The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble

Article

Accepted Version

Cafaro, C. ORCID: https://orcid.org/0000-0001-8063-4887, Frame, T. H. A. ORCID: https://orcid.org/0000-0001-6542-2173, Methven, J. ORCID: https://orcid.org/0000-0002-7636-6872, Roberts, N. and Bröcker, J. (2019) The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble. Quarterly Journal of the Royal Meteorological Society, 145 (721). pp. 1780-1798. ISSN 1477-870X doi: 10.1002/qj.3531 Available at https://centaur.reading.ac.uk/82958/

Published version at: https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3531
To link to this article DOI: http://dx.doi.org/10.1002/qj.3531

Publisher: Royal Meteorological Society

The added value of convection-permitting ensemble forecasts of sea breeze compared to a Bayesian forecast driven by the global ensemble

Carlo Cafaro[1,2,4]* | Thomas H. A. Frame[2] | John Methven[2] | Nigel Roberts[3] | Jochen Bröcker[1,2]

[1]Department of Mathematics and Statistics, University of Reading, RG6 6AX, United Kingdom

[2]Department of Meteorology, University of Reading, RG6 6BB, United Kingdom

[3]Met Office@Reading

[4]NCEO, Reading

Correspondence
Department of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading, RG6 6BB, UK
Email: c.cafaro@reading.ac.uk

Funding information
EPSRC, Grant/Award Number:EP/L016613/1.

Dynamical downscaling of ensemble forecasts to convection-permitting resolutions aims to improve forecast skill by explicitly resolving mesoscale dynamical features. The success of this approach is dependent on the ability of the model to spin up smaller features embedded in the larger scale flow and provide more local information than could be inferred from knowledge of the climatological response to the large scale flow alone. Here we test whether such additional information is obtained from the Met Office Global and Regional Ensemble Prediction Systems (MOGREPS) for the sea-breeze phenomena which is resolved properly only at convection-permitting resolutions but is driven by large-sale conditions that are well represented in the global driving model. The sea breeze is detected using a new automatic tracking algorithm suitable for use in convective-scale forecast data. The skill of probabilistic forecasts of sea-breeze occurrence from the high resolution ensemble is compared to that of a Bayesian forecast trained on paired high/low resolution ensemble members. This creates a statistical forecast of the high-resolution ensemble member given knowledge of the global forecast ensemble alone. The aim of this paper is twofold: -firstly to assess whether the information about sea breeze occurrence is encoded in a few large-scale parameters and can be forecast by a statistical method ; secondly to estimate what information is gained by run-

ning the high resolution forecast beyond that which is contained in these large-scale flow parameters. Comparison of the two forecasting methods using a variety of verification methods all lead to the same conclusion: although both the Bayesian forecast and convection-permitting ensemble provide information about sea-breeze occurrence, the convection-permitting ensemble is significantly more able to discriminate between sea-breeze events and non-events for all lead times.

KEYWORDS
ensemble forecasts, post-processing, sea breeze, information gain, convection-permitting models, verification

1 | INTRODUCTION

In recent years, numerical weather prediction (NWP) centres have started running short-range convection-permitting ensemble prediction systems (CP-EPSs) on limited-area domains with the aim of simulating with fidelity small scale phenomena, including high-impact weather and hence improving local weather forecasts. Some examples include UK Met Office (MOGREPS; Hagelin et al. 2017, Tennant 2015, Golding et al. 2014), the German weather service (DWD) (COSMO-DE-EPS; Gebhardt et al. 2011), Meteo-France (AROME-EPS; Raynaud and Bouttier 2016, Bouttier et al. 2016), Hungarian Meteorological Service (Wang et al., Horányi et al. 2011), the Italian Meteorological Service (COSMO-LEPS run by ARPA SIM; Montani et al. 2011). Although deemed beneficial, CP-EPSs place a heavy burden on the computational resources of forecasting centres. They also produce large amounts of data which needs to be rapidly digested and utilized by operational forecasters. Knowing when and how the convection-permitting ensemble is likely to provide useful additional information is key to successful real-time utilisation of these data. Similarly, knowing where equivalent information can be gained (even if partially) from the global ensemble using statistical/dynamical post-processing both extends lead time (due to faster production) and also potentially provides information in regions where no convection-permitting ensemble is available. There is therefore a good reason to assess CP-EPSs, not only against the "truth" (usually represented by observations), but also against lower-resolution (and thus cheaper) ensemble prediction systems (LR-EPS) to quantify any gain in the skill.

The extent to which CP-EPSs are able to provide better probabilistic forecasts is still not clear for some parameters. Therefore, CP-EPSs should be assessed in a way that is relevant to the purpose for which they have been constructed, which is the prediction of small-scale weather phenomena (Clark et al., 2016; Bowler et al., 2008). There have been many studies on the verification of CP-EPSs (Gilleland, 2017; Ebert et al., 2013, among others), but these have not focussed on information gain when exploiting the coarse resolution models as much as possible. Also, most studies have focussed on conventional variables that are directly represented on the model grid (e.g. surface temperature) and mainly on precipitation accumulation, since it can be evaluated against high resolution radar data and can be used as a proxy for convective storms.

To the best of our knowledge, there is no study focussing on quantifying the added value, relative to their coarser resolution version, of CP-EPS for specific weather phenomena over an extended evaluation period. Clark et al. (2016) reviewed the state-of-the-art convection-permitting models for forecasting rainfall, showing their benefit by considering few a cases studies in Summer 2014 with a qualitative comparison against radar images. Among the studies focussing on precipitation, Schellander-Gorgas et al. (2017) showed the benefit of using CP-EPSs for predicting convective precipitation over a 3-month summer period with

respect to a coarser resolution model. Also, Schwartz et al. (2010) showed the benefit of CP-EPSs for quantitative precipitation forecasts over 35 days of Spring 2007, stating that we can learn much about what type of information can be extracted from convection-allowing model output and how this information might best be used by operational weather forecasters.

With regard to studies comparing CP-EPSs to coarser resolution global models, Marsigli et al. (2008) compared the skill of COSMO-LEPS to Global EPS for precipitation, concluding that the extent to which COSMO-LEPS is more skillful is still not known. Marsigli et al. (2005) quantified the added value of COSMO-LEPS with respect to ECMWF-EPS for 24 h precipitation accumulation, using the Brier Score and the ROC (relative operating characteristic) area. They found that COSMO-LEPS performs better than ECMWF-EPS when comparing maximum precipitation values over boxes of $1.5° \times 1.5°$, whereas the reverse is true when considering the average precipitation values. Beck et al. (2016) have shown the benefit of combining CP-EPSs for gridded variables such as wind, temperature, relative humidity, and precipitation. Gowan et al. (2018) showed the benefit of the NCAR-EPS system (Schwartz et al., 2015) for orographic precipitation over Western USA for one extended winter season (October 2016 - March 2017). Woodhams et al. (2018) focussed on convective precipitation on a sub-daily timescale in the tropics over for a two-year period. They showed, for deterministic forecasts, the added value provided by a convection-permitting model with respect to a coarser resolution global model. Some studies have investigated specific weather phenomena using CP-EPSs. Barrett et al. (2016) verified rainfall against radar observations for four cases of quasi-stationary convective events. Trier et al. (2015) analysed a case study of convection initiation along a central Oklahoma dryline using a 10-member ensemble of convection-permitting simulations, finding that some members accurately reproduced both its timing and approximate location. Others have found useful skill in using CP-EPSs to derive probabilistic forecasts of tornadoes (Gallo et al., 2016; Sobash et al., 2016) and related high impact weather (e.g. strong wind gusts, large hail) (Sobash et al., 2016) based on updraft helicity as a surrogate for rotating thunderstorms.

In this study we aim to determine the degree to which CP-EPSs provide additional information about the occurrence of sea breezes beyond that which could be determined from the larger-scale environment. A sea breeze is a mesoscale circulation caused by the temperature contrast between the land and the sea, i.e. the differential heating due to different heat capacities of land and sea: the land heats up more rapidly than the sea during the day and cools down more rapidly at night. The sea breeze circulation can be explicitly resolved using CP-EPSs. Whereas at the resolution of the global ensembles only the large-scale land–sea temperature contrast that gives rise to sea breeze is well resolved. This is because in the situation when a process occurs at scales not much smaller than the grid length the resolved and unresolved scales are not well separated (Kalnay, 2002). Since sea breezes usually occur on a spatial scale between 10 to 100 km (Lin, 2007), it is not expected that LR-EPSs (with grid spacing larger than 20 km) are able to properly resolve the sea breeze circulation.

Sea breezes are an appealing choice of phenomena for model comparison because they are geographically constrained to initiate at the coast and occur quite frequently in summer months. Also, sea breeze forecasting is important for several reasons: its impact on air quality, since it affects transports of pollutants, health, relief from oppressive hot weather and as a possible trigger for convective storms, especially when interacting with other mesoscale and synoptic-scale flows (Warren et al., 2014; Birch et al., 2015). Furthermore, we can benefit from the fact that there have been lots of studies on the sea breeze, especially in the past 25 years (see Azorin-Molina and Tijm (2011, Table 1) for a list on studies on sea breeze forecasting and Crosman and Horel (2010, Table 1) on sea breeze numerical studies). Miller et al. (2003) state that forecasting sea breezes consists of three main aspects: occurrence, propagation speed and direction, and distance of inland penetration. In this study we focus solely on occurrence. We assess how much more information we can get from running CP-EPS compared with a Bayesian model based on LR-EPS model variables. The Bayesian model should perform better than the LR-EPS alone and is still cheap compared to running a CP-EPS, and therefore provides a much more stringent test of the benefit of the CP-EPSs. Although the sea breeze is locally forced by the temperature contrast, the synoptic-scale flow plays a fundamental role in controlling the initiation and the evolution of the sea breeze itself. There are of course other geographical factors involved, like coastline topography, surface roughness, latitude and season. However Azorin-Molina and Chen (2008) deem the impact of the large-scale flow to be greater than these,

even than the thermal gradient. Hence we focus on the temperature gradient and larger-scale flow in LR-EPS MOGREPS-G. We want to determine the source of sea-breeze predictability. More precisely we are investigating whether sea-breeze occurrence is due to better representation of local factors and sea-breeze dynamics in CP-EPSs or if it can be instead derived solely from the knowledge of the most influential large-scale conditions, namely the synoptic wind and temperature contrast between land and sea contained in the LR-EPS.

The reason for using an ensemble rather than just comparing deterministic forecasts is to get a better understanding of the link between the driving conditions and the sea breeze by increasing the sample size, and to obtain a probabilistic answer that contains information about the inherent uncertainties of the phenomenon, associated with variations of the environmental conditions. Therefore a more appropriate question to answer is "how much more information is the CP-EPS bringing and is this significant ?". The null hypothesis is that the two EPSs give us the same amount of information. Therefore sea breeze occurrence becomes a test, by which this hypothesis can be either rejected or accepted. To examine the information content for comparison, two statistical properties of an EPS are investigated: the reliability and resolution. Then two summary scores are calculated, Brier score and the ignorance score, measuring the ignorance possessed by the forecasts, after verification against the observed outcome.

