# ~~Objective identification~~ Identification of high-wind features within extratropical cyclones using a probabilistic random forest~~(RAMEFI)~~. Part I: Method and ~~illustrative~~ case studies

Lea Eisenstein[1], Benedikt Schulz[2], Ghulam A. Qadir[3], Joaquim G. Pinto[1], and Peter Knippertz[1]

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

**Correspondence:** Lea Eisenstein (lea.eisenstein@kit.edu)

**Abstract.** Strong winds associated with extratropical cyclones are one of the most dangerous natural hazards in Europe. These high winds are mostly associated with five mesoscale dynamical features, namely the warm (conveyor belt) jet (WJ), the cold (conveyor belt) jet (CJ), cold-frontal convective gusts (CFC), strong cold sector winds (CS) and - at least in some storms - the sting jet (SJ). The timing within the cyclone's lifecycle, the location relative to the cyclone core and some further

5   characteristics differ between these features and hence likely also the associated forecast errors. Here we present a novel objective identification approach for these high-wind features using a probabilistic random forest (RF) based on each feature's most important characteristics in near-surface wind, rainfall, pressure and temperature evolution. As CJ and SJ are difficult to distinguish in near-surface observations alone, these two features are considered together here. A strength of the identification method is that it works flexibly and independent of ~~spatial dependencies and~~ local characteristics and horizontal gradients, such

10   that it can be applied to irregularly spaced surface observations and to gridded analyses and forecasts of different resolution in a consistent way. As a reference for the RF, we subjectively identify the four storm features (WJ, CS, CFC, CJ+SJ) in 12 winter storm cases between 2015 and 2020 in both hourly surface observations and high-resolution reanalyses of the German COSMO model over Europe, using an interactive data analysis and visualisation tool. The RF is then trained on station observations only. The RF learns physically consistent relations and reveals mean sea-level pressure (tendency), potential temperature,

15   precipitation amount and wind direction to be most important for the distinction between the features. From the RF we get probabilities of each feature occurring at the single stations, which can be interpolated into areal information using Kriging. The results show a reliable identification for all features, especially for WJ and CFC. We find difficulties in the distinction of CJ and CS in extreme cases, as the features have rather similar meteorological characteristics. Mostly consistent results in observations and reanalysis data suggest that the novel approach can be applied to other data sets without a need of adaptation.

20   Our new ~~method~~ software RAMEFI (RAndom-forest based MEsoscale wind-Feature Identification) is made publicly available for straightforward use by the atmospheric community, and enables a wide range of applications, e. g., towards a climatology of these features for multi-decadal time periods (see Part II), analysing forecast errors in high-resolution COSMO ensemble forecasts and to develop feature dependent ~~postprocessing~~ post-processing procedures.

## 1 Introduction

25 In the mid-latitudes, extratropical cyclones ~~belong to~~ can produce some of the most severe natural hazards, especially during wintertime. These winterstorms can cause high wind speeds, heavy precipitation, storm surges and thus considerable damage. Prominent examples for Central Europe are *Lothar* (December 1999; Wernli et al., 2002) and *Kyrill* (January 2007; Fink et al., 2009). The development of extratropical cyclones is usually portrayed on the basis of the Norwegian (NC; Bjerknes, 1919) or the Shapiro-Keyser cyclone model (SKC; Shapiro and Keyser, 1990). Both types of cyclones evolve along a frontal wave and

30 include the formation of a warm front followed by a cold front with a warm sector in between. While in NCs the cold front slowly catches up with the warm front resulting in an occluded front, SKCs develop a frontal fracture (as illustrated in Fig. 1) and ultimately a warm seclusion when the warm front, then often referred to as a bent-back front, wraps around the cyclone centre.

High winds are typically associated with four mesoscale features within the synoptic-scale cyclone (Fig. 1): the warm

35 conveyor belt jet or ~~short~~, in short, warm jet (WJ), the cold conveyor belt jet or, in short, ~~short~~ cold jet (CJ), cold-frontal convection (CFC) and the sting jet (SJ), which only occurs within ~~SKC~~SKCs. Furthermore, high wind speeds are often detected within the cold sector (CS) without the formation of a distinct mesoscale feature. As the cold front itself, the cold sector is usually convectively active ~~(labelled pCFC in Fig. 1)~~, leading to downward momentum transport in the vicinity of showers or even thunderstorms. Distinctive characteristics will be discussed in Sect. 2.
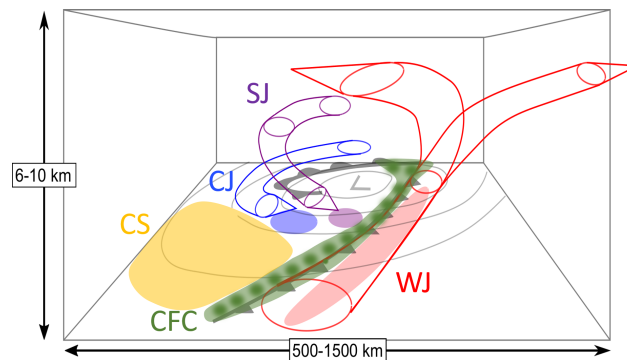
**Figure 1.** Conceptual model of the 3D structure of a SKC showing the WJ (red), CJ (blue) and SJ (magenta). In each case, the region of strong surface winds is indicated by an ellipse. Figure and caption adapted from Clark and Gray (2018) (their Fig. 7) to include CFC (green) and CS (gold) ~~with embedded post-cold-frontal convection (pCFC; light green)~~.

40 All features can cause damage due to strong gusts, such that it is important to accurately forecast them and their associated windfields. A widely employed approach to improve forecasts (not only of wind) is statistical post-processing (Vannitsem et al., 2021), where the model output is corrected on the basis of past forecast errors. The performance of various ensemble post-processing methods for wind gusts is discussed in a recent paper by Schulz and Lerch (2022), who find that approaches ranging from classical statistical methods to novel neural network-based ~~approaches~~ techniques significantly improve fore-

45 cast reliability and accuracy. However, Pantillon et al. (2018) ~~analysed the post-processing~~, who applied one of the classical

statistical methods to ensemble forecasts of wind gusts of and analysed the performance for several winter storms over Germany and, found that in some cases post-processing can actually considerably worsen the forecast, as it did, e. g., for storm *Christian* (October 2013), which is known to have developed an SJ (Browning et al., 2015). Similar results were also found for storm *Friederike* in January 2018 (not shown). This can generally come from a misprediction of the cyclone track or intensity but could also indicate that the characteristics of individual mesoscale high-wind features are not well represented. If that was the case, a feature-dependent post-processing could lead to further improvement, as it could take into account the specific dynamical characteristics and how they are treated in the forecast model. By developing an objective identification algorithm for the wind features shown in Fig. 1, this work lies the foundation for further exploring this idea.

Hewson and Neu (2015) analysed observations and reanalysis data of 29 wind storms with a focus on the three low-level jets, that is, WJ, CJ and SJ. They included CFC in their WJ analysis instead of treating it as an independent feature. They showed that the three wind features differ in their location of occurrence relative to the cyclone centre, in their timing within the cyclone lifecycle, as well as in their duration, strength, surface footprint and more. Furthermore, Earl et al. (2017) looked at the most common causes of high surface gusts in UK extratropical cyclones and, besides WJ, CJ and SJ, included several convection-induced high-wind features. They based the identification of SJs on satellite images, the location of the gusts within the cyclone and the deepening rate. Since no confirmation with Lagrangian trajectory analysis was done, the identified features are referred to as "potential SJs". Earl et al. (2017) found that although WJs and CJs are the most common causes of high winds, the strongest gusts are caused by CFC and potential SJs. Parton et al. (2010) categorised strong wind events captured by a wind-profiling radar in Wales over a 7-year period into cold frontal events (similar to our definition of CFC), warm sector winds (similar to WJ), tropopause folds/warm fronts, SJs and unclassified events. They found that warm sector events were the cause of around $40\%$ of all strong winds followed by cold frontal events with around $24\%$.

As all these approaches are purely subjective and relatively time consuming and thus hard to automate, we aim to develop an objective analysis of the different mesoscale wind features that can flexibly be applied to station and gridded data and thus serve as a basis for climatological studies, forecast evaluation and postprocessing post-processing development. The strategy we follow is to start with a subjective identification (as in previous studies) but to use the results to then train a probabilistic random forest (RF) to develop an objective procedure that can be applied to cases outside of the training data set. The identification is designed to be independent of horizontal gradients, hence resolution, and can principally be applied to observations from a single weather station. In addition, the identification is based on tendencies over $1\,\mathrm{h}$ only, making it applicable to timeseries with gaps. Our new developed method is referred to as RAMEFI (RAndom-forest based MEsoscale wind Feature Identification). Given that the provision of such a feature-dependent post-processing tool can enhance the forecasts of strong winds and wind gusts, it can potentially contribute towards better weather warnings and impact forecasting of such events (e. g., Merz et al., 2020). In this paper we will show examples using surface stations and COSMO reanalysis data. A full long-term climatology is the focus of Part II of this paper. The output of the RF are feature probabilities rather than binary identification, which allows an evaluation of how good individual data points fit the typical feature characteristics and the identification of hybrid features or transition zones.

The paper is structured as follows: First, a short description of each wind feature is given in Sect. 2 based on results from previous studies, before the used data sets and methods are discussed in Sect. 3. Section ~~?? then details the used methods~~4 then details our new software RAMEFI, starting from the subjective labelling of wind features through the training of the RF to the display of areal feature information. While Sect. 5 illustrates our approach with a case study, a statistical evaluation of the performance of the RF can be found in Sect. 6. Strengths and weaknesses of the approach are discussed on the basis of specific case studies in Sect. 7. In Sect. 8 we finally summarise our work and discuss future plans.

## 2 Characteristics of high-wind features

In this section the most important high-wind features within extratropical cyclones ~~(Fig. 1)~~ are shortly described on the basis of the existing scientific literature. Figure 1 exemplarily shows a typical SKC; differences to an NC are discussed in the following. While WJs and CJs are common to both cyclone models, SJs can only occur in SKC (Clark and Gray, 2018). On the other hand, SKC are rarely accompanied by (strong) CFC, since their cold front is often weak (Schultz et al., 1998). The ~~description is~~ following descriptions of individual features are mostly based on Hewson and Neu (2015), Earl et al. (2017) and Clark and Gray (2018). For more details, the reader is referred to these publications.

### 2.1 Warm ~~(conveyor belt)~~ jet

As the name implies, the WJ is associated with a warm air flow, typically ahead of and later ascending above the surface cold front, often referred to as the warm conveyor belt ~~(Fig. 1; Wernli and Davies, 1997; Eckhardt et al., 2004; Madonna et al., 2014).~~ (WCB; Wernli and Davies, 1997; Eckhardt et al., 2004; Madonna et al., 2014). During the ascent, the WCB splits into a cyclonic and anticyclonic branch as seen by the red tubes in Fig. 1. While the cyclonic part forms the cloud head and usually causes heavy precipitation along a narrow region, the anticyclonic part rises above the warm front and brings more moderate precipitation over a wider area. Overall, the WCB is the main cause for long-lasting precipitation (Catto, 2016). Furthermore, the WCB can be the cause of strong convection along the cold front (Hewson and Neu, 2015).

Contrary to Hewson and Neu (2015), we define the WJ as the region ahead of the cold front and its convection, hence ahead of the CFC feature (cf. Sect. 2.4), as displayed by the red shaded ellipse in Fig. 1. Located in the warm sector of the cyclone between the two rain-active fronts, the WJ is usually characterised by positive temperature anomalies, decreasing pressure with time and little or no precipitation. The WJ is usually associate with the first strong winds starting in an early stage of a cyclone. Maximum gusts of around $25\,\mathrm{m\,s^{-1}}$ are typical (Hewson and Neu, 2015). Since the warm conveyor belt is an ascending airstream, the winds at the surface weaken and disappear in later stages when the warm conveyor belt no longer affects the boundary layer. The jet is long-lived with a duration of $24\,\mathrm{h}$ to $48\,\mathrm{h}$ and can cause a large surface wind footprint with a width of $200\,\mathrm{km}$ to $500\,\mathrm{km}$ and length of up to $1000\,\mathrm{km}$ (Hewson and Neu, 2015). While the predictability was evaluated to be good with a relatively high coherence in space and time, Hewson and Neu (2015) found that the occurrence of very high winds within the warm sector are rather unusual, while Parton et al. (2010) associates $40\,\%$ of strong wind events with winds within the warm sector. Such a seeming contradiction can potentially be resolved with a climatological application of the identification

algorithm we develop in this paper. Due to generally stable conditions in the warm sector, wind speeds above the boundary layer are usually much higher than surface gust speeds. Compared to the CJ and SJ, the WJ is the most long-lasting, but – as already mentioned – typically does not cause the most destructive winds.

## 2.2 Cold ~~(conveyor belt)~~ jet

The CJ is associated with the main airflow of the cold conveyor belt that turns cyclonically around the centre of the low ~~.~~ (blue tube in Fig. 1). At first, the CJ moves around the northwestern flank behind the occluded or bent-back front beneath the cloud head. Since it is travelling against the motion of the low-pressure system, the CJ is hard to see in Earth-relative winds until it wraps around the cyclone centre. Once wrapped around, the CJ can cause strong surface gusts near the tip of the front as shown in Fig. 1 by the blue shaded ellipse. This usually happens around the time the maximum intensity is reached. The CJ weakens when the low decays or shortly before that. The jet mainly stays close to the ground, i. e., below $850\,\mathrm{hPa}$ (Smart and Browning, 2014) or is ascending slightly during its lifecycle (Martínez-Alvarado et al., 2014). With typical maximum gusts around $30\,\mathrm{m\,s^{-1}}$ the CJ is stronger than the WJ but with a typical lifetime of $12\,\mathrm{h}$ to $36\,\mathrm{h}$ does not last as long (Hewson and Neu, 2015). The impacted area expands with time, while the cold conveyor belt, and hence the CJ, wraps around the cyclone centre when finally the footprint can reach a width of around $100\,\mathrm{km}$ to $800\,\mathrm{km}$ and a length of up to $2500\,\mathrm{km}$ (Hewson and Neu, 2015). Although forming later in the lifecycle of the parent cyclone than the SJ, both jets can coexist. A damaging CJ is more common than an SJ or WJ over Europe (Hewson and Neu, 2015).

