Forecasting (WRF) model to predict ground/structural icing. In: *Proceedings of the 13th International Workshop on Atmospheric Icing of Structures (IWAIS XIII), Andermatt, Switzerland*. Available online at https://www.compusult.com/web/iwais/iwais-2009 [Accessed July 27, 2020].

Thompson, G., Berner, J., Frediani, M., Otkin, J.A. and Griffin, S.M. (2021) A stochastic parameter perturbation method to represent uncertainty in a microphysics scheme. *Monthly Weather Review*, 149, 1481–1497. https://doi.org/10.1175/MWR-D-20-0077.1.

Weigel, A.P. (2012). Ensemble forecasts. In I.T. Jolliffe and D.B. Stephenson (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, (Second edition), pp. 141–166: Chichester, UK: John Wiley & Sons, Ltd, DOI 10.1002/9781119960003.

Wilks, D.S. (2001) A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8, 209–219. https://doi.org/10.1017/S1350482701002092.

Wilks, D.S. (1997) Resampling hypothesis tests for autocorrelated fields. *Journal of Climate*, 10, 65–82. https://doi.org/10.1175/1520-0442(1997)010<0065:RHTFAF>2.0.CO;2.

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

How to cite this article: Strauss, L., Serafin, S. & Dorninger, M. (2022) Probability forecasts of ice accretion on wind turbines derived from multiphysics and neighbourhood ensembles. *Quarterly Journal of the Royal Meteorological Society*, 148(746), 2446–2467. Available from: https://doi.org/10.1002/qj.4311