The rest of the paper is organized into five sections. In § 2 we describe the dataset, in § 3 the method for detecting sea breezes in observations, CP-EPS and a Bayesian method predicting sea breezes from large scale conditions are presented, § 4 contains the result of model inter-comparison and verification against station observations. A summary of our results and conclusions are given in § 5.

2 | DATA SET

In this section the ensemble model data used for the probabilistic prediction and the observations used for the verification are described.

2.1 | Model data

The forecast data used in this paper come from the Met Office Global and Regional Ensemble Prediction System (MOGREPS), which consists of a global ensemble prediction system (MOGREPS-G) and a nested convection-permitting limited-area version (MOGREPS-UK). Both ensembles are constructed using the Met Office Unified Model (Wood et al., 2014; Davies et al., 2005). Descriptions of both configurations, the initial condition perturbations and stochastic physics can be found in (Hagelin et al., 2017),(Tennant, 2015) and (Bowler et al., 2008). These data also form part of the THORPEX Interactive Grand Global Ensemble (TIGGE) database (Swinbank et al., 2016). This database of operational ensemble forecasts from leading NWP centers was set up to enhance the development of probabilistic forecasting of high-impact weather. In 2014 a TIGGE-Limited Area Model (LAM) panel was also established to apply TIGGE concepts to limited-area model ensembles. Both the datasets are available for research and education purposes at the following link (`https://www.ecmwf.int/en/research/projects/tigge`[1]). The MOGREPS-G forecasts were produced 4 times a day, every 6 hours starting from 00 UTC, with 12 members (1 control + 11 perturbed members) on each cycle run to T+174 (7 days, 6 hours), using a horizontal grid spacing of 33 km and 70 vertical levels. MOGREPS-UK has been running routinely since 2012 and has a horizontal grid spacing of 2.2 km. It was run every 6 h up to 36 h , by dynamically downscaling starting from the 3 h forecast of the matching MOGREPS-G ensemble members. This means that the initial conditions for each MOGREPS-UK member were simply interpolated to the high-resolution grid using the 3 h MOGREPS-G forecast at 0300/0900/1500/2100 UTC. Note that this configuration of MOGREPS-UK was operational from 16 January 2013 until 15 March 2016, after which the MOGREPS-UK ensemble was generated by perturbing around the control

[1] Accessed online on 22 February 2016

forecast intialised using a high resolution analysis (Hagelin et al., 2017). Since there was no additional high resolution data assimilation in the period studied here (2013-2015), the information gain from CP-EPS can only come from better representation of the dynamics and/or ancillary fields such as the coastline, topography and the land-use dataset. The CP-EPS is therefore only constrained by observations at scales greater than 33 km through the global initial and boundary conditions. Since we are focussing on sea breeze, we use only summer (JJA) data for the years 2013-2015. Data from some months during this time period were either partially or completely missing from the archive so are not included in this analysis. These are June 2013, June 2014 and 1st-16th of July 2014. Forecasts of sea breeze are made within the time window 0900-2100 UTC, in which the sea breeze is expected to occur. Figure 1 shows a schematic diagram of the different lead times we have considered. For instance, the forecast labelled with "T+12" is the MOGREPS-G forecast issued at 00 UTC (black dots) or the MOGREPS-UK forecast initialised three hours later (red dots). The parameters for the Bayesian forecast are computed at 12 UTC valid time (black squares) Note that the T+36h forecast lead time is excluded from our analysis since the valid "time window" extends beyond it.

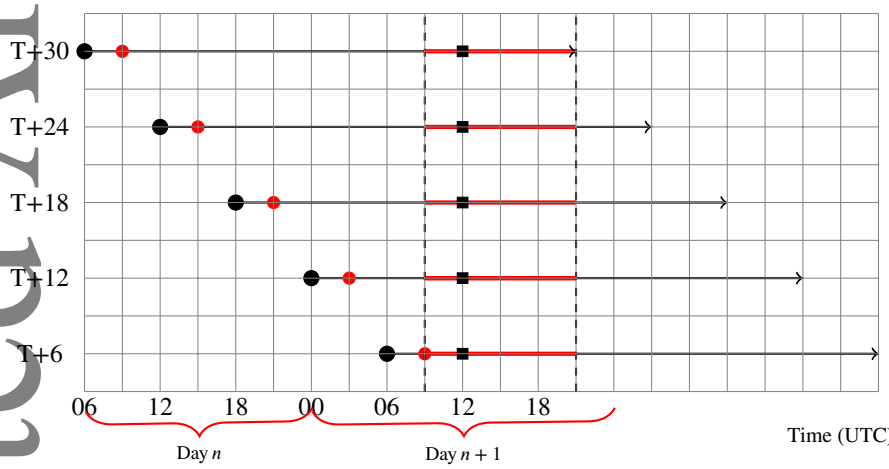


FIGURE 1 Schematic diagram of the forecast lead times. The black (red) dot indicates the MOGREPS-G (MOGREPS-UK) initialization time, whereas the red bar represents the sea-breeze detection window. The black square indicates the validity time at which the MOGREPS-G parameter are computed, for all the different initialization times. The solid black line with arrow covers the total MOGREPS-UK forecast length (36 hours).

2.2 | Observational data

The probabilistic forecasts of sea breeze occurrence from both MOGREPS-UK and the Bayesian forecast generated using MOGREPS-G are verified using hourly observations from four synoptic stations taken from Met Office Integrated Data Archive System (MIDAS) and available at this link: http://archive.ceda.ac.uk/ [2]. Table 1 shows the details of the stations used in this study. Figure 2 shows their location, with the rectangular box indicating the spatial domain where sea breeze has been detected and verified. The spatial sub-domain extends from −1.2 W to 0.2 E in the along the coast direction and from the coast up to 66 km inland (30 MOGREPS-UK grid boxes). In this study only the onshore side of the sea breeze is taken into account, with the

[2] Accessed online on 25 September 2017

aim of detecting the well defined inland frontal boundary. Therefore offshore sea breezes will be excluded from this analysis.

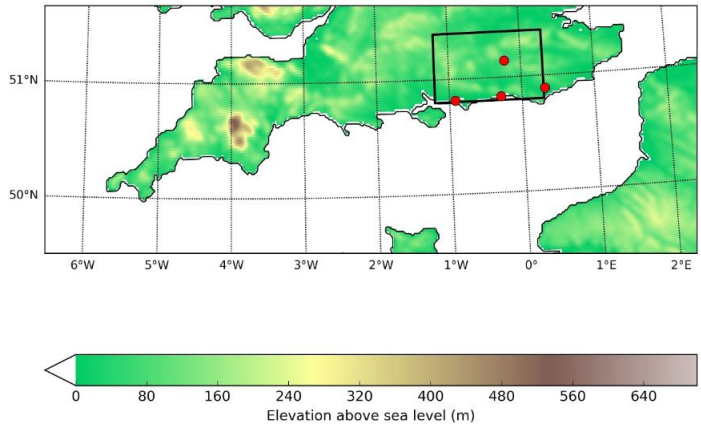


FIGURE 2 A map showing the orography over the south UK domain. Orography data are from MOGREPS-UK. The solid box encloses the subdomain used in this study with red dots indicating weather station observations.

3 | SEA BREEZE IDENTIFICATION AND PREDICTION

WMO number	Name	Location	Distance from the coast (km)
03876	Shoreham Airport	50.8356 N, 0.29194 W	1.3
03872	Thorney Island	50.8142 N, 0.92098 W	1.2
03882	Herstmonceux: West End	50.8904 N, 0.31818 E	9
03769	Charlwood	51.1435 N, 0.22786 W	35

TABLE 1 Information about surface weather stations used for verification. The shortest distance from the coast has been measured. See Figure 2 for their location on a map.

Before looking at the sea breeze occurrence prediction problem over a large sample it is necessary to be able to identify the sea breeze objectively. Many criteria have been developed for the sea breeze identification in observational data, however these are not tailored to automatic detection in high resolution gridded forecast data. These criteria involve meteorological variables which are affected by the sea breeze front passage. In (Azorin-Molina and Tijm, 2011, Table 1), there is a list of the past studies, each with the objective criteria described. Some are directly related to the sea breeze passage (air humidity, air temperature, wind speed and direction), others are used as predictors for the inhibition of sea breeze occurrence (total cloudiness, precipitation). Azorin-Molina and Tijm (2011) point out that these techniques are clearly dependent on the nature of test criteria adopted by each researcher and on the region of interest. They also rely significantly on the thresholds of the meteorological variables used. Coceal et al. (2018) proposed a new method for sea breeze detection in stations observations using fuzzy-logic[3], but still relying

[3] fuzzy-logic is a type of logic where the variable can take any real value between 0 and 1 and not just 0 or 1, as in the classical Boolean logic.

on thresholds of the meteorological variables affected by sea breeze passage. There is need for a more general identification method that makes use of a minimum number of key variables that can in principle be applied to any coastal region of the world, as long as a cross-coastal direction can be found. In this paper we focus on the south coast of UK (see Figure 2). We present two new sea breeze identification algorithms that do not rely significantly on the thresholds of the variables involved, and are useful for the current state of the art CP-EPS data and observations.

3.1 | Sea breeze identification in observations

Sea breezes were identified using the weather station data described above. Three stations have been chosen as they are located within a few kilometers of the coast, where the sea breeze phenomenon is found. The fourth station is inland by 35 km, approximately matching the scale of the MOGREPS-G forecast grid box size. Hourly 10 m wind speed and direction, station air pressure, 2 m dry bulb temperature, and 2 m dew point were used to calculate hourly changes of wind speed and direction, specific humidity and dew point depression. As an example of two cases of sea-breeze days detected in the station observations, Figure 3 shows a clear signal of an early and a relatively late onset of sea breeze front, in that there is a simultaneous shift in wind direction towards being more normal to the coast, an increase in wind speed and specific humidity and a reduction in dew-point depression. In order to construct the verification data set we developed an automatic algorithm for detecting sea breeze in stations observations. The algorithm can be summarized as follows:

1. We select wind direction, wind speed, specific humidity and dew point depression. We then check that the wind changes from having an offshore component to have an onshore component and at the same time the wind speed and specific humidity increase and dew point depression decrease. This is done for each station independently, for the three stations on the coast and the station inland.
2. If sea breeze is detected at least at one coastal station, then we check if it has also been detected inland. Then we require that time of the arrival of the sea breeze front inland is at least two hours after the onset time at the coast. Given that the station inland is approximately 35 km from the coast (see table 1), a 1-hour interval would allow propagation speeds that are too fast and therefore perhaps pick up spurious signals; a 3-hour interval would make identification more difficult because the sea breeze may have propagated well beyond the inland station at the time of measurement, and also gives less guarantee the same feature is being identified. This would imply an approximate propagation speed of $35/2$ km/h $\sim$ 17 km/h.
3. Finally, in order to avoid spurious detections (e.g. cold fronts), we also require that the sea breeze day is also a dry day at the coast, meaning that no rainfall accumulation is recorded in the time window relevant to the sea-breeze passage at the stations near the coast only. Thus the possibility of having convective rain further inland, maybe due to sea-breeze front interacting with other mesoscale and synoptic-scale flow, is not excluded in this analysis. (e.g. the case reported by Pedgley (2003).