## 2.3 Sting jet

The SJ is a distinct airstream that descends from mid-levels within the cloud head into the frontal fracture region of an SKC (magenta tube in Fig. 1). When the CJ wraps around the low, the SJ can be replaced by the CJ or merge with it. Hewson and Neu (2015) suggest an average surface footprint of less than $100\,\mathrm{km}$ in width and up to $800\,\mathrm{km}$ in length. SJs usually last just a few hours but can be active up to $12\,\mathrm{h}$ in extreme cases. A detailed review can be found in Clark and Gray (2018). The most common way to identify SJs is to compute Lagrangian trajectories (e. g., Volonté et al., 2018; Eisenstein et al., 2020), for which a high resolution is required, horizontally and vertically, but also in time. Recent work has tried to identify SJs with low-cost approaches. For example, Gray et al. (2021) introduced an instability-based precursor tool, while Manning et al. (2022) developed a kinematic approach looking for reversals in the vertical gradient of horizontal wind speed along streamlines. ~~The distinction between the SJ and CJ in surface observations alone is challenging as we will discuss in Sect. 4.1.~~

## 2.4 Cold-frontal Convection

The passage of cold fronts is often accompanied by heavy precipitation, sometimes in form of convective lines, which, in turn, can cause strong wind gusts associated with the downward transport of high momentum from above the boundary layer. CFC is displayed in Fig. 1 as an elongated line along the front with several darker spots representing individual convective events.

Cold frontal rain, snow and graupel is responsible for around $28\,\%$ of extreme precipitation events in the midlatitudes (Catto and Pfahl, 2013). The frontal zone is characterised by a marked change in wind direction and a decrease in temperature (Clark,

145 2013). In extreme cases, tornadoes can occur in association with CFC, causing even more hazardous winds (e. g., *Kyrill*; Fink et al., 2009). Earl et al. (2017) suggest a separation of CFC into convective lines and pseudo-convective lines. The latter do not strictly satisfy identification criteria for convective lines but show characteristics of organised, strong convection and may fulfil the criteria earlier or later. A convective line shows a clear signal in radar imagery at $3\,\mathrm{km}$ to $10\,\mathrm{km}$ height (Parton et al., 2010). In case of a kata cold front, CFC occurs at and ahead of the surface cold front (Lackmann, 2011). Although high winds

150 might then strictly speaking occur within the warm sector, we have decided not to associate it with the WJ due to the physically distinct characteristics. In case of an ana cold front, CFC occurs at and behind the cold front, and the distinction is therefore more straightforward. SKC usually have a rather weak cold front, hence high winds are less likely to be associated with CFC (Schultz et al., 1998).

## 2.5 Cold sector winds and post-cold-frontal convection

155 The cold sector is the region behind the cold front .(gold shading in Fig. 1). In this area high winds can be caused by post cold-frontal convection and also by a so-called dry intrusion. This region of a cyclone is generally being known for its instability and turbulent behaviour. A dry intrusion is, as an SJ, a descending airstream but in this case one that originates near the tropopause or even in the lower stratosphere. This way it brings dry air down to the middle and lower troposphere moving towards the cyclone centre (Raveh-Rubin, 2017). Dry intrusions – and later in the cyclone's lifecycle a dry slot near the cyclone centre –

160 can often be seen in water vapour satellite imagery. Raveh-Rubin (2017) state that dry intrusions can cause destabilisation and increased wind gusts. Furthermore, cold fronts show a stronger temperature gradient, higher winds and more precipitation when accompanied by a dry intrusion (Catto and Raveh-Rubin, 2019; Raveh-Rubin and Catto, 2019). Most publications consider all high-winds on the colder side of a cyclone to be a CJ or SJ and do not distinguish these features to other cold-sector winds (e. g., Manning et al., 2022). Given the relatively large area of the cold sector in many cyclones and the appearance of discernible

165 substructures, here we decided to specifically separate out CFC, CJ and SJ, and then label the remaining strong winds as CS.

## 3  Data and method

This section introduces the observation and gridded data sets as well as 12 winter storm case studies used for the training and evaluation of RAMEFI. Furthermore, it is described how to assess probability predictions obtained by the RF.

## 3.1  Surface observations

170 The main basis of our analysis is a data set of hourly surface observations from 2001 to mid-2020. This includes mean sea level pressure ($p$), 2 m air temperature ($T$), wind speed in at 10 m ($v$), wind direction in at 10 m ($d$) and precipitation amount ($RR$). Using $T$ and $p$, we further compute the surface pressure using the barometric height formula to then calculate the potential temperature ($\theta$). Our focus is on Europe, more specifically stations within the area of $-10\,^\circ\mathrm{E}$ to $20\,^\circ\mathrm{E}$, $40\,^\circ\mathrm{N}$ to $60\,^\circ\mathrm{N}$. Around

1700 stations are included of which, however, less than 400 stations on average observe all five parameters. For the training of
175 the RF (Sect. ??4), we focus on stations that measure at least three of the five parameters. The most frequent missing parameter in the hourly data is $RR$, as many stations only measure 3- or 6-hourly precipitation. However, many stations, especially over Germany, measure $RR$ only and, hence, are not usable for the training of the RFs (Sect, 4.2) but still helpful to inform our subjective labelling (Sect. 4.1). In addition, we exclude mountain stations, that is, those with a station height above $800\,\mathrm{m}$, as we suspect these to be dominated by orographic influences that may blur the feature characteristics we want to identify. This
180 leaves around 750 stations per time step.

In order to take into account the diurnal and seasonal cycles as well as location-specific characteristics (e. g., exposed stations in coastal regions) in $\theta$ and also $v$, we decided to normalise these parameters by their climatology. For $\theta$, this means $\tilde{\theta} = \theta/\theta_{50}$, where $\theta_{50}$ is the median for the specific location, time of day and day of the year $\pm 10$ days. This is done analogously for $v$ using the 98$^{\text{th}}$ percentile: $\tilde{v} = v/v_{98}$, as we are mostly interested in high winds in this work. The 98$^{\text{th}}$ percentile is used in analogy
185 to standard high-wind quantities such as the Storm Severity Index (SSI), which is computed from stations where measured gusts exceed the local 98th percentile and provides an integral indication for the strength of the cyclone and the associated potential damage (see Appendix C for details). Both $\theta_{50}$ and $v_{98}$ are computed for the time period 2001 to 2019. Moreover, we are interested in temporal tendencies of $p$, $\tilde{\theta}$ and $d$, here represented simply by the difference between the current and the prior time step ($\Delta p$, $\Delta\tilde{\theta}$ and $\Delta d$, respectively). All parameters and their descriptions are listed in Table 1.

**Table 1.** Overview of the variables considered for the objective identification using the probabilistic RF. The ~~third~~ fourth column indicates whether the variable is used as ~~a predictor~~ an input variable for the final version of the RF. The associated percentiles and medians are computed with respect to the location, time of day and day of the year $\pm 10$ days.

| Variable | Description | Unit | RF | Derivation |
|---|---|---|---|---|
| $v$ | 10m wind speed | $\mathrm{m\,s^{-1}}$ | - | Station observation. |
| $\tilde{v}$ | Normalised wind speed | 1 | ✓ | $\tilde{v} = v/v_{98}$, where $v_{98}$ is the assoc. 98$^{\text{th}}$ percentile. ~~Consider only $\tilde{v} \geq 0.8$~~ |
| $d$ | Wind direction | ° | ✓ | Station observation |
| $\Delta d$ | Tendency of wind direction | $\mathrm{°\,h^{-1}}$ | ✓ | $\Delta d = d - d_{-1}$, where $d_{-1}$ is the observation of the previous hour |
| $p$ | Mean sea level pressure | hPa | ✓ | Station observation |
| $\Delta p$ | Tendency of mean sea level pressure | $\mathrm{hPa\,h^{-1}}$ | ✓ | $\Delta p = p - p_{-1}$, where $p_{-1}$ is the observation of the previous hour |
| $T$ | 2m air temperature | K | - | Station observation |
| $\theta$ | Potential temperature | K | - | Derived from $T$ and $p$ |
| $\tilde{\theta}$ | Normalised potential temperature | 1 | ✓ | $\tilde{\theta} = \theta/\theta_{50}$, where $\theta_{50}$ is the assoc. median |
| $\Delta\tilde{\theta}$ | Tendency of norm. pot. temperature | $\mathrm{h^{-1}}$ | ✓ | $\Delta\tilde{\theta} = \tilde{\theta} - \tilde{\theta}_{-1}$, where $\tilde{\theta}_{-1}$ is the derived value of the previous hour |
| $RR$ | Precipitation | $\mathrm{mm\,h^{-1}}$ | ✓ | Station observation |

## 3.2 COSMO-REA6

As an example for a gridded data set, we use COSMO-REA6 data from the Hans-Ertel-Centre for Weather Research, which is a reanalysis based on the COSMO model from the German Weather Service (DWD) covering the European CORDEX domain with a grid spacing of $0.055\,\mathrm{deg}$, i. e., roughly $6\,\mathrm{km}$ (Bollmeyer et al., 2015). Several parameters, including wind, temperature, humidity and pressure, are assimilated using a nudging technique. Observations from radiosondes, aircraft, wind profilers and surface stations are used for the nudging. However, precipitation is not assimilated, which can lead to larger deviations between the reanalyis and observations by raingauges or radar (Bach et al., 2016; Hu and Franzke, 2020). The reanalysis is available from 1995 to 2019. This means that one of our case studies, namely storm *Sabine*, is not included (Table 2). The same surface parameters as mentioned in Sect. 3.1 are used. The data set contains $p$, $T$, $RR$ and the zonal and meridional wind components, from which we can compute $v$ and $d$. Again, we further calculate $\tilde{\theta}$ and the temporal tendencies $\Delta p$, $\Delta\tilde{\theta}$ and $\Delta d$. Due to computational cost, we compute $\theta_{50}$ and $v_{98}$ for the 10-year time period 2005 to 2015 only but this should have a negligible effect on the final outcome. The data, originally on a rotated grid, are regridded to a latitude-longitude grid with a grid spacing of $0.0625°$, i. e., roughly $7\,\mathrm{km}$.

## 4 Method

The objective identification of mesoscale high-wind features within European wintertime cyclones includes three steps described in the following subsections. First, we identify the features subjectively in surface observations in several selected case studies, such that each station is assigned to a specific feature. These labels are then used to train RFs for feature prediction on the basis of a cross-validation approach. In a final step, we obtain forecasts on a grid by interpolating the predicted probabilities using a Kriging approach. For the COSMO-REA6 data, the features are identified analogously. Instead of training separate RFs, we apply the RFs trained on the surface observations. As the COSMO-REA6 forecasts are already grid-based, the Kriging step is obsolete. We will refer to our newly developed method as RAMEFI (RAndom-forest based MEsoscale wind Feature Identification).

### 3.1 Subjective labelling using an interactive tool

Given the sometimes unclear distinction between the high wind features of interest in realistic cases, we decided to base our algorithmic development on how experienced meteorologists would identify the features on the basis of a wide range of parameters and their evolution in time and space. To accelerate and facilitate the subjective labelling of high-wind features, we developed an interactive tool using the open-source data visualisation package `bokeh` for Python (Bokeh Development Team, 2021) , where one can switch between various parameters in a graphical display and select an area to set labels using a mouse-controlled lasso tool. A screenshot is provided in the Appendix (Fig. D1). The chosen parameters, taken from station observations or gridded data, include $\tilde{v}$, $p$, $\Delta p$, $\tilde{\theta}$, $\Delta\tilde{\theta}$, $RR$, $d$ and $\Delta d$. All used parameters are independent of the location of the station/grid

220 ~~point and horizontal gradients such that, theoretically, the approach can later be applied to a single station. Furthermore, the approach is independent of temporal evolutions beyond 1 h.~~ for the area of $-10\,°\mathrm{E}$ to $30\,°\mathrm{E}$, $40\,°\mathrm{N}$ to $65\,°\mathrm{N}$.