3.2 | Sea breeze identification in MOGREPS-UK

In this subsection we will describe the sea breeze identification algorithm using CP-EPS data. The method is designed to identify, at each time step, the position of the leading edge of sea breeze front. Due to the different grid spacing of the observational network ($\sim$ 40 km, see again Figure 2), and the CP-EPS (2.2 km), we have developed a different algorithm, with the same aim to detect the sea-breeze initiation and its propagation inland.

The passage of the sea breeze front is typically marked by a drop in the air temperature, an increase in wind speed, a shift in the wind direction and increase in specific humidity. In Figure 4 the hourly rate of change of air temperature, wind direction and wind speed are shown for all ensemble members from an example T+12 forecast. A sea breeze front is clear in several of the

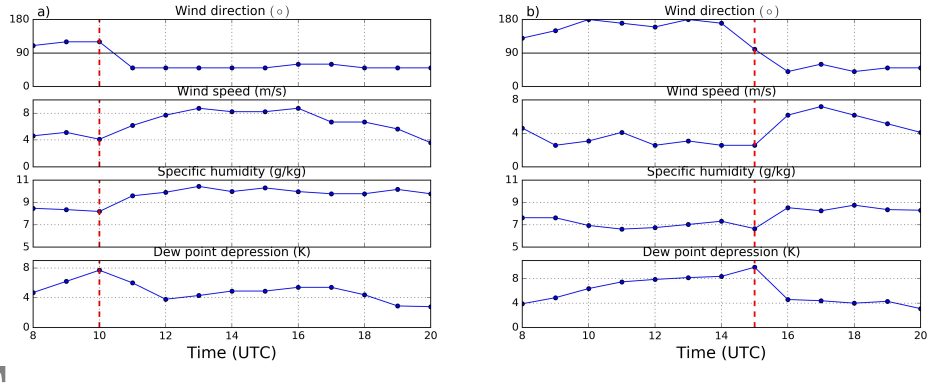


FIGURE 3 Shoreham airport station observation time series on a) 27 June 2015 and b) 23 June 2015. Wind direction is defined such that values of 0, 90 and 180° correspond to onshore, along the shore and offshore flow, respectively. The horizontal solid line indicates the separation between the northerly and southerly direction. Sea breeze onset time is indicated by the vertical dotted red line.

panels (see members 0, 4, 7, 9 in particular) where one can see the extrema of the hourly rate of change of all three variables in an elongated "strip" approximately parallel to the coast. Note also that the changes with respect to the different variables occur in the same location. The sea-breeze signal is less clear in other members, for instance members 1 and 11. Figure 5 shows the same variables at two different time steps for member 0 (the control member). It is clear that two hours later the sea breeze front has progressed further inland. The algorithm for sea-breeze detection in the MOGREPS-UK data is based on an analysis of the spatial patterns of the hourly rate of change in the 2 metre temperature and 10 metre windspeed and direction and then on the tracking of their propagation. A formal mathematical description is contained in supplementary material, but the method can be summarized as follows. We first define an along-coast and cross-coast coordinate. Since we are looking at the south-coast of England, these are taken to be zonal and meridional coordinates of the model's rotated equatorial grid, respectively. At each along-coast point in turn the sea breeze detection is performed as follows:

- We find the location of the minimum in the hourly change of the cosine of wind speed occurring within 30 grid points (66 km) from the coast in the cross-coast direction to identify the wind shift. If this occurs within four grid points (8.8km), and the minimum hourly change in temperature and maximum hourly change in wind speed are also found to be within four grid points of the coast at that time or one hour later, the time is flagged as a potential sea-breeze onset at that along-coast point.
- The next step is to test for propagation of the sea breeze. This done by checking that one hour later the minimum in cosine wind direction, the minimum in temperature change and maximum in wind speed change have moved between 1 and 10 grid points (up to 22 km) further inland. This threshold is different from the one based on observations (17 km/h). This is because the model grid spacing is Δx= 2.2 km and therefore the location of a sharp front at any snapshot is uncertain to at least Δx. The distance travelled between any two time frames is therefore uncertain to $2\Delta x$ and this motivates a model threshold that is larger by $2\Delta x$. Also, we check that the minimum of cosine of wind direction and temperature and maximum in wind speed move coherently in space inland, by requiring that their position is not further than 6 grid points apart. If this is satisfied for at least the subsequent two time steps from the onset time then it is confirmed that the potential sea breeze front propagates inland in a manner consistent with a sea breeze. If it is not satisfied then no sea breeze is detected. In Figure 6 the positions of the extrema hourly rate of change with respect to time are shown for a particular along-coast grid point.

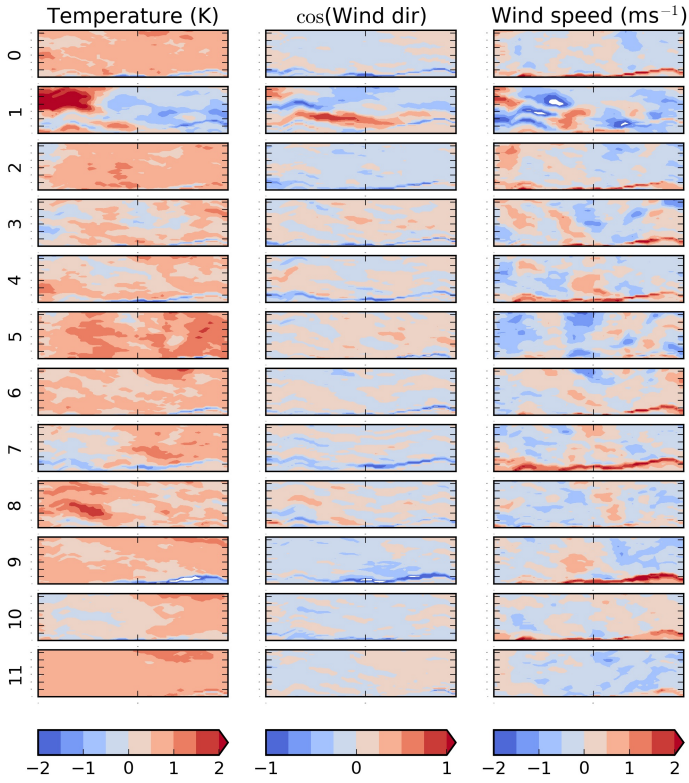


FIGURE 4 Hourly differences of the three variables (2 m temperature, 10 m wind speed and direction) of the 12 ensemble members on 27/06/2015 for the T+12 forecast, calculated between 1100UTC and 1200UTC. The rectangles correspond to the black box in Figure 2. The ticks on the y-axis are every 10 km (from 10 up to 60 km), whereas tick on the x-axis is at 60 km from the left edge of the domain.

- The final step is to test that the potential sea breeze is spatio-temporally coherent in the along-coast direction. This is done by requiring that the potential sea breeze occurs at at least four adjacent along coast points, meaning that it has an extent of at least 8.8km and that the range of the onset times of these coastal points is small (by requiring the standard deviation has to be less than 2 hours). If these criteria are satisfied then the detection of a sea breeze is confirmed. If it is not then no sea breeze is detected.

Sea breeze probabilistic forecast

The sea-breeze detection method described above provides a definition of sea breeze occurrence which is used to create a daily forecast of sea breeze. The detection algorithm is applied to each CP-EPS member over the twelve hour period starting at 9UTC

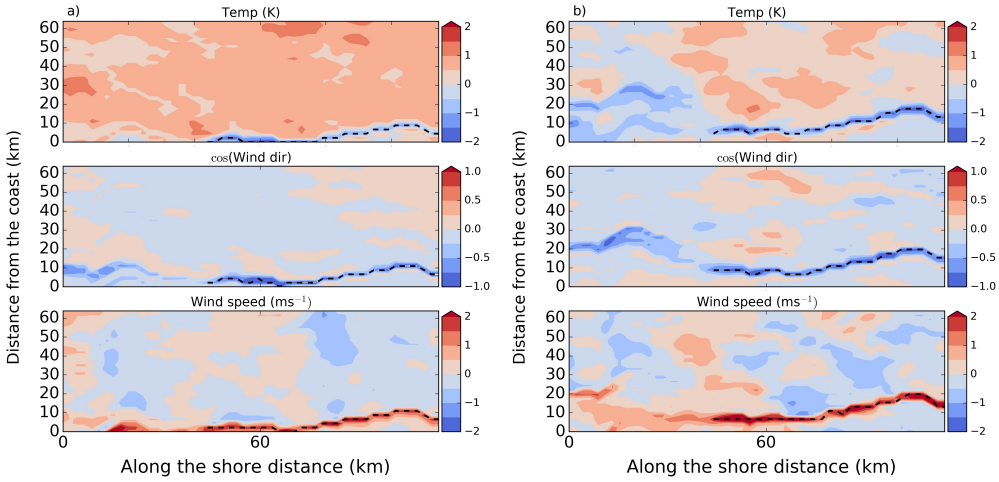


FIGURE 5 Snapshots map of hourly changes of temperature (temp), cosine of wind direction (wind dir) and wind speed for member 0 on 27/06/2015 for a) 11-12 UTC and b) 13-14 UTC. The dashed lines indicate approximately the position of the sea-breeze front for the along the coast grid points where the conditions of the sea-breeze detection algorithm are satisfied. The rectangles, as in Figure 4, correspond to the black box in Figure 2.

valid time. This defines a binary predictand

$$X(m,n) := \begin{cases} 1 & \text{sea breeze occurs,} \\ 0 & \text{otherwise} \end{cases},$$

where m is the ensemble member and n the valid day of the forecast. Using the standard formula for calculating probabilities from ensembles we define the forecast probability of a sea breeze occurring

$$\textbf{HR-DYN}(n) = \frac{1}{M}\sum_{m=1}^{M} X(m,n), \tag{1}$$

where M is the ensemble size (12 in the case of MOGREPS-UK). Due to the small ensemble size, events with very small or very large probability of occurrence are likely to result in forecast probabilities of zero or one. This can be problematic for information-based skill measures. Therefore, as in (Bröcker and Smith, 2008), we substitute the vanishing value of probability with a value of 1/(3M).

3.3 | Sea breeze prediction using MOGREPS-G

In this subsection we describe the methodology used to produce the Bayesian forecast of sea breeze occurrence using MOGREPS-G variables as input. Due to the coarser grid box size and inability to properly represent sea breezes explicitly, the method applied to MOGREPS-UK, described in the previous section, is not appropriate. Therefore we are seeking large-scale predictors of sea breeze that are related directly to the underlying dynamics and resolvable in the global model. Two candidates, as mentioned in

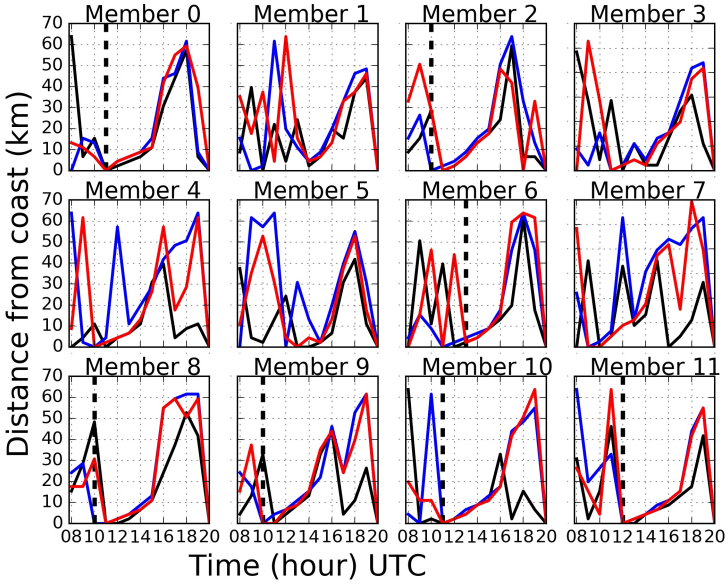


FIGURE 6 Postage stamp plot of the maxima hourly rate of change location against time for temperature (black), wind direction (blue) and wind speed(red) for T+12 MOGREPS-UK forecast on 27/06/2015 and for a particular coastal grid point. The vertical dotted line identifies the members satisfying sea-breeze condition and represents the onset time identified from the model .

§ 1, are the synoptic wind and the temperature contrast between the land and the sea. The temperature contrast is the driving mechanism of the sea breeze, whereas the synoptic wind exerts an important influence on its occurrence. In (Simpson and Britter, 1980) laboratory experiments have been carried out considering the effect of head/tail ambient wind. The effect of the head wind is to reduce the rate of advance of the sea breeze front by about three fifths of the value of the opposing flow. Arritt (1993) examined the effect of the ambient wind on the development of characteristic features of the sea breeze using a two dimensional numerical model. He found that both the direction normal to the coast (offshore or onshore) and the magnitude of the ambient wind exert a strong influence on the sea breeze behaviour. The synoptic regime associated with the inland penetration of the sea breeze consists of either calm conditions or weak ($\lesssim 6\,\mathrm{ms}^{-1}$) opposing (offshore) synoptic flow. With strong ($6-11\ \mathrm{ms}^{-1}$) or very strong ($\gtrsim 11\ \mathrm{ms}^{-1}$) opposing flow the sea breeze does not penetrate inland (remaining on the shoreline) or does not occur at all. Porson et al. (2007) investigated the effect of opposing wind on the sea-breeze inland propagation speed and occurrence. They developed an index for the sea-breeze occurrence, using time-integrated surface heat flux instead of land—sea temperature contrast. There exist two different threshold values of this index for the sea-breeze occurrence, depending on the existence or not of an opposing geostrophic wind. The case with onshore synoptic flow is more subtle. In (Crosman and Horel, 2010, their Table 4), a review of the studies investigating the effect of onshore synoptic flow on sea-breeze circulation is offered. Generally, an onshore flow tends to weaken the sea-breeze circulation, since it leads to frontolysis. For onshore geostrophic wind of $2-4\ \mathrm{ms}^{-1}$ sea breezes become almost indistinguishable from the large-scale flow. The temperature contrast between the land and the sea is needed to create the density and pressure gradient in the direction perpendicular to the shoreline. This is the same mechanism driving the so called *density* or *gravity currents*. There have been many studies showing sea breezes behave like atmospheric density currents (Simpson, 1969; Simpson and Britter, 1980). The comparison between sea breezes and density currents is

based on calculating the ratio between the sea breeze propagation speed and the density current speed, which is referred to in the literature as the *Froude number*.

To formulate a Bayesian forecast of sea breeze using large scale conditions we use as predictors $U|U|$ and c^2. $|U|$ is the magnitude of the cross-coast component of the synoptic scale wind and the predictor $U|U|$ is equal to the square of the cross-coast component of the wind multiplied by its sign (defined to be positive for offshore wind and negative for onshore wind). c^2 is defined as follows:

$$c^2 = \alpha \frac{gH\Delta T}{T}. \tag{2}$$

c represents the density current propagation speed, $\frac{g\Delta T}{T}$ is the reduced gravity (von Kármán, 1940; Keulegan, 1957; Benjamin, 1968), $\Delta T = T_{land} - T_{sea}$, and H the fluid depth, in this case the height of sea breeze front head. α is a density current coefficient to be calculated experimentally. For instance, Atkins and Wakimoto (1997) calculated α for offshore ambient winds cases, finding an average value of 0.7. The results of this paper do not depend on the specific value chosen for α because the predictors are rescaled by their covariance matrix as part of the calculations used to create the forecast[4]. Therefore, for simplicity, we set $\alpha = 1$. In static ambient conditions, c gives directly the speed of inland propagation of the sea breeze (once it has initiated). However, in the presence of offshore winds sea breeze progress is slowed. The effect of the ambient wind is taken into account by the other predictor $U|U|$. Another factor leading to sea breezes propagating slower than the predicted c has been investigated by (Robinson et al., 2013). They found, by means of 2D numerical simulations, that a gradual input of heat from a continental surface, rather than an initially specified density contrast as typical in laboratory analogs, can explain real sea breezes propagating at slower rate than the one predicted by the formula. However, this does not represent a major concern in this study, since we are only interested in the ability of c (which ultimately depends on the temperature contrast), to discriminate between the occurrence and non-occurrence of a sea breeze and not prediction of its propagation speed.

The two predictors are computed from MOGREPS-G at 12 UTC valid time. To compute c^2, defined in Equation 2, we need to find ΔT. For this we use the 2 m temperature from the model, above the land and the sea at a fixed distance from the coast, averaged along the coast direction. For H and T we use standard reference values ($H = 1000$ m and $T = 300$ K), i.e. they are not defined by model variables. Note, as with the choice of α the results do not depend on the specific values chosen for H and T, however values are chosen to be physically realistic to aid physical interpretation of Figure 7. $H = 1000$ has been used for instance by (Miller et al., 2003), whereas Simpson and Britter (1980) used H=700, an empirical mean value over 54 cases. Also, Zhong and Takle (1993) show the sea breeze depth starting at 500-600 m, increasing to 1000 m and then decreasing to 800 m later. For U, we use the model wind output at 850 hPa pressure level (well above the boundary layer height), with the magnitude being averaged over the area of interest and taken with positive or negative sign if the mean direction is offshore or onshore respectively. These two predictors define a two dimensional parameter space. The aim is to calculate the conditional probability of sea breeze occurrence given these predictors. It is therefore natural to apply a Bayesian framework. Firstly we need to calculate the distribution of the occurrences and non-occurrences of sea breeze in this parameter space. This has been done in the past studies using mainly observational data (e.g. (Frysinger and Lindner, 2003)). In this study, in order to train the Bayesian model, we are using MOGREPS-UK data. This is because the main aim of this study is not to produce the most skillful Bayesian forecast, but to examine whether, with only few predictors, it is possible to retrieve the information content of a much higher resolution model from its coarser resolution parent. In particular, given $X(m,n)$ from MOGREPS-UK, we look at the distribution of $\{X(m,n) = 1\}$ (occurrences) and $\{X(m,n) = 0\}$ (non-occurrences) in the parameter space defined by $U|U|(m,n)$ and $c^2(m,n)$ from the corresponding MOGREPS-G member. Scatter plots of $U|U|(m,n)$ and $c^2(m,n)$ for occurrences and non-occurrences are shown in Figures 7a and 7b, respectively. What is immediately apparent from these figures is that the

[4]Only the 1:1 line shown for reference in Figure 7 is affected by the choice of scaling.

distribution of occurrences is concentrated in the *wedge* between the 1:1 line and $c^2 = 0$, defined by $0 < U|U| < c^2$. This is where, through physical reasoning, we would expect conditions to be favourable for sea breezes: the temperature contrast is positive and the opposing ambient wind is not strong enough to prevent its initiation. By contrast non-occurrences are more widely scattered in parameter space. A 2-dimensional Kolmogorov-Smirnov test (Press and Teukolsky, 1988) gives a vanishingly small probability ($\sim 10^{-72}$) that such differences would be observed if the data were drawn from the same distribution. To further

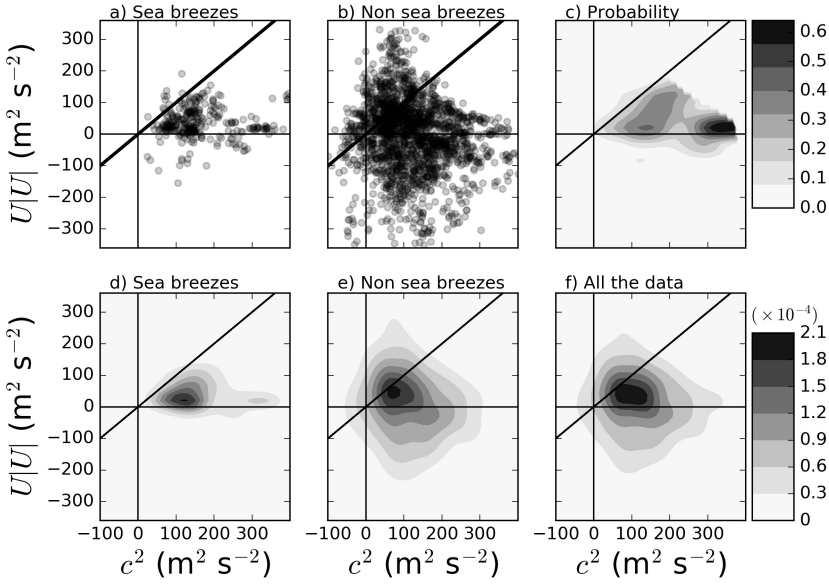


FIGURE 7 Distribution of a) sea breeze occurrences and b) non occurrences in the parameter space defined by c^2 and $U|U|$ for all the years of our study at 12 hours ahead. d), e) and f) show the kernel density estimation for a) and b) and the total distribution respectively. c) represents the conditional probability values $\pi(\theta_{sb}|x)$ in this parameter space. As kernel bandwidth we have used the one proposed by (Scott, 1979). Contour fields are limited to values of the kernel density estimation higher than 10^{-6}. The solid line represents the 1:1 line. A similar plot for the observations is available in supporting information.

quantify the differences between occurrences and non-occurrences, Table 2 is a contingency table showing the number of data points inside and outside the wedge for occurrences and non-occurrences. It is clear that the ratio of points inside the wedge to those outside is much higher when a sea-breeze occurs than when it does not, implying a strong relationship between location within the parameter space and the probability of a sea breeze occurring. A χ^2-test of independence applied to Table 2, gives a vanishingly small p-value $\sim 10^{-73}$, implying there is enough evidence to reject the null hypothesis that the rows and columns are independent and we may therefore conclude there is a relationship between being in the wedge and the occurrence of sea breeze.