~~To distinguish the features, the locations of the fronts and the cyclone centre are important, such that analysis charts of the DWD and the UK Met Office were used for orientation. For a more detailed analysis, we further used the 3D front detection of the interactive visualisation software Met.3D (Rautenhaus et al., 2015). We only label stations with a high wind speed $\tilde{v} \geq 0.8$,~~
225 ~~that is, reaching 80% of their $98^{\text{th}}$ percentile.~~

~~The guiding principles for the labelling were extracted from the scientific literature (cf. Sect. 2). As the WJ occurs within the warm sector, we expect warmer temperatures in this region, mostly westerly winds and almost no precipitation. Furthermore, $p$ decreases ahead of the cold front (Lackmann, 2011). With the arrival of the cold front, $p$ starts increasing again, temperatures cool and we find a veering shift in wind direction. As the name implies, CFC is mainly marked by a line of (heavy) precipitation.~~
230 ~~The CJ occurs on the colder side of the cyclone near the cyclone centre during maximum intensity, such that $p$ is generally lowest for this feature. It is often described to have a hook-shaped footprint in surface wind along the tip of an occluded or bent-back front. An SJ is labelled, where model-based trajectories analogous to Eisenstein et al. (2020) confirm a descending airstream. The area behind the cold front that is not associated with the CJ or SJ is labelled as the CS. An example of the reasoning behind the labelling is described for one time step of storm *Burglind* in Sect. 5.~~
235 ~~The subjective labelling was done by the first author for~~

## 3.1 Case studies

In this work we focus on 12 ~~extratropical cyclone case studies~~ winter storm case studies between the years 2015 and 2020 listed in Table 2. The selection was based on their ~~Storm Severity Index (SSI )~~ SSI over Germany (see Appendix ~~C~~C), caused damage and impacted area~~, while overall attempting~~. This includes the most devastating winter storms occurring in this time
240 period but also some more moderate ones to capture a healthy diversity of cyclones and features. The selected cases occurred during the extended winter half year between the end of September and end of March. They vary in terms of their cyclone tracks (Fig. 2) and occurring high-wind features, and include both ~~NC and SKC~~NCs and SKCs. Two case studies developed ~~an SJ~~SJs, namely Egon (Eisenstein et al., 2020) and Friederike~~, which are included in the CJ feature here as explained in Sect. ??.~~. We also include two storms, named *Herwart* and *Sabine*, ~~where high wind speeds are enhanced due to a strong pressure~~
245 ~~gradient~~ with an exceptional large pressure gradient leading to a stronger background wind field, such that it is more difficult to distinguish the features and the contribution of them to the storm's wind footprint. Further, *Sabine* stands out to be an extremely deep cyclone with a minimum core pressure of $944\,\mathrm{hPa}$ during its lifetime. ~~More information on can be found in Sect. 7. In total 282 time steps have been analysed.~~

~~As~~

## 3.2 Assessing probability predictions for multiple wind features

Probability predictions of three or more classes, such as the wind features, are typically evaluated by downscaling to two-class problems, of which the one-against-all and all-pairs approaches are two well-known examples (Zadrozny and Elkan, 2002)

**Table 2.** Selected winter storm cases from 2015 to 2020 over Central Europe (names as given by FU Berlin), date, maximum observed gust speed (location), SSI over Germany and associated high-wind features including exceptional large pressure gradients (SPG). The cyclone tracks are displayed in Fig. 2

| Case | Date | max. gust [km h$^{-1}$] above 800 m | max. gust [km h$^{-1}$] below 800 m | SSI | Feat... |
|------|------|-------------------------------------|-------------------------------------|-----|---------|
| Niklas | 31 Mar 2015 | 192 (Zugspitze, D, 2964 m) | 148 (Weinbiet, D, 553 m) | 20.8 | WJ, |
| Susanna | 09 Feb 2016 | 158 (Patscherkofel, AT, 2247 m; Pilatus, CH, 2106 m) | 158 (Île de Groix, FR, 46 m) | 3.6 | WJ, |
| Egon | 12-13 Jan 2017 | 150 (Fichtelberg, D, 1231 m) | 148 (Weinbiet, D, 553 m) | 5.9 | WJ, |
| Thomas | 23-24 Feb 2017 | 158 (Brocken, D, 1134 m) | 152 (Capel Curig, UK, 216 m) | 3.0 | WJ, |
| Xavier | 05 Oct 2017 | 202 (Sněžka, CZ, 1602 m) | 141 (Saint-Hubert, BE , 563 m) | 6.3 | WJ, |
| Herwart | 29 Oct 2017 | 176 (Fichtelberg, D, 1231 m) | 144 (List/Sylt, D, 26 m) | 15.2 | WJ, |
| Burglind | 03 Jan 2018 | 217 (Feldberg, D, 1490 m) | 150 (Wädenswil, CH, 463 m) | 15.2 | WJ, |
| Friederike | 18 Jan 2018 | 204 (Brocken, D, 1134 m) | 144 (Hoek Van Holland, NL, 7 m) | 18.3 | WJ, |
| Fabienne | 23 Sep 2018 | 141 (Feldberg, D, 1490 m) | 158 (Weinbiet, D, 553 m) | 4.6 | WJ, |
| Bennet | 04 Mar 2019 | 181 (Cairngorn, UK, 1245 m) | 151 (Cape Corse, FR, 106 m) | 5.1 | WJ, |
| Eberhard | 10 Mar 2019 | 194 (Sněžka, CZ, 1602 m) | 141 (Weinbiet, D, 553 m) | 10.1 | WJ, |
| Sabine | 09-10 Feb 2020 | 195 (Sněžka, CZ, 1602 m) | 219 (Cape Corse, FR, 106 m) | 20.0 | WJ, |
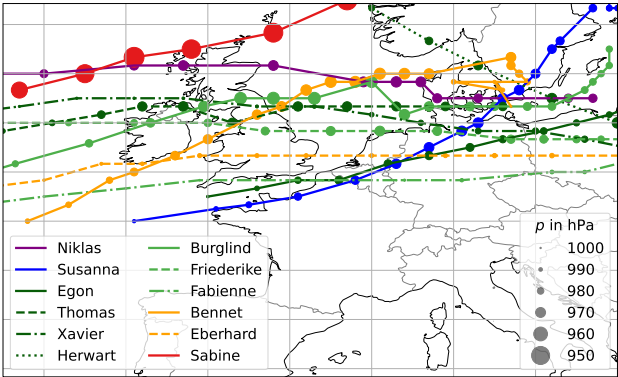
**Figure 2.** Cyclone tracks of case studies from Table 2. Colour indicates the year (purple = 2015, blue = 2016, dark green = 2017, green = 2018, orange = 2019, red = 2020), linestyle indicates the order in that year (solid - dashed - dashdot - dotted). The size of the markers correspond to the minimum mean sea level pressure $p$.

. While the one-against-all approach compares the occurrence of one wind feature against all others grouped together, the all-pairs approach considers the conditional probabilities for each pair of classes, for example, the conditional probabilities of

255    the WJ and the CJ when one of the two features materialises. The one-against-all approach is used to evaluate how well one specific wind feature is forecast, the all-pairs approach to evaluate the ability to discriminate between two wind features.

    The probabilities are evaluated based on the paradigm of Gneiting et al. (2007) that a prediction should aim to maximise sharpness subject to calibration. Calibration refers to the consistency of the prediction and the observation, while sharpness is a property of the prediction alone and refers to the associated uncertainty. In a nutshell, a probability $f$ is called calibrated if
260    the conditional event probability (CEP; conditional on $f$) matches $f$, for example, if a $20\,\%$ forecast is issued 100 times, the event should occur about 20 times. Further, a probability prediction is said to be sharper, the more confident the prediction is, that is, the closer to 0 or 1. Both calibration and sharpness can be assessed qualitatively via reliability diagrams, which display the calibration curve (Sanders, 1963; Wilks, 2011). The calibration curve is a plot of the CEP dependent on the probability $f$, which is close to the diagonal if the predictions are calibrated. In addition to the calibration curve, the frequency of the
265    probabilities is illustrated by a histogram. The more U-shaped the histogram is, the closer the predictions are to 0 and 1 and thus the sharper. The reliability diagrams shown in this paper are based on the novel CORP-approach (Consistency, Optimality, Reproducibility, PAV-algorithm based) of Dimitriadis et al. (2021), which yields optimal calibration curves and eliminates the need of implementation decisions for the calculation of the calibration curve.

    Quantitatively, calibration and sharpness can be assessed using the Brier score (BS; Brier, 1950). The lower the score, the
270    better the prediction. Here, we compare our RF probabilities with the class frequencies observed in the training data of the RF using the Brier skill score (BSS), which denotes the improvement of the RF over a prediction based on the class frequencies in the RF training data, where a negative percentage corresponds to worse predictions, 0% to no improvement and 100% to an optimal prediction. Details on the evaluation of probability predictions are provided in Appendix A1 (practical implementation) and B1 (mathematical formulation).

## 4  RAMEFI

Our new method RAMEFI focuses on strong but not exceptionally high wind speeds. The latter are usually indicated by the 98th percentile. To obtain a sufficiently large storm area and to base that on a widely used reference, we decided to include stations reaching $80\,\%$ of their 98th percentile, i. e., $\tilde{v} \geq 0.8$. To capture usually narrow and fast-moving features such as CFC, RAMEFI requires hourly data. All used parameters are independent of the location of the station/grid point and horizontal gradients, such
280    that, in principle, the approach can be applied to a single station and data sets with differing horizontal resolution. Furthermore, the approach is independent of temporal evolutions beyond $1\,\mathrm{h}$.

    RAMEFI includes three steps described in the following subsections. First, we identify the features subjectively in surface observations in several selected case studies, such that each station is assigned to a specific feature. These labels are then used to train RFs for feature prediction on the basis of a cross-validation approach. In a final step, we obtain forecasts on a grid
285    by interpolating the predicted probabilities using a Kriging approach. For the COSMO-REA6 data, the features are identified analogously. Instead of training separate RFs, we apply the RFs trained on the surface observations. As the COSMO-REA6 forecasts are already grid-based, the Kriging step is obsolete.

## 4.1 Subjective labelling using an interactive tool

Given the sometimes unclear distinction between the high wind features of interest in realistic cases, we decided to base our algorithmic development on how experienced meteorologists would identify the features on the basis of a wide range of parameters and their evolution in time and space. To accelerate and facilitate the subjective labelling of high-wind features, we developed an interactive tool using the open-source data visualisation package `bokeh` for Python (Bokeh Development Team, 2021), where one can switch between the available parameters (Table 1) in a graphical display and select an area to set labels using a mouse-controlled lasso tool. A screenshot is provided in the Appendix (Fig. D1).

The guiding principles for the labelling were extracted from the scientific literature (cf. Sect. 2) and are mainly based on the location relative to the cold front and cyclone core. For this, analysis charts of the DWD and the UK Met Office were used for orientation. For a more detailed analysis, we further used the 3D front detection of the interactive visualisation software Met.3D (Beckert et al., 2022; Rautenhaus et al., 2015). In our surface parameters, a cold front is then mostly identified through the characteristic change of the sign of $\Delta p$. It is labelled CFC, if a larger area of precipitation along it is observed, while high winds ahead of the front within the warm sector are labelled WJ. The CJ is mostly detected through its hook-shaped wind footprint at the tip of a wrapped-around occlusion or bent-back front as well as through its proximity to the cyclone centre. An SJ is labelled, where model-based trajectories analogous to Eisenstein et al. (2020) confirm a descending airstream. The area behind the cold front that is not associated with the CJ or SJ is labelled as the CS. An example of the reasoning behind the labelling is described for one time step of storm *Burglind* in Sect. 5.

The subjective labelling was done for the introduced 12 extratropical cyclone case studies (cf. Sect. 3.1). In total 282 time steps have been analysed. As mentioned above, we excluded mountain stations and stations where less than three of the given parameters were measured. This leaves around 750 stations per time step for the subjective labelling. Overall for the 12 case studies, we have 77.517 data points where $\tilde{v} \geq 0.8$, of which 19.200 (24.77 %) are not associated with a feature (NF), ~~19.501 (25.16 %~~21.809 (28.13 %) were labelled as ~~a WJ, 3.800 (4.9 %) as CFC~~CS, 19.501 (25.16 %) as WJ, 11.705 (15.1 %) as CJ, 3.800 (4.9 %) as CFC and 1.502 (1.94 %) as ~~SJ and 21.809 (28.13 %) as CS. Since~~ SJ. However, the SJ is a small, short-lived and rare feature and the characteristics of SJs and CJs in surface parameters are very similar due to the proximity in both time and space~~, it is included~~. A first training with SJ and CJ as separate features showed that a clear distinction is not possible with the information at hand and that the SJ is mostly detected as CJ. Therefore, we decided to include it in the more frequent CJ feature, increasing the values for CJ to 13.207 data points (17.04 %).

The features were further labelled in all case studies using the interactive tool for COSMO-REA6 data – except for *Sabine* that occurred outside of the reanalyis time period. These labels are used to evaluate the predictions generated by the station-based RFs for a grid-based data set (Sect. 4.2 and 6.1). For computational reasons, i. e., as labels are set for every grid point rather than an area, we downsampled the COSMO grid to every third grid point in the zonal and meridional directions, resulting in a grid spacing of 0.1875 deg, i. e., around 21 km. Moreover, we excluded ocean grid points, as the characteristics of the high-wind features might be different from land due to different surface friction, surface heat fluxes, etc. Regions with high wind speed not directly associated with a winter storm, especially over Italy and the Balkans, were not labelled.

## 4.2 Probabilistic Random Forest

An RF (Breiman, 2001) is a popular, ~~non-parametric~~ robust machine learning method for classification and regression problems that does not rely on ~~distributional~~ parametric assumptions but instead is based on the idea of decision trees (Breiman, 1984). ~~A decision tree subsequently splits the predictor space based on order criteria at~~ Given data for which we want to generate predictions, a decision tree operates by assigning each sample to one of its so-called ~~nodes and thereby generates a partitioning. The order criteria are chosen automatically such that the resulting subsets of observations associated with a partition are as diverse as possible, where diversity is defined through a chosen splitting rule. The subspaces induced by the partitioning, the so-called~~ leaves, ~~include a set of observations in the training set from which an analogue forecast is derived. In case of multi-class probability forecasts , the class frequencies within the terminal leave are taken as forecast.~~ which is done by subsequent queries at the so-called nodes, where we check whether one of the input variables exceeds a certain threshold. Each leaf is associated with that subset of the training data that satisfies the criteria of the corresponding nodes. Tailored to the specific task, a prediction is then derived from the associated subset. Here, we obtain probability forecasts by using the frequencies of the observed wind features among the samples in the corresponding leaf. The queries at the nodes, that is, the specific input variables and corresponding thresholds, are chosen automatically such that the terminal leaves are as diverse as possible. The maximal depth, the minimal node size and the diversity criterion are tuning parameters and can be chosen by the user. An RF builds a randomised ensemble of decision trees, where the generation of each tree is based on a different subsample of the training data and that of each node on a different subset of ~~predictor~~ input variables. To obtain a final ~~probability forecast~~ prediction from the ensemble ~~of decision trees~~, the individual predictions of the decision trees are aggregated. For further details on RFs, we refer to Breiman (2001) and Hastie et al. (2009). In a meteorological context, probabilistic RFs have already been applied to predict damaging convective winds (Lagerquist et al., 2017) and severe weather (Hill et al., 2020), but also in a general form for a wide range of applications such as ensemble post-processing of surface temperature and wind speed (Taillardat et al., 2016).