	Inside	Outside
Occurrences	260	93
Non-occurrences	509	1544

TABLE 2 Number of data points used in the figure 7 according to the sea-breeze definition and the position in the parameter space. Inside and outside refer to the *wedge*, the region of the parameter space between the x-axis and the $1:1$ line.

In order to calculate the probability density for specific values of the parameters, we apply *Gaussian kernel density estimation* (Parzen, 1962; Rosenblatt, 1956), with the bandwidth defined by Scott's rule (Scott, 1979), to the data. This produces smooth estimates of the probability density functions (PDFs) which can be evaluated at any point in parameter space. As part of the standard kernel density estimation procedure the data are rescaled by their covariance matrix which eliminates any dependence of the smoothed PDF on the choice of scaling applied to the physical parameters (i.e. the values chosen for α, H and T). Figures 7d and 7e show the smoothed equivalents of Figures 7a and 7b, respectively. These two distributions can be used to calculate the probabilistic forecast of sea breeze conditioned on the information given by $x = (c^2, U|U|)$. We define θ_{sb} to be a discrete binary variable, representing the event $X(m, d) = 1$, i.e. the sea breeze has occurred in the MOGREPS-UK forecast. As prior knowledge we have also $\pi(\theta_{sb})$, the climatological probability of sea breeze occurring in the MOGREPS-UK forecasts. The aim is then to update this information, after observing x. Therefore it is natural to calculate the posterior distribution using Bayes' formula:

$$\pi(\theta_{sb}|x) = \frac{f(x|\theta_{sb})\pi(\theta_{sb})}{f(x)}. \tag{3}$$

The unconditional distribution in the data in parameter space (shown in Figure 7f) is given by

$$f(x) = f(x|\theta_{sb})\pi(\theta_{sb}) + f(x|\theta_{nsb})\pi(\theta_{nsb}), \tag{4}$$

where $\pi(\theta_{nsb}) = 1 - \pi(\theta_{sb})$, $f(x|\theta_{sb})$ and $f(x|\theta_{nsb})$ are the PDFs of the occurrences (Figure 7d) and non occurrences (Figure 7e) respectively. In particular $f(x|\theta_{sb})$ is the so called likelihood function. The distribution of $\pi(\theta_{sb}|x)$ in the parameter space is shown in Figure 7c. It can be seen the higher values of probability correspond with high values of c^2 (and so, high values of ΔT) and low positive values of $U|U|$. A second maximum is evident for lower values of c^2. However this is most likely be due to sampling issues rather than multi-modality of the conditional distribution, since a gap in the data can be seen in the scatter plots of both occurrences and non-occurrences (Figures 7a and b, $c^2 \approx 200$ and $U|U| \gtrsim 0$). In order to produce a Bayesian prediction of the sea breeze occurrence θ_{sb} in the MOGREPS-UK forecast using the MOGREPS-G forecast the following steps are performed. The parameters x are calculated for the MOGREPS-G member. The conditional probability of a sea breeze $\pi(\theta_{sb}|x)$ is evaluated at x, using a training dataset of paired MOGREPS-G and MOGREPS-UK forecast members. In the results presented here we use the *leave-p-out* cross validation technique (Shao, 1993; Celisse and Robin, 2008), producing and validating forecasts for one season using training data from the other two seasons. For each MOGREPS-G member in the ensemble we then have a probability of a sea breeze occurrence. The final probabilistic forecast of sea breeze occurrence (hereafter called **LR-BAY**) is then given by ensemble mean probability

$$\textbf{LR-BAY}(n) := \frac{1}{M}\sum_{m=1}^{M} \pi(\theta_{sb}|x(m, n)), \tag{5}$$

where M is the number of ensemble members, 12 in case of MOGREPS-G. Then, similarly with **HR-DYN**, we exclude the probability values of zeros substituting them with the smallest non zero values. The use of a Bayesian model and/or kernel density estimation to calculate conditional probabilities is not new. For instance Hamill and Church (2000) used the Bayesian framework to calculate conditional probabilities of tornadic storms without using a kernel approach, whereas, more recently, Taszarek et al. (2017) used the kernel approach but not Bayes' rule. The scatter plots in Figure 7a,b are essentially identical to Frysinger and Lindner (2003, Fig. 3). In that case the aim was to find a proper threshold of the ratio $\frac{U|U|}{c^2}$, such that for any value below this threshold most of the non occurrences were excluded. By using this kernel approach it is not necessary to find such a threshold, so again it can be *in principle* generalized to any part of the world. In operational contexts, where these CP-EPS data are not available for some reason, coastal station observational data could be used instead. An example of the same plot, using

observations rather than model data is shown in the supporting information file (figure S1). It is worth noting that it gives very similar results to those produced by training using high-resolution data.

4 | COMPARISON AND VERIFICATION

In this section we will show the results of the comparison between the skill of **LR-BAY** and **HR-DYN** forecasts of sea breeze occurrence (first qualitatively and then quantitatively). In what follows, within verification equations, we will denote **HR-DYN**, **LR-BAY** and observations with $\mathbf{h}$, $\mathbf{l}$ and Y respectively.

Figure 8 shows the T+12 time series of forecast probabilities for the summer season 2015. It also shows whether a sea breeze was detected in the verifying observation. The probability values on occasions when a sea breeze is detected in observations are plotted above the zero line, whereas when a sea breeze is not detected in observations the values are plotted below zero line. It can be seen that HR-DYN is a better detector of sea-breeze occurrence, with a greater proportion of blue bars above the 0.5 line with respect to LR-BAY. However, it should be noted that LR-BAY never exceeds 0.5. Therefore in order to make a first comparison we count the number of occasions when sea breeze is observed (or not) and probability values exceed (or not) a threshold. This threshold is chosen to be half of the maximum forecast probability issued by each forecasting method (denoted t_h and t_l). The results are shown in table 3. It can be seen that when a sea breeze is observed the proportion of HR-DYN probabilities exceeding the threshold is $24/53 \approx 0.45$, whereas for LR-BAY it is $30/53 \approx 0.57$. When sea breeze is not observed the proportions are more different. HR-DYN has a proportion of $144/147 \approx 0.98$ of probabilities below threshold, whereas LR-BAY has $106/147 \approx 0.72$. This already suggests HR-DYN and LR-BAY are equally able to discriminate events, but HR-DYN is more able to discriminate non-events. This property will be described and quantified later in the section. Two properties

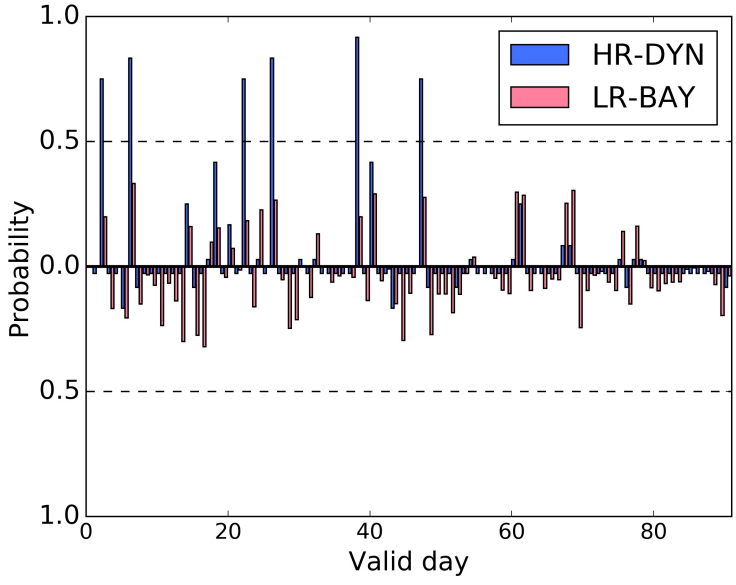


FIGURE 8 Time series of HR-DYN and LR-BAY for T+12 (hours) forecast lead time. The values above the zero line refer to the *observed* sea breeze occurrence, whereas the values below correspond to days when sea breeze has not been detected.

	SB observed	SB not observed
HR-DYN $> t_h$	24	3
HR-DYN $\leq t_h$	29	144
LR-BAY $> t_l$	30	41
LR-BAY $\leq t_l$	23	106

TABLE 3 Number of occurrences and non-occurrences of sea breeze in relation to probability values exceeding half of the maximum probability value. This refers to the whole dataset for the T+12 forecast.

that probabilistic forecasts should have are *reliability* and *resolution*. A qualitative description of these two properties can be found in Murphy (1993). Broadly speaking, reliability measures how close the distribution of the observations, conditional on the forecasts, is to the forecast distribution. Resolution is the ability of a forecast to discriminate between events and non-events.

4.1 | Reliability

The reliability diagram is a tool to visualize the reliability property, by plotting the conditional observed frequencies against the predicted probabilities (see (Murphy and Winkler, 1977) for example). For a reliable system points should lie along the diagonal; meaning, for example over all the occasions in which an 80% probability of an event is forecast, the event should occur 80% of the time. In order to plot the reliability diagrams, the probabilistic forecasts are first partitioned into bins (here we use 4 bins). Then observed frequencies and the mean value of the probabilities forecasts are computed for each bin. Bröcker and Smith (2007) propose a new method to take into account sampling fluctuations in the reliability diagram, since these can lead to misleading evaluation, even for a perfectly reliable forecast. Due to a limited number of verification cases, it is not expected that all the points lie along the diagonal line. Therefore this method helps to assess whether any deviation from the diagonal line is due to small sampling or to an unreliable forecasting system. The method consists of resampling from the original data with replacement forecasts paired with their verifying observation to form a surrogate data set. The observed frequencies of the surrogate data set are then recorded. This procedure is repeated N_r times and the range of the possible surrogate observed frequencies is plotted for each bin. The reliability diagrams for **HR-DYN** and **LR-BAY**, computed with the method described, are shown in Figure 9, for the T+12h forecast. It can be seen that most of the observed frequencies lie within the 90% confidence interval for both forecasting systems. However, in the lowest probability bin they appear to be under-confident; i.e. the probability value is significantly smaller than the observed frequency.

4.2 | Resolution

To visualize the resolution we use the Receiver Operating Characteristics (ROC) curve. This comprises hit rates plotted against false alarm rates for different probability thresholds. A detailed description can be found in the supporting information file. We evaluate the relative performance of the two probabilistic forecasts by calculating the area under the ROC curve (AUC, Egan, 1975; Mason and Graham, 2002) of both and testing the significance of their difference. ROC curves for both the forecasts and for T+12 forecast lead time, are shown in Figure 10a. The diagonal line represents the condition of no resolution. An AUC of 0.5 indicates a forecast with no resolution. For forecast with perfect resolution the the ROC curve would pass through the top left corner (the point (0,1)), and the AUC would be equal to 1. What can be seen immediately in Figure 10a is that the ROC for both forecasts is above the diagonal indicating both have resolution, but that the ROC of the convection-permitting forecast is closer to the the point (0,1) indicating it has greater resolution.