Machine learning methods such as the RF are often referred to as black boxes due to a lack of interpretability, although there exist several techniques to understand what the models have learnt and how the predictions are related to the input variables, typically referred to as predictors (McGovern et al., 2019). We will apply two of these predictor importance techniques, one to find the most relevant predictors and one that illustrates the effect of the predictor values on the ~~forecast~~ RF probabilities. The first is the permutation importance of a predictor (Breiman, 2001). Proceeding separately for each predictor, the values ~~within the test set~~ of that predictor are shuffled randomly within the test data in space and time such that the ~~link to the target variable~~ physical relation to the observed wind feature is broken. Then, based on these permuted ~~predictors~~ predictor values, new predictions are generated and compared to ~~the predictions~~ those obtained with the original data. The worse the ~~forecasts~~ predictions become (with respect to an evaluation measure), the more important the predictor. Here, we measure ~~forecast~~ predictive performance with the ~~Brier score (BS; Brier, 1950).The~~ BS that was introduced in Sect. 3.2, the importance measure is referred to as BS permutation importance. The second technique is given the partial dependence plot (PDP; Greenwell, 2017), which illustrates the effect of a predictor on the ~~predictions can be illustrated with a partial dependence plot (PDP; Greenwell, 2017)~~

prediction. Given a fixed predictor, a PDP shows the expected ~~probability forecast as a function of~~ RF probability dependent on the value of the predictor variable while averaging out the effects of the other predictors. Hence, a PDP illustrates how the ~~probability forecasts~~ RF probabilities depend on the value of a specific predictor variable, on average. For more details, we refer to McGovern et al. (2019).

In this study, we apply RFs to generate ~~probabilistic forecasts~~ probabilities of the wind features presented in Sect. 2. The ~~parameters~~ input variables used are listed in Table 1. ~~As mentioned in Sect. 4.1, all parameters are independent spatially as well as temporally beyond 1 h. This local approach gives us more flexibility but can cause detriments. Advantages and disadvantages of this approach are discussed in Sect. 7.4.~~ For the station-based observations, we use a cross-validation scheme on the different winter storm cases, that is, for each winter storm the predictions are ~~based on~~ generated by an RF that is trained on the data of the remaining 11 winter storms. Training RFs in a similar cross-validation scheme for the COSMO-REA6 data becomes computationally infeasible as the underlying data sets become too large. ~~As~~ Since the underlying processes should coincide for both the station- and model-based data, we instead apply the station-based RFs in the same ~~cross-validation~~ scheme to generate ~~probability forecasts~~ probabilities using the COSMO-REA6 data. Details on the implementation including the choice of the ~~hyperparameters~~ tuning parameters can be found in ~~the Appendix~~ Appendix A2.

Due to normalising $\theta$ and $v$, the trained RF is fairly independent of location-specific information, such that it can hopefully be applied successfully to other midlatitude regions around the world affected by extratropical cyclones. However, before doing that we recommend a thorough sanity check, particularly when using it over the ocean and mountainous regions.

## 4.3 Kriging

As it is difficult to envision a coherent area of a certain wind feature from probabilities at single stations that are distributed irregularly over the study area, we interpolate the station-based ~~probability forecasts~~ probabilities to a regularly spaced grid in order to visualise the results. In geostatistics, this is generally achieved by ~~methods based on the best linear unbiased prediction (BLUP), commonly referred to as~~ Kriging (Matheron, 1963). In principle, the Kriging predictions (here on the grid) are the weighted averages of the input data (here the station data), where the specification of the weights is driven by the covariance of the underlying random process. Under the assumption ~~that the input data is a partial realisation of a Gaussian process~~ of Gaussianity (Rasmussen and Williams, 2005), Kriging provides the optimal full predictive distribution. The key requirement for the implementation of Kriging in the context of Gaussian processes is the specification of the mean and the covariance function. ~~Mathematical details regarding the Kriging approach taken here are provided in the Appendix.~~

In this study, we perform univariate Kriging to obtain probability maps for each wind feature, where we specify the mean and covariance function by a constant mean function and the stationary Matérn covariance function (Matérn, 1986; Guttorp and Gneiting, 2006). For the estimation, we resort to the method of maximum likelihood estimation for Gaussian processes. However, since the input data are, in our case, ~~multi-class~~ probabilities and thus deviate from the Gaussianity assumption, we perform a data transformation for approximate Gaussianity. For the production of probability maps, we independently perform Kriging on each of the class probabilities, ~~and execute probability sum normalisation on the predicted~~ hence univariate Kriging, and normalise the resulting probabilities for each grid point such that, across ~~multiple classes, the predicted~~ the multiple wind

390   feature, the probabilities sum to 1. Note that the Kriging predictions are only obtained for areas over land, where our winter storms occurred and where a sufficient amount of data was available for a reliable interpolation. More details ~~on the practical implementation of Kriging are also provided in the Appendix~~ regarding the Kriging approach are provided in Appendix A3 (practical implementation) and B2 (mathematical formulation).

## 5   Illustrative case study: Storm *Burglind*

395   In this section a full case study for Storm *Burglind* is presented to illustrate the functionality of the new feature detection method RAMEFI. *Burglind* is relatively close to a "textbook" cyclone and shows a feature evolution largely in concordance with the literature. ~~Further examples and a general discussion on strengths and weaknesses of RAMEFI can be found in Sect. 7.~~

### 5.1   Synoptic evolution

400   Storm *Burglind* (~~int.~~ also known elsewhere as $Eleanor$) developed as a secondary cyclone on 02 January 2018 over the North Atlantic and reached the British Isles at the end of that day. The core pressure dropped by more than $27\,\mathrm{hPa}$ in $24\,\mathrm{h}$ and, thus, exceeds the criteria for an explosive cyclogenesis after Sanders and Gyakum (1980). The minimum pressure occurred just east of the English North Sea coast around 03 UTC on 03 January. The cyclone then tracked mostly eastward across the North Sea and Baltic Sea before heading northeastward in later stages (Fig. 2). The cold front crossed France and Germany in the first
405   half of 03 January and caused high winds due to CFC. Ahead of the front, high winds were associated with the WJ. Later, when the occlusion front wrapped around the cyclone centre, the CJ dominated as the cause of high winds in addition to CS winds further away from the cyclone core.

### 5.2   Application of feature identification

  Figure 3 shows the most important parameters, namely $p$, $RR$, $\Delta\tilde{\theta}$, to distinguish the high-wind features for one selected time
410   step, i. e., 03 January 2018, 06 UTC, to illustrate the process of the subjective labelling ~~. In addition to the station data shown in Fig. 3, decisions for the labels were also aided by satellite imagery and synoptic charts (not shown). The set labels are shown in Fig. 3d. The~~ as described in Sect. 4.1. The cyclone centre was located to the east of the UK over the North Sea (red x in Fig. 3). The cold front stretched from northwestern Germany to France (see Fig. D4 in the Appendix). Highest winds were observed ahead of and along the cold front and over western France (contours in Fig. 3). ~~As described in Sect. 4.1, $p$ decreases ahead of~~
415   ~~the cold front and starts to increase again afterwards. This can be seen in~~ Fig. 3a ~~with a strong decrease of mostly more than~~ shows $p$ decreasing strongly with values of mostly above $3\,\mathrm{hPa}$ over southern and eastern Germany ahead of the cold front. The increase after the ~~cold~~ front is a lot weaker with some stations even still showing a weak decrease, which can for example be caused by small-scale processes like convection. Nevertheless, coinciding with the location of the front, several stations observe a $p$ increase of around $2\,\mathrm{hPa}$. This region also coincides with heavy rain with values of around $5\,\mathrm{mm\,h^{-1}}$ to more
420   than $8\,\mathrm{mm\,h^{-1}}$ (Fig. 3b). Slightly lower amounts were observed along the occluded front and northern part of the warm front.
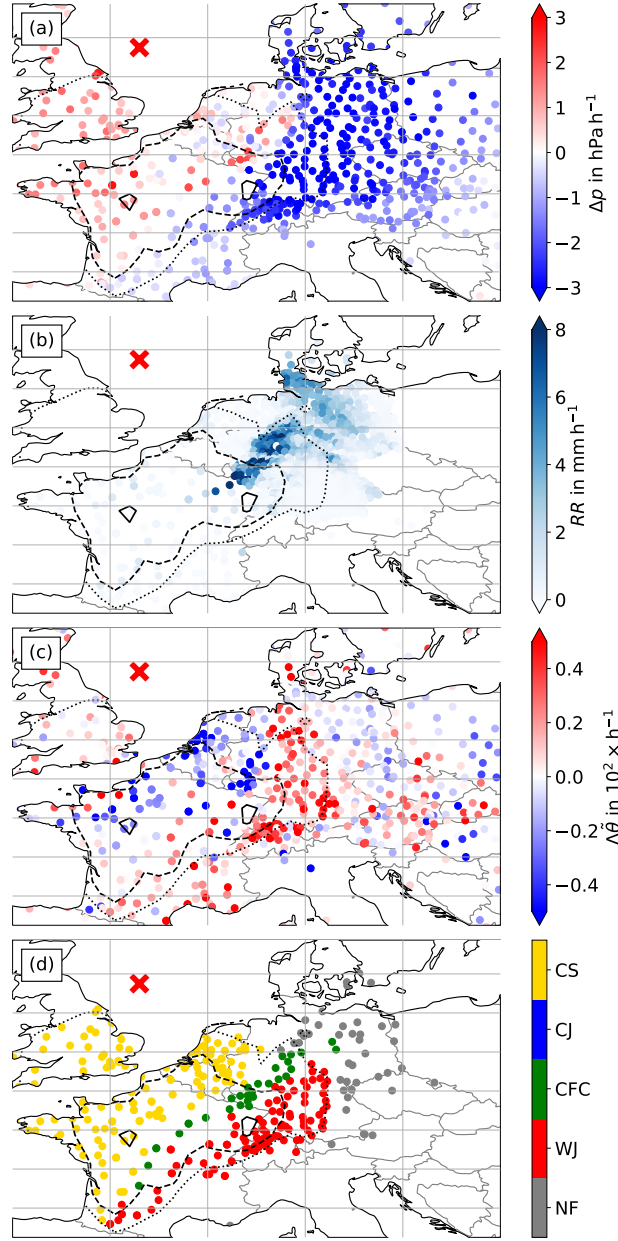
**Figure 3.** Storm *Burglind* on 03 January 2018, 6 UTC. Scattered dots show station observations for (a) $\Delta p$, (b) $RR$, (c) $\Delta\tilde{\theta}$ and (d) ~~set labels~~ subjectively identified wind features for stations where $\tilde{v} \geq 0.8$ and at least three of the five initial parameters are measured (cf. Sect. 4.1). The contours show the interpolated $\tilde{v}$ (dotted $\tilde{v} = 0.8$, dashed $\tilde{v} = 1$, solid $\tilde{v} = 1.2$) to display the regions where the highest winds occurred. The red x indicates the cyclone centre

Furthermore, we note a change in $\tilde{\theta}$ (Fig. 3c) with an abrupt decrease in the frontal region and a shift of $d$ from southwesterly to westerly winds (not shown). While $\Delta\tilde{\theta}$ indicates large tendencies over northern France and western Germany, it shows noisier behaviour further away from the highest winds. Following this, we set labels for a WJ in the region of negative $\Delta p$, positive $\Delta\tilde{\theta}$ and ahead of the high values of $RR$. CFC is labelled in the frontal region, where the heavy precipitation occurred. As ~~signs~~
425 ~~for a CJ are missing~~ the occluded front is not wrapped around the core yet (cf. Fig. D4), implying that a CJ is not yet occurring at this point in time, the region behind the cold front is entirely labelled as CS. All set labels are displayed in Fig. 3d.

The forecasts of feature probabilities by the RF, which is trained on the 11 other cases (Table 2), for 06 UTC and other time steps in a 3 h interval are shown in Fig. 4 after Kriging was applied to generate a gridded field of probabilities. Moreover, $\tilde{v}$ and $p$ contours are added for orientation. An animation for the entire lifetime of *Burglind* is provided in the supplementary material
430 (Eisenstein et al., 2022b).

Comparing Fig. 3d (subjectively identified features) and Fig. 4b (RF predictions) shows the features mostly in consistent areas with high confidence. CFC is identified in a smaller region, which is partly due to missing precipitation observations, since this is the most important variable to predict CFC as will be discussed in Sect. 6.2.

As ~~described in Sect. 2~~stated, the WJ is the first feature to occur during the lifecycle of a cyclone, as it is the case for *Burglind*.
435 Figure 4a shows 03 January 2018, 03 UTC where high probabilities of a WJ are predicted for most of Central France. This region is followed by a smaller region of CFC along the front and CS behind it. As the cyclone evolves further, the cyclone centre and the identified features coherently move further east, while the area affected by the WJ diminishes (Fig. 4b). 3 h later the WJ dissolves north of the Alps, while still being followed by a line of identified CFC (Fig. 4c). At this time step, which is also the time of minimum core pressure, the CJ is identified at the coasts of Belgium and the Netherlands, while the CS is
440 detected further away from the cyclone centre. Highest $\tilde{v}$ are observed along the CFC and remainder of the WJ. The WJ and CFC vanish completely until 12 UTC, when the CJ and also the region of high $\tilde{v}$ extends to western Germany (Fig. 4d) and moves further east following the cyclone centre during the next 3 h (Fig. 4e). At 18 UTC, when the storm and $\tilde{v}$ weaken, the CJ starts to diminish and the probabilities of both CJ and CS decrease as well.

Even though the RF is only dependent on few meteorological parameters and their development over the last hour, looking at
445 all time steps together, the features are largely coherent, both spatially and temporally, and behave as described in the literature (cf. Sect. 2). While the WJ and CFC appear in earlier stages of the lifecycle (Fig. 4a-c), they disappear in later stages, when CJ and CS dominate as the cause of high wind speeds. NF is mostly detected at the ~~edges of the boundary or cyclone in all~~ peripheries of a cyclone or not connected to one at all during all time steps, indicating that most of the high wind speeds in the vicinity of the cyclone are in fact associated with the introduced wind features and identified accurately.

450 **5.3 Comparison to gridded data**

One interesting application of our new algorithm is the comparison of gridded ~~forecasts~~data sets with station observations. To do this successfully, we need to ascertain that such data sets can be fairly compared.

To provide a visual impression of the differences between station- and reanalysis-based results, Fig. 5 shows the example of *Burglind* again, such that it can be directly compared to Fig. 4b. Results are mostly consistent with even higher probabilities
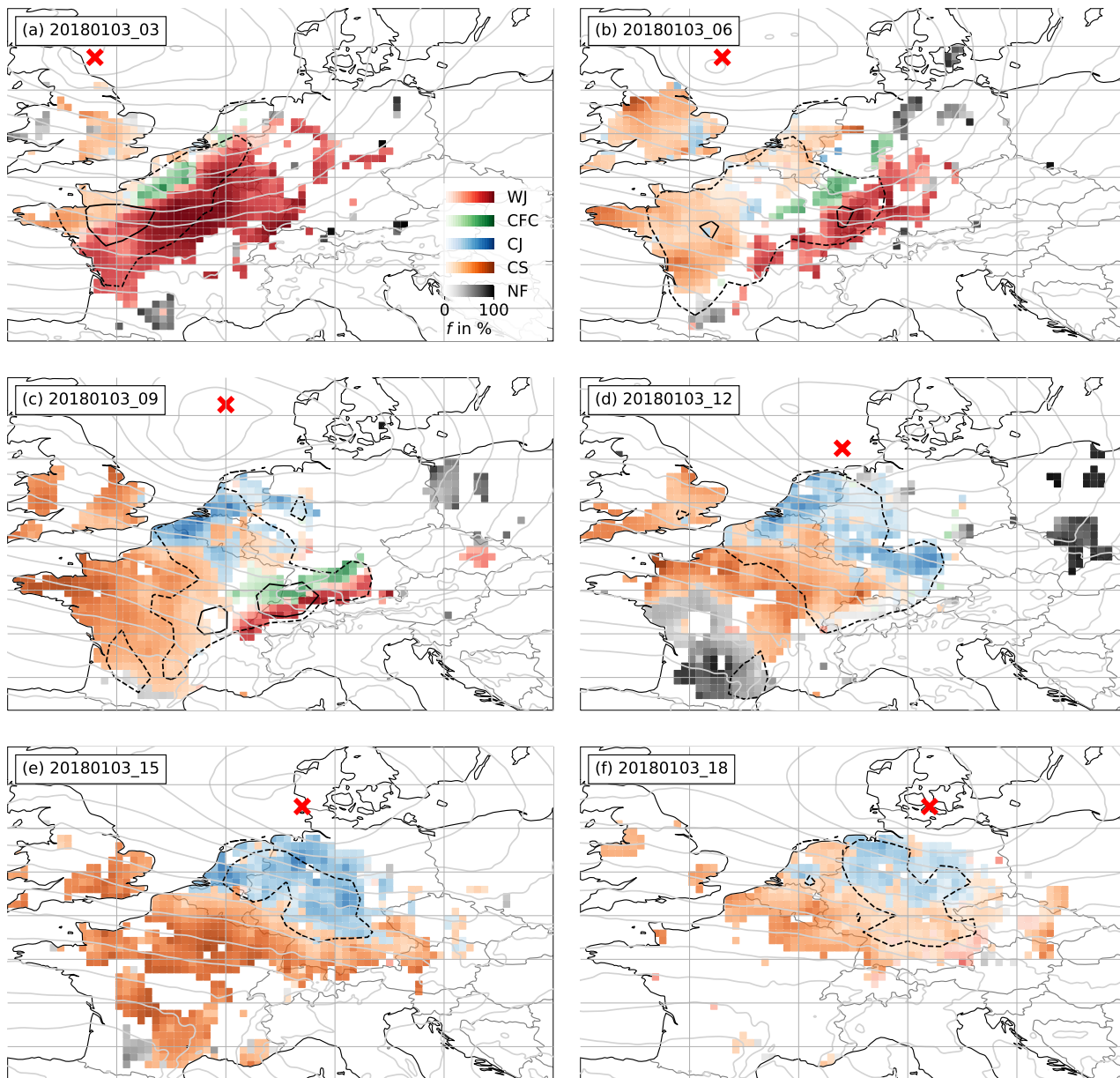
**Figure 4.** ~~RF-derived~~ RF probabilities $f$ for each high-wind feature after Kriging for Storm *Burglind* on 03 January 2018: (a) 03 UTC, (b) 06 UTC, (c) 09 UTC, (d) 12 UTC, (e) 15 UTC and (f) 18 UTC. The dashed contour shows $\tilde{v} = 1$, the solid contour $\tilde{v} = 1.2$). The red x indicates the cyclone centre. Light grey contours show COSMO-REA6 $p$ in a 4 hPa-interval.
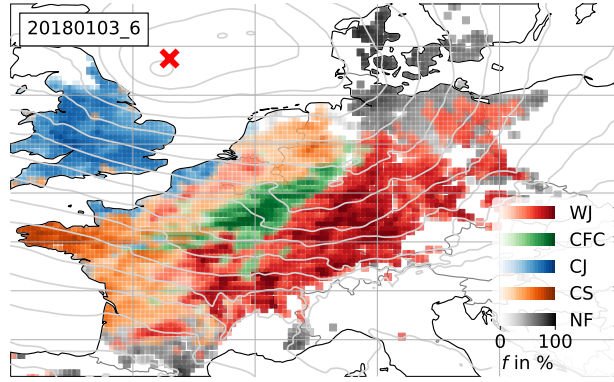
**Figure 5.** RF-derived probabilities $f$ for each feature for Burglind, 03 January 2018, 06 UTC as Fig. 4b but for COSMO-REA6.

most of the time in the reanalysis. Note that COSMO-REA6 provides complete information on a dense regular grid in contrast to the irregularly distributed stations that have to be interpolated. This leads to more coherent areas here. Particularly CFC covers a larger region with higher probabilities. This is due to the spatially higher resolved $RR$ field in COSMO-REA6, the most important parameter for the detection of the CFC feature. Over the UK, high probabilities of a CJ are predicted by the RF. This contrasts with the identified CS in station data and the subjective ~~labelling~~analysis, where a CS was labelled, since no hook-shaped structure of high winds was discernible yet. However, the occurrence of a hook-shaped structure cannot be accounted for in the spatially independent approach of RAMEFI making it difficult to distinguish these otherwise similar features. This is consistent with the more common ~~overforcast~~ overforecast of CJ in the COSMO reanalysis than in observations as will be discussed in Sect. 6.1.

## 6   Statistical evaluation

In ~~this section, we first describe how we generally evaluate the probabilistic forecasts generated by the RF and then~~ Section 3.2, we described how we evaluate probability predictions for the wind features. Here we first apply this concept to the ~~RF-based predictions~~ RF probabilities for the station data and the COSMO reanalysis. At the end of the section, we investigate the relationship ~~of~~ between the predictors and the ~~predictions.~~ RF probabilities.

### 6.1   ~~Assessing multi-class probability forecasts~~

### 6.1   Evaluation of the RF probabilities

~~Multi-class probability forecasts of three or more classes are typically evaluated by downscaling to two-class problems, of which the one-against-all and all-pairs approaches are two well-known examples (Zadrozny and Elkan, 2002). While the one-against-all approach compares the occurrence of one wind feature against all others grouped together, the all-pairs approach considers the conditional probabilities for each pair of classes, for example, the conditional probabilities of the WJ and the CJ~~

475 ~~when one of the two features materialises.~~ The ~~one-against-all approach is used to evaluate how well one specific wind feature is forecast, the all-pairs approach to evaluate the ability to discriminate between two wind features.~~

~~The binary predictions are evaluated based on the paradigm of Gneiting et al. (2007) that a forecast should aim to maximise sharpness subject to calibration. Calibration refers to the consistency of the forecast and the observation, while sharpness is a property of the forecast alone and refers to the associated uncertainty. In a nutshell, a binary probability forecast $f$ is called~~

480 ~~calibrated if the observed relative frequency (given $f$) matches $f$, for example, if a $20\%$ forecast is issued 100 times, the event should occur about 20 times. Further, a probability forecast is said to be sharper, the more certain it is, that is, the closer to 0 or 1. Both calibration and sharpness can be assessed qualitatively via reliability diagrams (Sanders, 1963; Wilks, 2011), which display the calibration curve that is close to the diagonal if the forecasts are calibrated. In addition to the calibration curve, the frequency of the probability forecasts is illustrated by a histogram. The more U-shaped the histogram is, the closer the forecasts~~

485 ~~are to 0 and 1 and thus the sharper. The reliability diagrams shown in this paper are based on the novel CORP-approach of Dimitriadis et al. (2021), which yields optimal calibration curves and eliminates the need of implementation decisions such as the number of bins that the calibration curve is based on. Quantitatively, calibration and sharpness can be assessed using the BS. Here, we compare our multi-class probability forecasts with a climatological benchmark using the multivariate Brier skill score. Details on the evaluation of multi-class and binary probability forecasts are provided in the Appendix.~~

490 ## 6.2 ~~Evaluation of the RF forecasts~~

~~The~~ evaluation of the station-based RF ~~forecasts~~ _probabilities_ is split into three parts. First, we quantitatively compare the RF forecasts with ~~a climatological benchmark forecast~~_the class frequencies in the training data_, then we assess how well the RFs predict the individual wind features in the one-against-all approach and finally we check how well the ~~forecasts~~ _predictions_ distinguish two features with the all-pairs approach. For each ~~fold of the cross-validation scheme~~_storm that we predict_, the

495 class frequencies ~~within the training set~~ _of the other 11 storms_ are used as a ~~climatological forecast~~_benchmark prediction_. As expected, we find that the RF ~~forecasts outperform the climatology~~ _probabilities outperform the benchmark_ in terms of the ~~multivariate~~ BS for the prediction of each winter storm. The overall improvement is $24.7\%$, while for the different storms it ranges from $11.8\%$ to $34.7\%$ with $11.8\%$ being the skill for _Xavier_, which is discussed in some detail in Sect. 7.4.

Figure 6 shows the reliability diagrams of the ~~binary probability forecasts~~ _RF probabilities_ in the one-against-all approach

500 for the occurrence of NF and the four specific wind features (WJ, CFC, CJ and CS). We observe that the ~~probability forecasts~~ _probabilities_ are in general well-calibrated for all five cases, as the calibration curves closely follow the diagonal. The ~~forecasts~~ _predictions_ are generally reliable, especially for small probabilities, which are most frequent in this ~~multi-class~~ setting, as the peaks of the histograms illustrate. Therefore, the RFs identify the non-occurrence of a specific wind feature with high confidence (Fig. 6a). For larger ~~forecast~~ probabilities, the predictions of NF, the WJ and the CFC are well-calibrated, as the

505 calibration curves stay reasonably close to the diagonal (Fig. 6a–c), while for the CJ and CS (Fig. 6d,e) larger deviations are evident. In both cases, the RF overforecasts the events, that is, the predicted probability is generally too large.