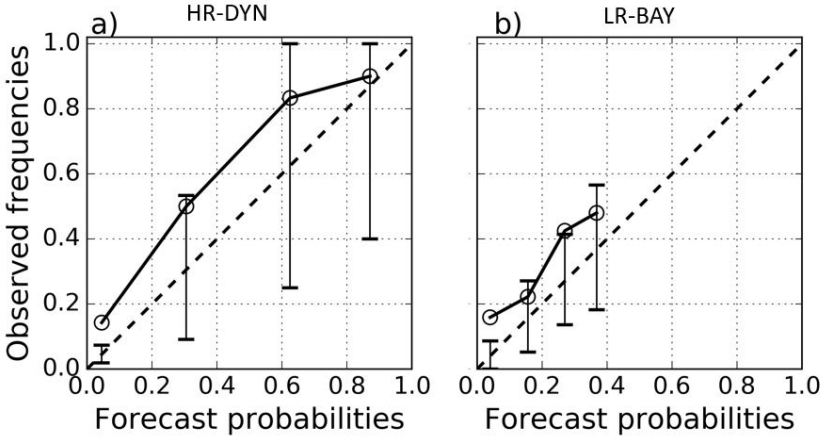


FIGURE 9 Reliability diagram of a) HR-DYN, b) LR-BAY for T+12 h forecast. The dots represent the calculated values from our sample, whereas error bars indicates the 5-95 % confidence interval for the consistency resampling, as described in (Bröcker and Smith, 2007).

Figure 10b shows the evolution of the AUC with lead time for both the forecasts. It can be seen that the convective-scale forecast has a greater AUC than the Bayesian forecast for all lead times. It can be also noted that at 6 hours ahead **HR-DYN** has lower AUC than at other lead times. This could be due to the initial spin-up period of the convective-scale model required to develop small-scale structures. To test whether difference between the AUC for the Bayesian and convection-permitting forecasts is significant we use the following statistic:

$$\frac{\mathrm{AUC}_h - \mathrm{AUC}_l}{\sqrt{\mathrm{Var}(\mathrm{AUC}_h - \mathrm{AUC}_l)}}, \tag{6}$$

where

$$\mathrm{Var}(\mathrm{AUC}_h - \mathrm{AUC}_l) = \mathrm{Var}(\mathrm{AUC}_h) + \mathrm{Var}(\mathrm{AUC}_l) - 2\mathrm{Cov}(\mathrm{AUC}_h, \mathrm{AUC}_l). \tag{7}$$

The subscripts h and l refer to **HR-DYN** and **LR-BAY** respectively.
For the variance and covariance in Equation 7 we have used expressions by (DeLong et al., 1988). The results of the significance testing of the difference between the two areas under the ROC curve are shown in Figure 10c. It can be seen that the difference in AUC, $\mathrm{AUC}_h - \mathrm{AUC}_l 0$, is positive at all lead times. Furthermore the 90% confidence intervals, computed using Equation 6, indicate that this difference is significant for all lead times. This suggests that **HR-DYN** is more skillful than **LR-BAY** in discriminating between events and non-events, since it has greater resolution.

4.3 | Summary scores

In this subsection we compute two probabilistic scores, namely the Brier Score (BS: (Brier, 1950)) and Ignorance score (IS: (Roulston and Smith, 2002)). First we compute the classic decomposition of the scores into reliability and resolution terms, to examine whether the conclusions drawn in the previous subsections about reliability and resolution can be confirmed. Then we

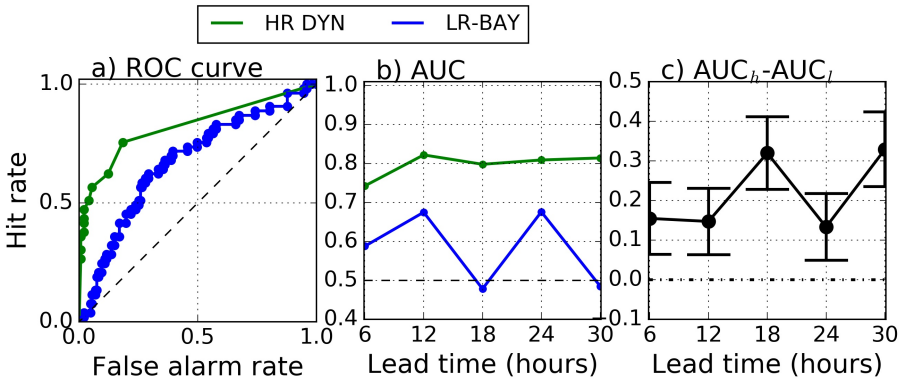


FIGURE 10 a) ROC curve for **HR-DYN** and **LR-BAY** at 12 hours lead time, b) Area under the ROC curve against forecast lead time and c) difference between the two areas. Positive (negative) values indicate the area for the HR-DYN (LR-BAY) forecast is greater. The error bars show the 5-95 % confidence interval based on Equation 6 .

compute the differences in the scores for each forecast relative to the climatological forecast. The scores and relative differences (BSD and ISD) with respect to a *reference* (ref) system are defined as follows:

$$\mathrm{BS}_p = \frac{1}{N}\sum_{n=1}^{N}(p(n) - Y(n))^2 \tag{8}$$

$$\mathrm{IS}_p = \frac{1}{N}\sum_{n=1}^{N} D_{KL}(p(n), Y(n)) \tag{9}$$

$$\mathrm{BSD} = \mathrm{BS}_{ref} - \mathrm{BS}_p \tag{10}$$

$$\mathrm{ISD} = \mathrm{IS}_{ref} - \mathrm{IS}_p. \tag{11}$$

Here $p(n)$ is the probabilistic forecast to be evaluated, $Y(n)$ is the observation on day n, N is total length of the dataset. $D_{KL}(p, Y) = Y \log\left(\frac{Y}{p}\right) + (1 - Y)\log\left(\frac{1-Y}{1-p}\right)$ is the Kullback-Leibler divergence (Kullback and Leibler, 1951). BS is a mean-squared error and measures the distance between forecasts and observations. IS measures the ignorance possessed (or the information content) by a forecast after the verification against the true outcome. The amount of ignorance decreases monotonically with increasing probability value. [5] Both IS and BS are negatively oriented, which means the lower the score the more skillful the forecast. Therefore BSD> 0 or ISD> 0 implies that the convective-scale forecast is more skillful than the reference system, whereas if BSD < 0 or ISD< 0 the reference system is more skillful. Both BS and ISDN can be decomposed

[5] For instance if $p(n) \to 1$ and $Y(n) = 1$ (the event actually occurs), then IS $\to 0$.

into reliability (REL), resolution (RES) and uncertainty (UNC) components (Bröcker, 2009; Weijs et al., 2010) as follows:

$$\mathrm{BS} = \mathrm{BS}_{REL} - \mathrm{BS}_{RES} + \mathrm{BS}_{UNC} \tag{12}$$

$$= \frac{1}{N}\sum_{k=1}^{K} n_k(p(k) - \overline{Y_k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{Y_k} - \overline{Y})^2 + \overline{Y}(1-\overline{Y}) \tag{13}$$

$$\mathrm{IS} = \mathrm{IS}_{REL} - \mathrm{IS}_{RES} + \mathrm{IS}_{UNC} \tag{14}$$

$$= \frac{1}{N}\sum_{k=1}^{K} n_k D_{KL}(p(k), \overline{Y_k}) - \frac{1}{N}\sum_{k=1}^{K} n_k D_{KL}(\overline{Y_k}, \overline{Y}) - \left(\overline{Y}\log(\overline{Y}) + (1-\overline{Y})\log(1-\overline{Y})\right). \tag{15}$$

Here, K is the number of categories (bins) which the probabilistic distribution is divided into and $p(k)$ is the value of each category k. n_k is the number of occurrences and $\overline{Y_k}$ is the observed frequency for each category, respectively. $\overline{Y} = \frac{1}{N}\sum_{n=1}^{N} Y(n)$ is simply the observed frequency over the entire data set. In the calculation of the decomposition and the score differences we have assumed no serial correlation. This is a reasonable assumption, since we are verifying sea breeze occurrence every 24 hours, which is longer than the typical sea breeze time scale. This assumption allows us to neglect correlations when calculating the standard error of the differences with the bootstrap method.

The results of the decomposition are shown in Figures 11a–11d. Recalling that for an ideal forecast REL = 0 and RES > 0, it can be seen that both the forecasts have reliability significantly different from 0 but have significant resolution, for all lead times. It is also clear, that the resolution term for the convective-scale forecast is greater than the Bayesian one although not significantly so at all lead times. This implies that the Bayesian forecast is less able than the convective-scale forecast to discriminate between events and non-events but more able to do so than a climatological forecast.
Figures 11e and 11f show the differences between the BS and IS scores obtained for the two forecasting methods, using Equations 10 and 11 with **LR-BAY** taken as reference and **HR-DYN** as p. The results are shown in Figure 11e, f. It can be seen that for both ISD and BSD the difference in the score is significant (except for ISD at 6 hours ahead), implying that the convective-scale forecast has more value than the Bayesian forecast for sea-breeze occurrence.

After comparing and quantifying the skill of the convection-permitting forecast with respect to the Bayesian forecast, it is interesting to assess how both the systems perform with respect to a third reference, the climatological forecast. By computing the resolution terms in the previous section, we have already shown that both the forecasts have significantly more resolution than the climatological forecast but less reliability. Therefore, this further comparison is useful to assess whether the probabilistic forecasts are performing better than the climatological forecasts, or the lack of reliability is masking their additional resolution. The climatological forecast, i.e. the percentage of times sea breeze occurs over all the data, has been calculated using either observations or convection-permitting model data, leaving one-year-out for the verification itself (for instance, if we are validating on 2013 then 2013 is not included in the calculation of the climatology). The difference in the skill has been quantified using the Brier score and ignorance score, using Equations 10 and 11, with climatology as a reference, in place of LR-BAY. The results are shown in Figure 12. In terms of Brier score difference, the convective-scale forecast is significantly more skillful than the climatological forecast (based either on model or observations), for all lead times, except 6 hours ahead. On the other hand, the Bayesian forecast is shown to be generally less skillful than climatology, although significantly only at 18 hours ahead. With regard to the ISD, the conclusions are essentially same. The convection-permitting forecast is still more skillful than the climatology (although significantly only compared to model climatology for T+12, T+24 and T+30) and the Bayesian forecast significantly worse than the climatology not only at 18 hours ahead, but also at 30 hours ahead. These slight differences between ISD and BSD could be due to the fact that the two scores assign rewards and penalties to the forecast in a different way, because

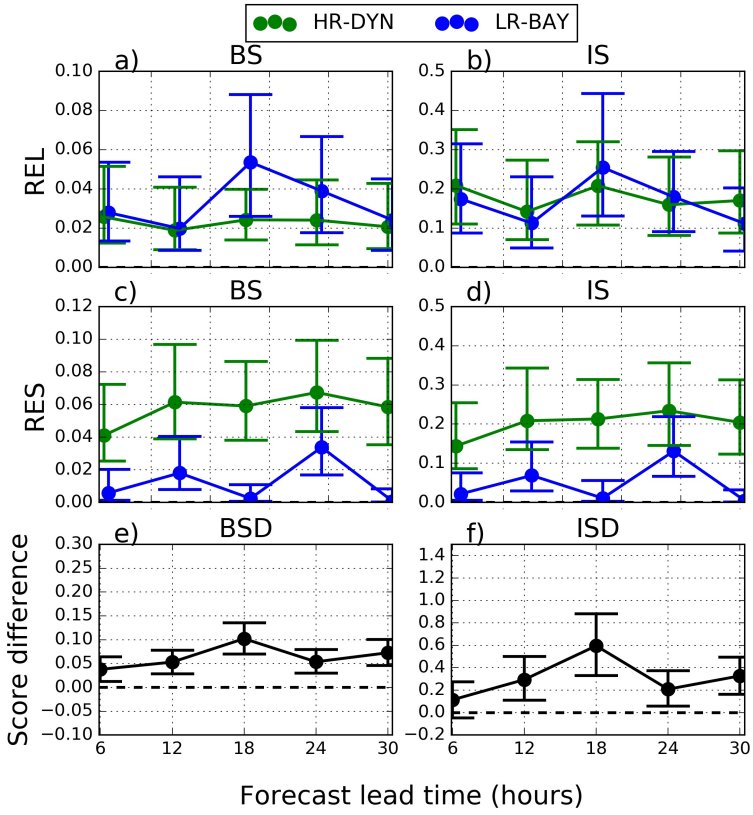


FIGURE 11 Reliability (a) and b)) and resolution (c) and d)) terms calculated for both the forecasts calculated using Brier Score (BS) and ignorance score (IS). e) and f) are the Brier and ignorance score differences respectively based on Equations 10 and 11. Error bars represent the 90% bootstrap confidence interval.

of the different function they are based on. Roulston and Smith (2002) show that IS penalizes sharp [6] forecasts much more than BS. Since even the Bayesian forecast is sharp compared to a climatological forecast, it is likely that this is the reason for the differences in the ISD and BSD measures. Although the Bayesian forecast has more resolution than a climatological forecast it is not sufficient to compensate for its lack of reliability.

5 | CONCLUSIONS AND FUTURE WORK

We have presented work to quantify the added value provided by a convection-permitting ensemble forecast of sea breeze occurrence compared to a simple Bayesian forecast driven by the same global model ensemble. The sea breeze forecasts have been verified over three different summer periods (2013-2015), for lead times from 6 hours up to 30 hours. Extracting the right information from both the models is the first crucial step in the prediction and evaluation process. In § 3 we have shown how the sea breeze signal has been extracted from both the models and then used to produce probabilistic forecasts.

[6] Sharpness is the tendency of a forecast to predict extreme values (0 or 1).

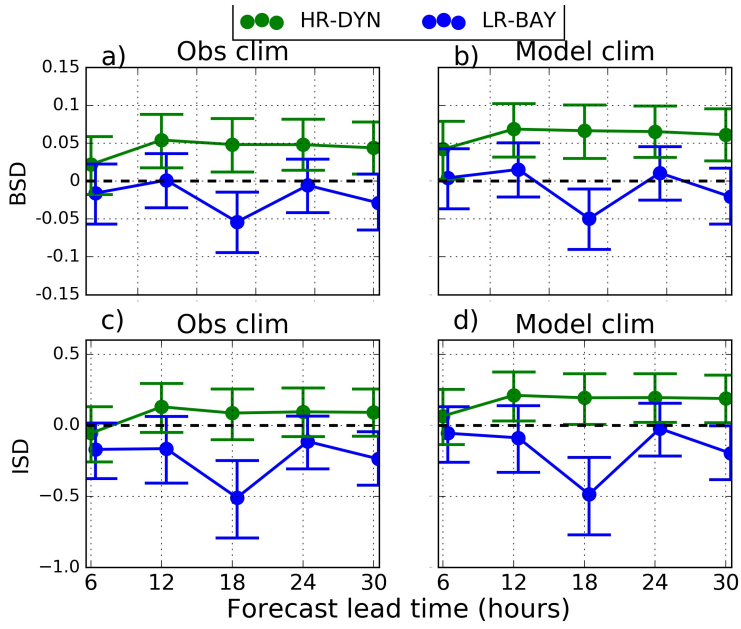


FIGURE 12 Score differences between each probabilistic forecast and a climatological forecast (estimated either from the observations or the model). a) and b) show these differences calculated using the Brier score, whereas c) and d) show the score differences calculated with the . Positive (negative) differences indicate that the probabilistic forecast is more (less) skillful than the climatological forecast. Error bars represent the 90 % confidence interval calculated using bootstrap technique.

The methodologies used to detect a sea breeze are different for the two models and for the observations, due to different grid box sizes and to the low density of the observation network respectively. More precisely, for the convection-permitting ensemble MOGREPS-UK the probability can be extracted from the explicit representation of the sea breeze on the model grid. For the coarser-resolution ensemble (MOGREPS-G), a Bayesian probability of detection is calculated from variables describing the larger-scale meteorological conditions that favour sea breeze formation. For the observation sites a time series of key meteorological variables is used. The second step consisted of translating this information into probabilistic forecasts, whose value was then assessed against station observations. The comparison of the two forecasts against the observational dataset leads us to the following conclusions:

1. The two large-scale predictors used for the Bayesian model are shown to contain more information (in terms of resolution) about the sea breeze occurrence predictability than a climatological forecast. The convective-scale forecast has been shown, however, to provide more information than the Bayesian forecast. This could be due to the convective-scale model's ability to explicitly represent the sea-breeze circulation and fine-scale topography which affects its dynamics. The results show that sea breeze occurrence depends on more than two large-scale predictors and that by representing dynamics explicitly the high-resolution model can extract additional predictive information from the large scale.
2. The convection-permitting model is significantly more skillful than the Bayesian model, with respect to Brier score and ignorance score. The classic score decomposition in reliability and resolution highlights that the convection-permitting model is more reliable and has more resolution for all lead times.
3. The convection-permitting model has significant skill also with respect to the climatological forecast. On the other hand, the

additional resolution possessed by the Bayesian forecast with respect to climatology is not sufficient to compensate for its lower reliability.

4. Forecast skill does not decrease with lead time, instead showing a dependence on the initialisation time. A negligible dependence of the forecast skill on the lead time was found also by Woodhams et al. (2018) for short-range convective precipitation forecasts. In this case, the T+18h forecast performs relatively poorly compared to other lead times. To ascertain the precise reasons for this would need more investigation. However, it is worth noting that forecasts of different lead times are initialized at different initial times (see again Figure 1). The T+18h forecast is initialised at 18UTC for the coarse resolution global model and 21UTC for the convection-permitting model. This might indicate that the relatively lower skill at T+18h is due to different observations being available for assimilation at different times of day (e.g. radiosondes ascents occur only at 00 UTC and 12 UTC).

It is important to perform verification of forecasts of weather phenomena, especially the ones leading to high impact. These phenomena need to be identified first and the necessity to have a large enough sample size necessitates the development of automatic methods to detect them. Such methods naturally come with some limitations, due to the difficulty of detecting with 100% accuracy. It is possible that our methods could miss some weaker sea breezes, or fail to distinguish a sea breeze from another phenomena (e.g. convective cold pools). It is also possible that some observed sea breezes could be missed when there are only a few observation sites. This study offers some insights into how ensemble forecasts (at convective-scale and coarser resolution) can be exploited to extract (and use) coherent meteorological phenomena over multiple grid squares, to provide information for users, from the forecasters to the general public. The Bayesian model can be used as a computationally cheap tool to produce probabilistic forecast of sea breeze occurrence. Its performance could be improved, for instance, by training it on a much larger training data set, using not only model data but also observations. Here, as previously stated, the Bayesian is trained on model data in order to test how much of the information content of much higher resolution forecast comes purely from simple large scale constraints. To conclude, it would be interesting to apply this methodology to other coastal locations around the world with a similar setup of the numerical models or to a severe weather phenomenon (e.g. wind gusts, aircraft turbulence), which emerges from dynamics on smaller scales than those resolved by global models, but is still slaved to the larger-scale environment and dynamics.

ACKNOWLEDGEMENTS

CC was funded by the Engineering and Physical Sciences Research Council Mathematics of Planet Earth Centre for Doctoral Training (grant reference EP/L016613/1) and by the NERC National Centre for Earth Observation.
The station observation data were provided by Met Office Integrated Data Archive System (MIDAS) Land and Marine Surface Stations Data.
This paper benefitted much from the input of three anonymous reviewers.

SUPPORTING INFORMATION

The following supporting information is available as part of the online article:
Appendix S1. Sea breeze detection algorithm using convection-permitting models and observations.
Figure S1. Same as Figure 7 with observational data.
Appendix S2 Equivalence between ROC curves and resolution property.

CONFLICT OF INTEREST

None

ENDNOTES

REFERENCES

Arritt, R. W. (1993) Effects of the Large-Scale flow on Characteristics Features of the Sea Breeze. *J. Appl. Meteorol.*, **32**, 116 – 125.

Atkins, N. T. and Wakimoto, R. M. (1997) Influence of the synoptic-scale flow on sea breezes observed during CAPE. *Monthly Weather Review*, **125**, 2112–2130.

Azorin-Molina, C. and Chen, D. (2008) A climatological study of the influence of synoptic-scale flows on sea breeze evolution in the Bay of Alicante (Spain). *Theor. Appl. Climatol.*, **96**, 249 – 260.

Azorin-Molina, C. and Tijm, D. C. S. (2011) Development of selection algorithms and databases for sea breeze studies. *Theoretical Applied Meteorology*, **106**, 531 – 546.

Barrett, A. I., Gray, S. L., Kirshbaum, D. J., Roberts, N. M., Schultz, D. M. and Fairman, J. G. (2016) The utility of convection-permitting ensembles for the prediction of stationary convective bands. *Monthly Weather Review*, **144**, 1093–1114.

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

Benjamin, T. B. (1968) Gravity currents and related phenomena. *Journal of Fluid Mechanics*, **31**, 209–248.

Birch, C. E., Roberts, M. J., Garcia-Carreras, L., Ackerley, D., Reeder, M. J., Lock, A. P. and Schiemann, R. (2015) Sea-breeze dynamics and convection initiation: The influence of convective parameterization in weather and climate model biases. *Journal of Climate*, **28**, 8093–8108.