The reliability diagrams of the all-pairs approach are displayed in Fig. 7, which show that the RFs yield well-calibrated ~~forecasts~~ _probabilities_ for the distinction of all feature pairs but one. When the RF predicts that the CJ is more likely to occur
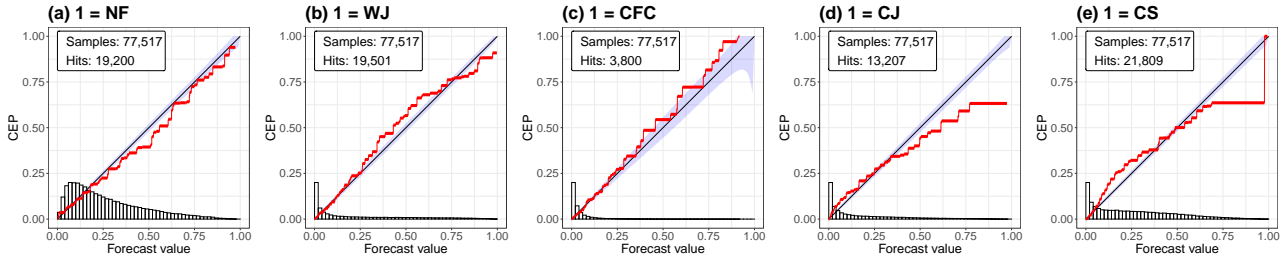
**Figure 6.** CORP reliability diagrams of the ~~probability forecasts~~ RF probabilities for the individual wind features in the one-against-all approach ~~with uncertainty quantification via consistency bands (under the assumption of calibration) at the 90%-level~~including all 12 storms.

than the CS (in case one of those two materialises), the RFs overforecast the CJ, meaning that the CS occurs more often than ~~identified~~ predicted (Fig. 7j). This is consistent with the results from the one-against-all approach, where we found that the CJ and CS ~~forecasts~~ predictions were not well-calibrated for high probabilities, indicating that the RF fails to distinguish them for large conditional probabilities of the CJ. Further, the histogram of this pairwise comparison shows that the RF cannot discriminate between the two features with high confidence. This issue can be seen best for the storms *Herwart* and *Sabine*, which both did not develop a CJ, however a CJ was identified by the RFs (cf. Sect. 7.2). The main meteorological reason for this problem is the general similarity of the two features and that the hook-shaped structure, which is used for the subjective identification of a CJ, cannot be considered in the RF, such that the distinction is mainly based on $p$, as will be discussed in Sect. 6.2. Other than that, the calibration curves of the other pairs follow closely the diagonal. Moreover, we note that the WJ is distinguished well from the CJ and CS, as the U-shaped histograms of the ~~forecast~~ probability distributions show (Fig. 7f,g).

For the predictions derived from the COSMO-REA6 data, the RF ~~forecasts are also~~ probabilities are also well able to distinguish the features~~well~~, although the ~~used RFs~~ RFs used were trained on station-based data. The ~~forecasts~~ predictions exhibit similar characteristics and perform, as expected, only slightly worse than for the station data. As before, the skill of the ~~multivariate~~ BS is calculated with respect to a ~~climatological forecast, which~~ benchmark prediction based on the class frequencies and is 19.6 % for all storms~~is 19.6 %~~. For eight of the selected storms, we observe improvements ranging from 11.0 % to 37.5 %, however for *Herwart* and *Susanna* the skill scores are $-0.8$ % and $-11.6$ %, respectively, indicating a decrease in ~~forecast~~ predictive performance. For *Susanna*, this is due to a larger high-wind region ahead of but not directly connected to the cyclone for multiple time steps. While the predictions for *Herwart* look consistent in both data sets at first sight, less stations over Poland, where the CJ was overforecast (cf. Sect. 7.2), are available, such that the overforcast in the gridded data carries more weight compared to the station data.

Further, we find at times high probabilities of mostly WJ in COSMO-REA6 in regions where winter storms are uncommon and where no features were labelled at all, as e. g., Italy and the Balkans (not shown). However, on the synoptic scale, the trough still affects some parameters in the region, that is, decreasing $\Delta p$ on the eastern side of the core and $d$ as the wind follows the isobars (not shown), which are the most important parameters to distinguish NF and WJ (Fig. 9). Therefore, high
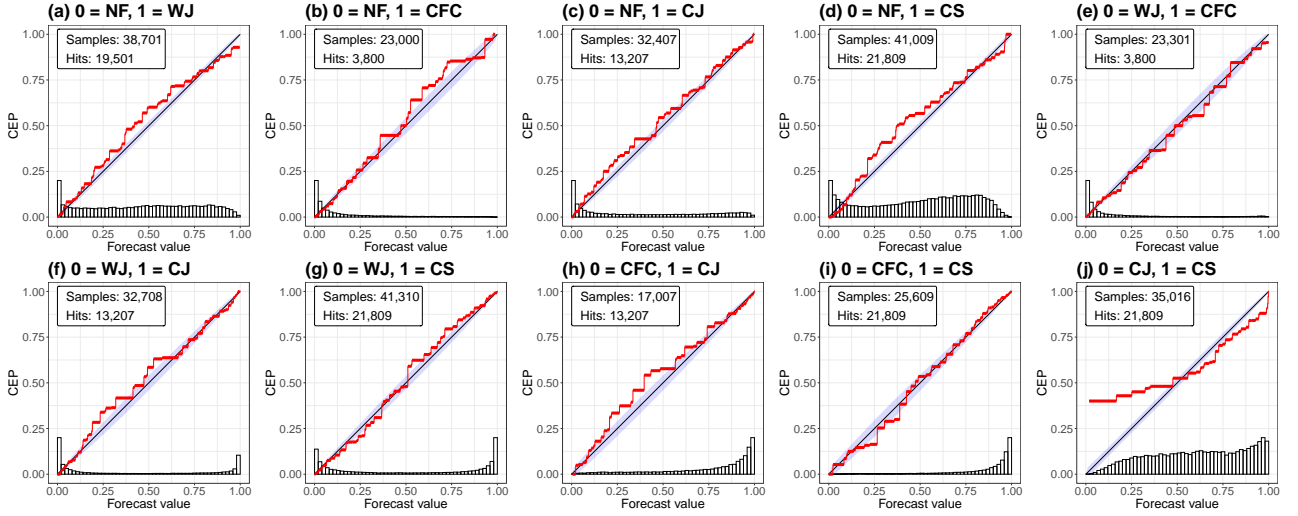
**Figure 7.** CORP reliability diagrams of the conditional ~~probability forecasts~~ RF probabilities comparing two wind features in the all-pairs approach ~~with uncertainty quantification via consistency bands (under the assumption of calibration) at the 90%-level~~ including all 12 storms.

winds caused by, e. g., mountainous effects, such as the foehn effect or land-sea breeze, might be falsely identified as a WJ. Thus, the RF should only be applied to regions affected by extratropical cyclones. As these regions have not been labelled (cf.
535 Sect. 4.1), we excluded them from our evaluation.

The reliability diagrams of the one-against-all approach for the COSMO-REA6 data (see Fig. D2 in the Appendix) show that the calibration curves deviate more from the diagonal than for the station-based data (Fig. 6) but are still reasonably close to calibrated. For the WJ and the CJ, we observe slight overforecasting (Fig. D2b,d), whereas we observe underforecasting for the CFC (Fig. D2c). For the CS, we observe a similar calibration curve to the station-based data (Fig. D2e). The
540 distinction of the individual features, which we assess via the all-pairs approach in Fig. D3, results in mostly well-calibrated ~~forecasts~~probabilities. The largest deviations from calibration are observed again for the distinction of the CJ and the CS, as discussed above, and for the distinction of the WJ and the CFC (Fig. D3e), where the WJ is identified more frequently than observed. This might be caused by a spatially extended area of precipitation further into the warm sector at times due to missing data assimilation for that parameter (cf. Sect. 3.2). Overall, the predictions based on the COSMO-REA6 data are satisfactory
545 considering that the RF models were trained on data from the station observations.

## 6.2 Predictor importance

To identify the predictors most relevant for the prediction of the wind features and the discrimination between two features, we calculate the BS permutation importance for the one-against-all and all-pairs approach. The BS permutation importance in the one-against-all approach is displayed in Fig. 8. In general, $\Delta p$ is the most important predictor variable, especially for the
550 WJ. Only for CFC it is not an important predictor, as it can occur slightly ahead of the cold frontal pressure trough, hence in a
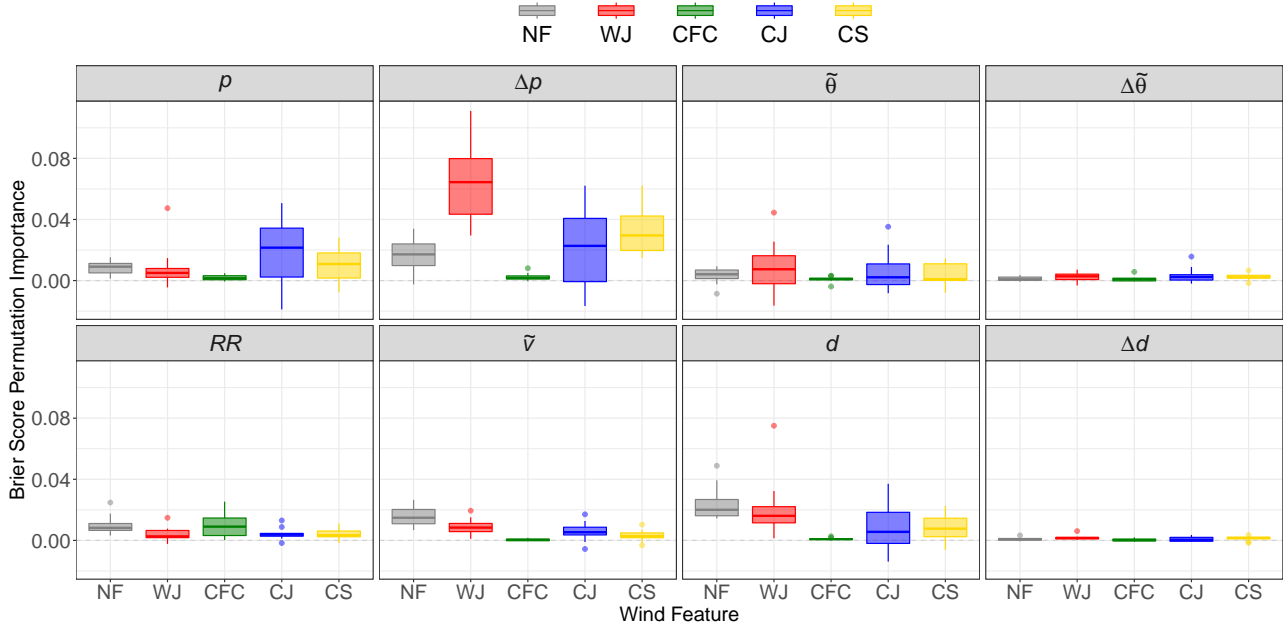
**Figure 8.** Boxplots of the ~~Brier Score~~ BS permutation importance of the ~~probability forecasts~~ RF probabilities for the individual wind features and predictor variables in the one-against-all approach. The boxplots are calculated over the ~~folds of the cross-validation, that is, the~~ individual winter storms.

region of positive $\Delta p$, as described in Sect. 2.4. On the other hand, the absolute $p$ values seem to be of less importance for WJ and NF, which occur further away from the cyclone centre than CJ and CS, for which $p$ indicates the proximity to the cyclone centre. For CFC, we find instead that $RR$ is the most relevant predictor variable as expected, while being less important for WJ, CJ and CS. $d$ seems to be relevant for most features, as it is a characteristic for the location relative to the cyclone centre. This also leads to a high importance for NF occurring more frequently north or west from the cyclone centre. However, $d$ is not important for CFC, probably as convection leads to a more variable wind direction and due to the characteristic jump in $d$ at cold fronts. To the contrary, $\Delta d$ is of minor relevance for all features as well as $\Delta \tilde{\theta}$. A more important temperature-based predictor seems to be $\tilde{\theta}$, although again being less relevant for CFC. Lastly, $\tilde{v}$ shows its highest importance for NF, as higher wind speeds are less likely to be found at the boundary of a cyclone.

In the all-pairs approach (Fig. 9), we can attribute the importance of the predictor variables more accurately. The key to distinguish the WJ from all other features is $\Delta p$, especially from the CJ and CS. This is consistent with the one-against-all discussion above. The large outlier in $\Delta p$ in WJ vs. CJ is related to storm *Herwart* as further discussed in the following section. Of secondary importance is $d$, particularly when compared to CJ, CS and NF. Temperature also plays some smaller role in the distinction of the WJ. For CFC the by far most important predictor is $RR$ but when compared against the CJ $p$, $\Delta p$, $\tilde{\theta}$, $\Delta \tilde{\theta}$ and $d$ also contribute. The positive outlier in $RR$ is related to storm *Fabienne* (not shown). The distinction of the CJ to other
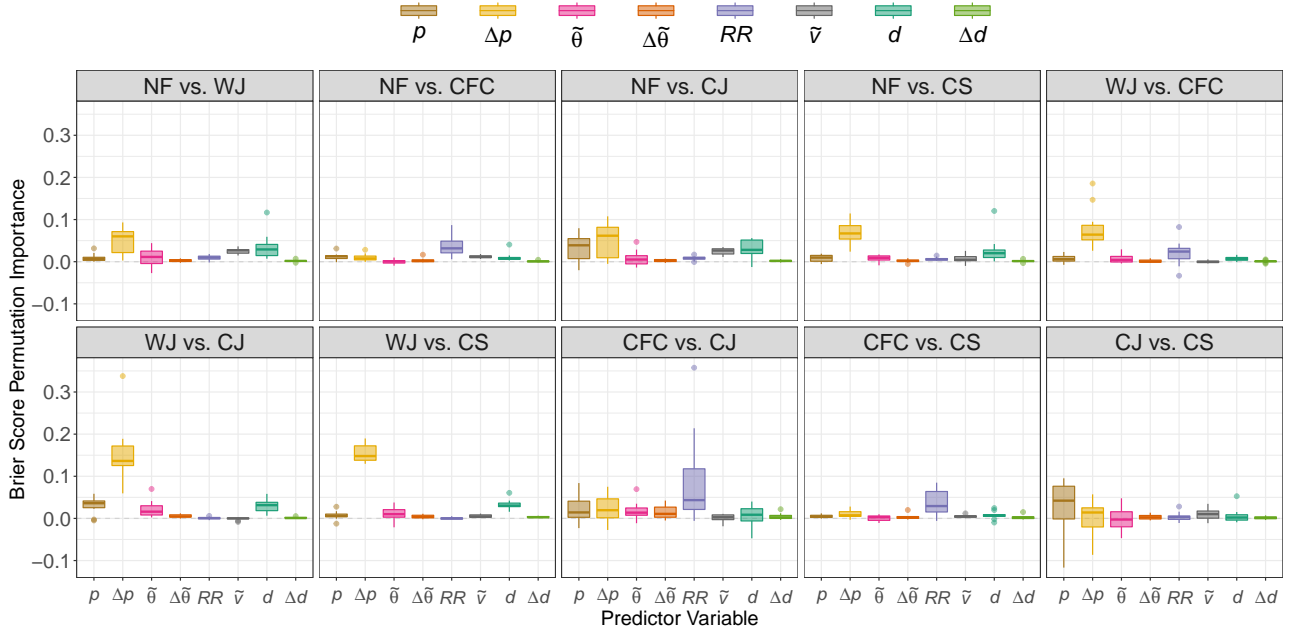
**Figure 9.** Boxplots of the ~~Brier Score~~ BS permutation importance of the ~~probability forecasts~~ RF probabilities comparing two wind features for the predictor variables in the all-pairs approach. The boxplots are calculated over the ~~folds of the cross-validation, that is, the~~ individual winter storms.

features is more complex. $p$ is relevant in all CJ-pairs, as already discussed. The distinction of CJ from NF additionally hinges upon $\Delta p$, $\tilde{\theta}$ and $d$

The shortcomings of the RFs to distinguish CS and CJ is also reflected in Fig. 9 by partly negative values for $p$, $\Delta p$ and $\tilde{\theta}$. A negative value indicates that the RF ~~forecasts~~ probabilities perform better, when we break the link to the target variable
570 by randomly permuting the predictor values. As discussed further in the following section, this is mostly due to storm *Sabine*, which reached an unusually low minimum core pressure of less than $950\,\mathrm{hPa}$ over the Norwegian Sea (cf. Fig. 2). Because of this, $p$ values in the CS over continental Europe were similar to values typical of a CJ.