Bouttier, F., Raynaud, L., Nuissier, O. and Benjamin, M. (2016) Sensitivity of the AROME ensemble to initial and surface perturbations during hymex. *Quarterly Journal of the Royal Meteorological Society*, **142**, 390–403.

Bowler, N., Arribas, A., Mylne, K. R., Robertson, K. and Beare, S. E. (2008) The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **134**, 703 – 722.

Bröcker, J. (2009) Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, **135**, 1512–1519.

Brier, G. W. (1950) Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Bröcker, J. and Smith, L. A. (2007) Increasing the reliability of reliability diagrams. *Weather and Forecasting*, **22**, 651–661.

— (2008) From ensemble forecasts to predictive distribution functions. *Tellus Series A: Dynamic Meteorology and Oceanography*, **60**, 663 – 678.

Celisse, A. and Robin, S. (2008) Nonparametric density estimation by exact leave-*p*-out cross-validation. *Comput. Statist. Data Anal.*, **52**, 2350–2368. URL: `https://doi.org/10.1016/j.csda.2007.10.002`.

Clark, P., Roberts, N., Lean, H. S. P., Ballard, S. P. C. and Charlton-Perez, C. (2016) Review: Convection-permitting models: a step-change in rainfall forecasting. *Meteorol. Appl.*, **23**, 165 – 181.

Coceal, O., Bohnenstengel, S. I. and Kotthaus, S. (2018) Detection of sea-breeze events around London using a fuzzy-logic algorithm. *Atmospheric Science Letters*.

Crosman, E. T. and Horel, J. D. (2010) Sea and lake breezes: A review of numerical studies. *Boundary-Layer Meteorology*, **137**, 1–29.

Davies, T., Cullen, M. J. P., Malcom, A. J., Mawson, M. H., Staniforth, A., White, A. A. and Wood, N. (2005) A new dynamical core for the Met Office's global and regional modelling of the atmosphere. *Q. J. R. Meteorolog. Soc.*, **131**, 1759 – 1782.

DeLong, E., DeLong, D. and Clarke-Pearson, D. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837—845. URL: https://doi.org/10.2307/2531595.

Ebert, E., Wilson, L., Weigel, A., Mittermaier, M., Gill, P. N. P. P., Göber, M., Joslyn, S., Brown, B., Fowler, T. and Watkins, A. (2013) Progress and challenges in forecast verification. *Meteorological Applications*, **20**, 130–139.

Egan, J. P. (1975) *Signal detection theory and ROC-analysis*. New York : Academic Press.

Frysinger, J. R. and Lindner, D. B. L. (2003) A Statistical Sea-Breeze Prediction Algorithm for Charleston, South Carolina. *Weather and Forecasting*, **18**, 614 – 625.

Gallo, B. T., Clark, A. J. and Dembek, S. R. (2016) Forecasting tornadoes using convection-permitting ensembles. *Weather and Forecasting*, **31**, 273–295.

Gebhardt, C., Theis, S., Paulat, M. and Bouallègue, Z. B. (2011) Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research*, **100**, 168 – 177. URL: http://www.sciencedirect.com/science/article/pii/S0169809510003455.

Gilleland, E. (2017) A new characterization within the spatial verification framework for false alarms, misses, and overall patterns. *Weather and Forecasting*, **32**, 187–198. URL: https://doi.org/10.1175/WAF-D-16-0134.1.

Golding, B. W., Ballard, S., Mylne, K., Roberts, N., Saulter, A., Wilson, C., Agnew, P., Davis, L. S., Trice, J., Jones, C., Simonin, D., Z, L., Pierce, C., Bennett, A., Weeks, M. and Moseley, S. (2014) Forecasting Capabilities for the London 2012 Olympics. *Bull. Amer. Meteor. Soc*, **95**, 883 – 896.

Gowan, T. M., Steenburgh, W. J. and Schwartz, C. S. (2018) Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the Western United States. *Weather and Forecasting*, **33**, 739–765.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W. (2017) The Met Office convective-scale ensemble, MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2846–2861.

Hamill, T. M. and Church, A. T. (2000) Conditional Probabilities of Significant Tornadoes from RUC-2 Forecasts. *Weather and Forecasting*, **15**, 461 – 475.

Horànyi, A., Mile, M. and Szucs, M. (2011) Latest developments around the ALADIN operational short-range ensemble prediction system in Hungary. *Tellus A: Dynamic Meteorology and Oceanography*, **63**, 642–651.

Kalnay, E. (2002) *Atmospheric modelling, data assimilation and predictability*. Cambridge University Press.

von Kármán, T. (1940) The engineer grapples with nonlinear problems. *Bull. Amer. Math. Soc.*, **46**, 615–683.

Keulegan, G. H. (1957) An experimental study of the motion of saline water from locks into fresh water channels. *13th Progress Report on Model Laws for Density Currents, Rep. 5168*.

Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *Ann. Math. Statist.*, **22**, 79–86. URL: https://doi.org/10.1214/aoms/1177729694.

Lin, L.-Y. (2007) *Mesoscale Dynamics*. Cambridge University Press.

Marsigli, C., Boccanera, F., Montani, A. and Paccagnella, T. (2005) The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlinear Processes in Geophysics*, **12**, 527–536.

Marsigli, C., Montani, A. and Paccangnella, T. (2008) A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Meteorological Applications*, **15**, 125–143.

Mason, S. J. and Graham, N. E. (2002) Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, **128**, 2145–2166.

Miller, S. T. K., Kelm, B. D., Talbot, R. W. and Mao, H. (2003) Sea breeze: structure, forecasting and impacts. *Rev. Geophys.*, **41**, 1/1–31.

Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T. (2011) Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A*, **63**, 605–624.

Murphy, A. H. (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–293.

Murphy, A. H. and Winkler, R. L. (1977) Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 41–47.

Parzen, E. (1962) On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**, 1065–1076.

Pedgley, D. E. (2003) The Bracknell hailstorm of 7 May 2000. *Weather*, **58**, 171–182.

Porson, A., Steyn, D. G. and Schayes, G. (2007) Formulation of an index for sea breezes in opposing winds. *Journal of Applied Meteorology and Climatology*, **46**, 1257–1263.

Press, W. H. and Teukolsky, S. A. (1988) Kolmogorov-smirnov test for two-dimensional data. *Computers in Physics*, **2**, 74–77.

Raynaud, L. and Bouttier, F. (2016) Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quarterly Journal of the Royal Meteorological Society*, **142**, 854–866.

Robinson, F. J., Patterson, M. D. and Sherwood, S. C. (2013) A numerical modeling study of the propagation of idealized sea-breeze density currents. *Journal of the Atmospheric Sciences*, **70**, 653–668.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

Roulston, M. S. and Smith, L. A. (2002) Evaluating Probabilistic Forecasts Using Information Theory. *Mon. Weather Rev.*, **130**, 1653 – 1660.

Schellander-Gorgas, T., Wang, Y., Meier, F., Weidle, F., Wittmann, C. and Kann, A. (2017) On the forecast skill of a convection-permitting ensemble. *Geoscientific Model Development*, **10**, 35–56.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C. and Wandishin, M. S. (2010) Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather and Forecasting*, **25**, 263–280.

Schwartz, C. S., Romine, G. S., Sobash, R. A., Fossell, K. R. and Weisman, M. L. (2015) NCAR's experimental real-time convection-allowing ensemble prediction system. *Weather and Forecasting*, **30**, 1645–1654.

Scott, D. W. (1979) On optimal and data-based histograms. *Biometrika*, **66**, 605 – 610.

Shao, J. (1993) Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **88**, 486–494.

Simpson, J. and Britter, R. (1980) A laboratory model of an atmospheric mesofront. *Q. J. R. Met. Soc.*, **106**, 485 – 500.

Simpson, J. E. (1969) A comparison between laboratory and atmospheric density currents. *Q. J. R. Met. Soc.*, **95**, 758 – 765.

Sobash, R. A., Schwartz, C. S., Romine, G. S., Fossell, K. R. and Weisman, M. L. (2016) Severe weather prediction using storm surrogates from an ensemble forecasting system. *Weather and Forecasting*, **31**, 255–271.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H. A., Wilson, L. and Yamaguchi, M. (2016) The TIGGE Project and its achievements. *Bull. Amer. Meteor. Soc*, **97**, 49 – 67.

Taszarek, M., Brooks, H. E. and Czernecki, B. (2017) Sounding-derived parameters associated with convective hazards in europe. *Monthly Weather Review*, **145**, 1511–1528.

Tennant, W. (2015) Improving initial condition perturbations for MOGREPS-UK. *Weather and Forecasting*, **141**, 2324 – 2336.

Trier, S. B., Romine, G. S., Ahijevych, D. A., Trapp, R. J., Schumacher, R. S., Coniglio, M. C. and Stensrud, D. J. (2015) Mesoscale thermodynamic influences on convection initiation near a surface dryline in a convection-permitting ensemble. *Monthly Weather Review*, **143**, 3726–3753.

Wang, Y., Martin, B., Christoph, W., Martin, S., Florian, W., Alexander, K., Stjepan, I., Weihong, T., Xulin, M., Simona, T. and Eric, B. () The Central European limited-area ensemble forecasting system: ALADIN-LAEF. *Quarterly Journal of the Royal Meteorological Society*, **137**, 483–502.

Warren, R. A., J., K. D., S., P. R. and W., L. H. (2014) A 'Boscastle-type' quasi-stationary convective system over the UK southwest peninsula. *Quarterly Journal of the Royal Meteorological Society*, **140**, 240–257.

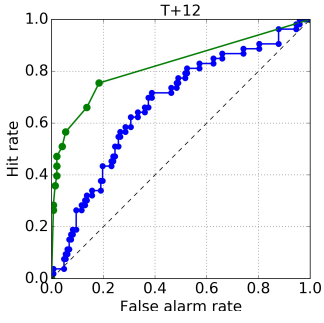
Weijs, S. V., van Nooijen, R. and van de Giesen, N. (2010) Kullback–leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, **138**, 3387–3399.

Wood, N., Staniforth, A., White, A., Allen, T., Diamantakis, M., Gross, M., Melvin, T., Smith, C., Vosper, S., Zerroukat, M. and Thuburn, J. (2014) An inherently mass-conserving semi-implicit semi-lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1505–1520.

Woodhams, B., Birch, C., Marsham, J., Bain, C., Roberts, N. and Boyd, D. (2018) What is the added-value of a convection-permitting model for forecasting extreme rainfall over tropical East Africa? *Monthly Weather Review*.

Zhong, S. and Takle, E. S. (1993) The effects of large-scale winds on the sea land-breeze circulations in an area of complex coastal heating. *Journal of Applied Meteorology*, **32**, 1181–1195.

GRAPHICAL ABSTRACT



The added value of probabilistic forecast of sea breeze occurrence derived from a convection-permitting ensemble (green) relative to a Bayesian forecast (blue) conditioned on coarser resolution predictors is quantified.
In particular we examined reliability and resolution properties of the two probabilistic forecast to assess whether the convective-scale one can provide additional information.