We do not only want to identify the most relevant predictors, but also investigate their effect on the predictions, which is illustrated for the eight predictor variables by the PDPs in Fig. 10. Again, the largest impact is found for $\Delta p$. The probability
575 of observing a WJ is largest for small values of $\Delta p$ and declines rapidly as the tendency increases and switches signs, while the probabilities of the CS and CJ increase. Probabilities for NF decrease slightly, while changes for CFC are small. For little $RR$ the probability of a CFC is close to zero, but consistently increases with increasing precipitation. In turn, probabilities for other features slightly decrease with increasing precipitation. In general, CJ and CS show high probabilities for low $p$ values consistent with their occurrence during the most intense stage of a cyclone (cf. Sect. 2). However, surprisingly CS
580 shows higher probabilities than CJ between $970\,\mathrm{hPa}$ to $980\,\mathrm{hPa}$, although the CJ is usually closer to the cyclone centre. This is again associated with the unusual behaviour of storm *Sabine* with its deep pressure minimum but no subjectively identified
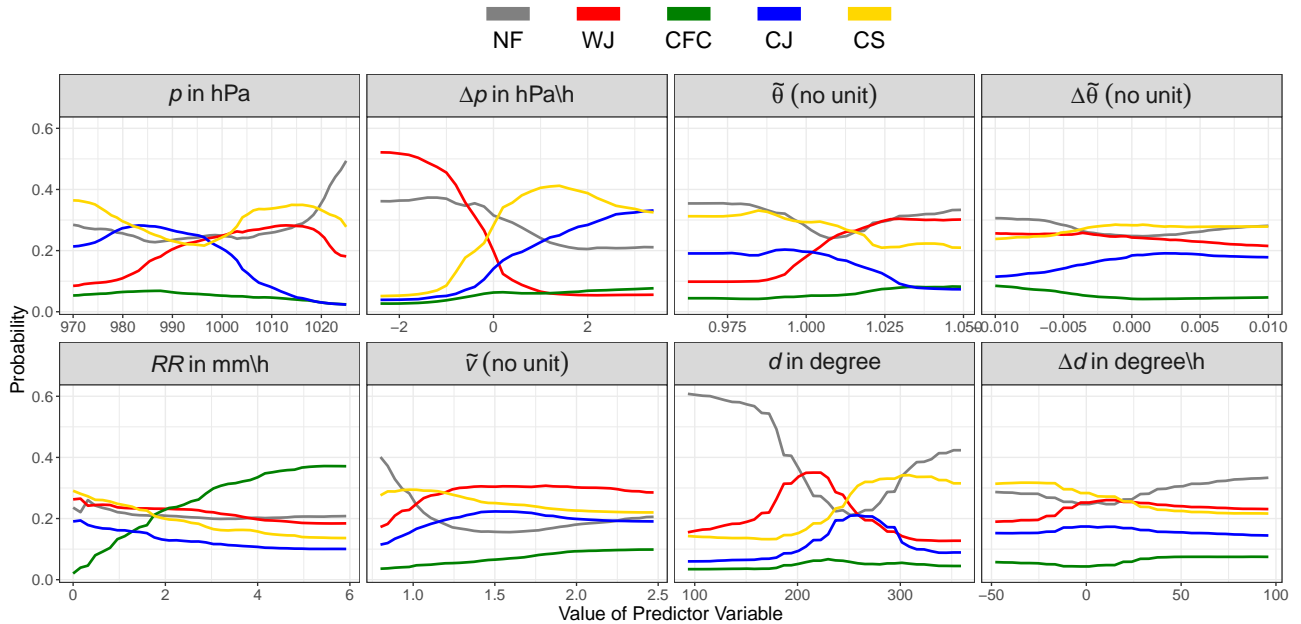
**Figure 10.** Partial dependence plots for the predictor variables and wind features.

CJ. As such intense cyclones are rare, we are confident that the RF performs well in most more ordinary cases. As discussed previously, $d$ is dependent on the location relative to the cyclone centre. As the introduced features are all located south to west of the cyclone, we focus on values from $90°$ to $360°$ only. Within the warm sector $d$ values mostly show south-westerly winds and do not change drastically. Probabilities for CFC increase with a positive wind shift, leading to more westerly and north-westerly winds for CFC but also following features, i. e., CJ and CS. $\Delta d$ shows almost no change in probabilities for all features consistent to its low BS permutation importance. $\tilde{\theta}$ shows an increasing trend for the WJ, while the probabilities decrease for the other features, most strongly for the CJ, as one would expect. For $\Delta\tilde{\theta}$ we see indications of the air mass change at the cold front and thus higher probabilities in CFC for negative values. The CJ shows a slightly positive trend, while all the others are flat.

Overall, investigating the importance of the predictor variables on the predictions, we find that the RFs largely learn physically consistent relations as described in Sect. 2 and Sect. 4.1.

## 7   Discussion

Section 5 showed a successful application of the introduced method to storm *Burglind* both in station- and grid-based data sets. However, extratropical cyclones rarely follow the textbook NC model exactly. For example, sometimes high winds are ~~related to a~~ mainly related to an exceptionally strong synoptic-scale pressure gradient rather than associated with the mesoscale features that we have developed an objective identification algorithm for.

As described in Sect. ~~??~~4, we selected 12 windstorm case studies to train the RF based on surface observations with the diversity of cyclones and features in mind, such that the RF is representative for a climatology over a longer time period. ~~General information about these storms can be found in Table 2, the cyclone tracks and core pressure evolutions are shown in Fig. 2.~~

This section discusses how the trained RF deals with some well-known deviations from idealised cyclone models such as double fronts and convergence lines (Sect. 7.1), large background pressure gradients (Sect. 7.2), as well as the specific characteristics of SKC and SJ (Sect. 7.3). In addition, we further discuss advantages and disadvantages of not using spatial dependencies in the feature identification (Sect. 7.4). A complete set of results for all case studies can be accessed in the supplementary material (Eisenstein et al., 2022b).

## 7.1 Double fronts and convergence lines

Real-world frontal structures of extratropical cyclones can differ considerably from idealised conceptual models, e. g., in terms of complex vertical structure, strong tilts, or a secondary frontal zone parallel to the main front. Here, we are interested in synoptic systems with double cold fronts and convergence lines with high winds. While a cold front is associated with a second low pressure trough, a convergence line develops where two airflows collide and can occur independently of a cyclone. The area between a primary and secondary cold front can have characteristics of a warm and/or cold sector and thus high-wind features are predicted with higher uncertainty by the RF. This can also be the case for the area between a cold front and a convergence line.

The example from our selection of 12 cyclones that illustrates this best is *Bennet*. On 04 March 2019 at 12 UTC the cyclone centre was located over the North Sea to the west of Denmark (Fig. 11a). The primary front is located at the north-eastern border of Germany and Bavaria and has already weakened (see Fig. D5 in the Appendix). A secondary strong temperature gradient can be found over northwestern Germany, Luxembourg and France. However, it is uncertain, if this should be classified as a front or convergence line. While synoptic charts from DWD show a convergence line, Met Office charts show an upper-level cold front at 06 UTC and an occluded front $6\,\mathrm{h}$ later. As this feature shows characteristics typical of CFC (cf. Sect. 2.4), it was ultimately labelled as such by the first author and is referred to as a secondary front.

Figure 11a shows the mesoscale wind features identified by the RF. At this time *Bennet* causes ~~rather weak wind speeds, indicated by the small area of $\tilde{v} > 1$~~ its highest winds slightly ahead of the secondary front and a smaller region behind the primary front (see dashed line in Fig. 11a). For the area between the fronts, the RF predicts both CS and WJ with medium confidence. This is to be expected, as some parameters, especially $\Delta p$, $\tilde{\theta}$ and $\Delta\tilde{\theta}$, show behaviour of both features with a tendency towards CS. RF predictions along the secondary front ~~/convergence line~~ show only low probabilities for CJ and CFC. However, at earlier and later timesteps, that is, ahead of the primary cold front and behind the secondary one, the prediction of the WJ, CJ and CS are accurate, as can be seen in the full animation provided in the supplementary material (Eisenstein et al., 2022b). So looking at the entire lifetime of the storm, satisfactory identification can be obtained from RAMEFI.
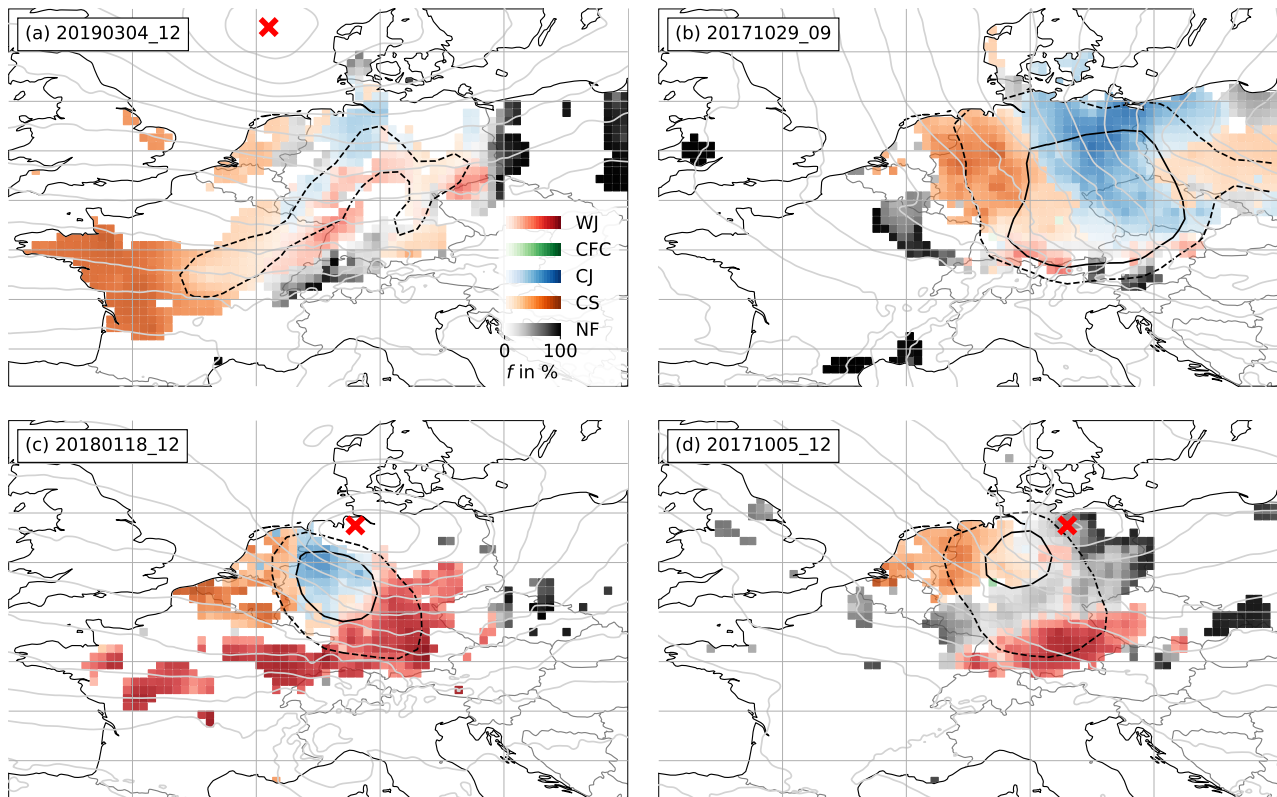
**Figure 11.** As Fig. 4 but for (a) storm *Bennet*, 04 March 2019, 12 UTC; (b) storm *Herwart*, 29 October 2017, 09 UTC; (c) storm *Friederike*, 18 January 2018, 12 UTC; and (d) storm *Xavier*, 05 October 2017, 12 UTC. Note that at the time shown, storm *Herwart* had already exited the plot area to the east.

## 7.2 Strong background pressure gradients

Very intense cyclones are often accompanied by a strong large-scale pressure gradient (e. g., Fink et al., 2009), which in turn causes high wind speeds unconnected to one of the four mesoscale wind features under study ~~(or enhance the wind speeds associated with one~~ but can be enhanced by them. With an underlying strong wind field, the detection and distinction of the features ~~)~~might be more complicated. Good examples from our list of case studies to illustrate this are the storms *Herwart* and *Sabine*. Figure 11b shows *Herwart* in a late development stage on 29 October 2017 at 09 UTC when the pressure gradient is already weakening (see light grey contours). Around this time, the cyclone centre travelled over Poland (outside of the area shown in the figure). An occurrence of a CJ seems unlikely in that region, since a typical hook-shape structure cannot be seen in wind observations (not shown). The occluded front was rather weak and did not fully wrap around the cyclone centre (see Fig. D6 in the Appendix). So ultimately the ~~highest wind speeds are enhanced~~ already high wind speeds caused by a strong background pressure gradient over Germany (black lines in Fig. 11b) ~~, making~~ make the subjective labelling of additionally

occurring features quite challenging. The RF shows high probabilities of a CJ for several hours in this region, although it was originally labelled to be CS. The main reason for this is that the proximity to the cyclone centre, reflected in low $p$, is the most important predictor to distinguish the CJ and CS (Fig. 9). Nevertheless and even in this unusual case, the prediction by the RF is still reasonable, since both a subjective and objective identification of the two features here is ambiguous in surface
645 observations alone.

In the case of *Sabine* the cyclone centre did not cross continental Europe but moved through the Norwegian Sea (Fig. 2). The minimum pressure reached less than $950\,\mathrm{hPa}$. Stations over Central Europe still observed $p$ below $970\,\mathrm{hPa}$, which is lower than the cyclone centres of most of our other case studies, making *Sabine* a quite unusual case. As discussed already in Sect. 6, this causes difficulties to distinguish the CJ and CS of *Sabine*, somewhat similar to *Herwart*. Although a CJ is identified in an
650 area of low $p$ values, this region is not in the vicinity of the cyclone centre, as it was the case for *Herwart*. An animation of the feature identification for all time steps of *Sabine* is provided in the supplementary material (Eisenstein et al., 2022b). In this case the CJ-CS distinction issue could have been avoided to some extent by including spatial dependencies in the identification algorithm at the expense of losing the capability for flexible application as discussed in Sect. ??4. We will return to this issue in Sect. 7.4.

655 ### 7.3  Shapiro-Keyser cyclones and sting jets

As already explained in Sect. 2, cold fronts of SKC are usually weaker than those of NCs, such that CFC wind features hardly occur. A good example for this is storm *Friederike*. Figure 11c shows 12 UTC on 18 January 2018, when the cyclone centre just reached the North Sea coast of Germany. High probabilities of a WJ occur over Central Germany, while a CJ is identified by the RF over northeastern Germany. CFC probabilities are very low along the entire cold front. In the western part of the high
660 wind area (see black lines in Fig. 11c) a CS feature is detected. Lagrangian trajectories confirmed an SJ in the region where the CJ is identified during that time. As the CJ and SJ are considered together in this work (cf. Sect. 4.1), the area was labelled as CJ and, hence, identified accurately. This area also coincides with the highest wind speeds associated with Friederike (see black lines in Fig. 11c). West of the identified CJ region, a CS feature is detected. This shows that CJ and SJ indeed show similar characteristics in surface parameters and can be considered as one feature in this context.

665 ### 7.4  Spatial independence

The decision not to use spatial (nor temporal beyond $1\,\mathrm{h}$) dependencies in the identification algorithm makes our method highly flexible in its application but the local approach can also cause issues where features deviate from their stereotypical characteristics. One example for the problem is the CJ of *Sabine* as discussed in Sect. 7.2. Another example is storm *Xavier*, where for several hours many points within the vicinity of the cyclone show the highest probability for NF rather than for any
670 of the mesoscale wind features (Fig. 11d). The main reason for this appears to be that *Xavier* was characterised by unusually cool $\tilde{\theta}$ and high $p$ (not shown), generally two of the most important parameters to distinguish features (Fig. 8). In such a case, normalising $p$ by the core pressure could help detecting the region close to the cyclone centre, while the warm sector can be detected more reliably by comparing $\tilde{\theta}$ to the mean state over Europe during the period of the storm. While one predictor

behaving in an unusual way could be compensated, as e. g., in the case of *Fabienne* (cf. Eisenstein et al., 2022b), two anomalous behaviours unsurprisingly result in considerably greater uncertainty.

A possible solution to the issues described here on the basis of *Sabine*, *Xavier* and *Fabienne* is to not only regard anomalies from diurnal and seasonal cycles but also to include some kind of spatial background, e. g., by ~~subtracting a spatial mean~~ normalising $p$ by the core pressure to detect the region close to the cyclone centre, or comparing $\tilde{\theta}$ to the mean state over Europe during the period of the storm to detect the warm sector. However, such a step would bring its own set of problems. Any spatial mean would require an arbitrary decision about the considered area, which may vary greatly from cyclone to cyclone. Moreover, spatial means computed from surface observations are not representative due to the irregular spacing of the stations. Essentially, as the features identified by the RF still occur in the expected areas as described in Sect. 2, we conclude that a flexible local approach offers more advantages than disadvantages overall.

## 8   Conclusions

High wind and gust speeds can be caused by distinct mesoscale features within extratropical cyclones, which occur during different stages of the cyclone lifecycle, in varying regions relative to the cyclone centre and have distinctive meteorological characteristics (e. g., Hewson and Neu, 2015). These differences likely imply differences in hazardousness, forecast errors and, hence, risk to life and property.

To better understand, monitor and predict these mesoscale features, we developed RAMEFI, a first-ever objective identification method that is able to reliably distinguish the four most important features, that is, the WJ, the CJ, CFC and CS. The rare and often short-lived SJ is included in the CJ category, as their surface characteristics are often rather similar and 3D trajectories are required for a clean distinction (Gray et al., 2021).

The first step was to build a browser-based, interactive tool to subjective label surface stations over Europe for 12 selected winter storm cases between 2015 and 2020. Based on the outcome, we trained a probabilistic RF based on the eight predictors $\tilde{v}$, $p$, $\Delta p$, $\tilde{\theta}$, $\Delta\tilde{\theta}$, $RR$, $d$ and $\Delta d$. Being independent of spatial behaviour or gradients and only requiring $1\,\mathrm{h}$ tendencies, the approach is very flexible and can be applied to single stations or grid points and various data sets with differing grid spacing. To obtain areal information from irregular station data, Kriging was applied on the station-based probabilities generated by the RF.

The trained RFs are generally well-calibrated. Merely, the distinction between CJ and CS is more challenging, since the two features show similar characteristics in most parameters except for lower $p$ in the CJ being located nearer to the cyclone centre. Overall, the RFs learn physically consistent relations reflected in the importance of individual predictors. For example, while $\Delta p$ appears to be most important for WJ, CJ and CS, $RR$ is substantial for the identification of CFC.

A detailed analysis of the RF feature probabilities for the selected cases shows a high consistency with the subjectively set labels with only few disagreements, mostly in cases of large deviations from standard cyclone models. While the identification of WJs has the highest confidence, the identification of CFC is least certain due to relatively few surface stations reporting hourly precipitation and thus less training data. Even the distinction between the relatively similar CS and CJ works well in

most cases and time steps. In some cases, however, high probabilities of CJs are predicted by the RF in areas where no CJ was identified subjectively due to a missing hook-shaped structure and occlusion front, or too large distance from the cyclone centre (e. g., *Herwart*, *Sabine*). Despite the spatial independence of the method, putting the predicted probabilities together on a horizontal map and following the storm evolution in time shows a high degree of coherence for each feature, demonstrating the success of our method.

The station-based RFs are also applied to COSMO reanalysis data without any adaptations to the new data set. Nevertheless, the obtained results are mostly consistent and only slightly less calibrated. This demonstrates that the method ~~should~~ could be readily applicable to other analysis and forecast datasets. Although applying RAMEFI over regions other than that used in the training has not been examined yet, relying on location-independent predictors suggests that it should be possible with no or only little modification.

Now that the RAMEFI method is fully developed~~and tested~~, it enables a number of exciting follow-on studies (see Part II). The next step we plan is to use the objective identification approach to compute a long-term climatology over Europe based on station observations and COSMO reanalysis data. Although, previous literature discussed different causes of winds within extratropical cyclones, their climatologies were based on more subjective categorisations for a limited sample size (e. g., Hewson and Neu, 2015; Earl et al., 2017). RAMEFI will for the first time allow a statistically substantiated analysis of the characteristic of the mesoscale wind features in terms of size, lifetime, position relative to the cyclone core, occurrence relative to the lifecyle of the cyclone and wind characteristics. Furthermore, a systematic forecast error analysis will reveal to what extent forecast errors differ between the identified features and whether there any significant, systematic deficits in their representation in models. Based on the outcome we plan to subsequently work on a feature-dependent ~~postprocessing~~ post-processing approach using the methods discussed in Schulz and Lerch (2022). This can ultimately improve the forecasts of strong winds and gust, and thus support the provision of better warnings regarding high wind gusts and of ultimately timely forecasts of the associated impacts.

*Code availability.* The interactive visualisation and labelling tool, as well as the trained RF and Kriging code are available at https://gitlab. physik.uni-muenchen.de/Lea.Eisenstein/ramefi, where it will be updated in future studies, and is archived at https://doi.org/10.5281/zenodo. 6541303 (Eisenstein et al., 2022a).

*Data availability.* COSMO-REA6 data are available under https://reanalysis.meteo.uni-bonn.de (Hans-Ertel-Centre for Weather Research, 2019). Daily maximum gusts over Germany used for the computation of the SSI are available under https://cdc.dwd.de/portal/ (Deutscher Wetterdienst). Further observation data over Europe was provided by DWD for this work and cannot be made freely available. Please contact the DWD directly regarding this data

*Video supplement.* A video supplement showing the probability maps for all 12 selected case studies (Table 2) can be freely accessed at https://doi.org/10.5281/zenodo.6541277 (Eisenstein et al., 2022b).

## Appendix A: Implementation details

740  In this section, we provide technical details on the ~~probabilistic RFs, the Kriging and the evaluation of the RF forecasts.~~ evaluation of the RF probabilities, the probabilistic RFs and the Kriging.

### A1  Evaluation of the RF probabilities

The CORP reliability diagrams were generated using the `reliabilitydiag`-package (Dimitriadis et al., 2021), the PDPs using the `pdp`-package (Greenwell, 2017). Each reliability diagram is based on the RF probabilities and observations of all storm cases. For the PDPs, one partial dependence curve has to be calculated for each RF generated in a fold of the

745  cross-validation, that is, for each winter storm. The final curves are then obtained by a weighted average depending on the sample size of the folds.

### A2  Random Forests

RF is implemented via the `ranger`-package (Wright and Ziegler, 2017) in R (R Core Team, 2021). Table 1 summarises the predictors used, Table A1 the chosen ~~hyperparameters~~ tuning parameters. One question in the implementation is the handling

750  of missing values, which an RF cannot process. The station-based samples frequently miss values of one or more predictor variables, especially precipitation is affected. We tried different strategies to handle missing values such as leaving out instances with missing values or replacing the missing values with a mean value and found similar results. Therefore, we decided to replace the missing values in order to use the largest sample size possible, which is desirable for the evaluation and the Kriging step. In each fold of the employed cross-validation scheme, the missing values (both in the training and test set) are replaced

755  by the mean value of the associated predictor variable in the training set.

**Table A1.** Overview of the ~~hyperparameters~~ tuning parameters of the probabilistic RF.

| ~~Hyperparameter~~ Tuning Parameter | Value |
| --- | --- |
| Number of trees | 1,000 |
| Number of predictors considered at each split | 2 |
| Minimal node size | 10 |
| Maximal depth | Unlimited |
| Splitting criterion | Gini |

## A3   Kriging

In our practical implementation of Kriging, we employ the R-packages ~~:~~`fields` (Douglas Nychka et al., 2017), `mvtnorm` (Genz et al., 2021), `maps` (Original S code by Richard A. Becker and Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn., 2018) and `maptools` (Bivand and Lewin-Koh, 2021). Additionally, we transform the probabilities by using the `bestNormalize` package (Peterson, 2021) to achieve approximate Gaussianity, which automatically chooses a suitable transformation from a set of commonly used transformations. The probabilities on the grid generated via the univariate Kriging need to be normalised such that they sum up to 1. However, at some grid cells distant from the cyclone track, the predicted probabilities are small for all of the wind feature and normalisation results in unrealistic predictions. Thus, we only perform the normalisation at grid cells where the accumulated probability is larger or equal to 20 %. For the visualisation, we further drop the grid cells where the largest normalised probability is smaller than 20 % (which includes the grid cells for which no normalisation was performed).

## A4   ~~Evaluation of the RF forecasts~~

~~The CORP reliability diagrams were generated using the `reliabilitydiag`-package (Dimitriadis et al., 2021), the PDPs using the `pdp`-package (Greenwell, 2017). Each reliability diagram is based on the probability forecasts and observations of all storm cases. For the PDPs, one partial dependence curve has to be calculated for each RF generated in a fold of the cross-validation, that is, for each winter storm. The final curves are then obtained by a weighted average depending on the sample size of the folds.~~

## Appendix B: Mathematical details

In this section, we provide a more detailed description of the ~~Kriging approach and the~~ assessment of probability ~~forecasts.~~ predictions and the Kriging approach.

## B1   ~~Kriging~~

~~Let $\{X(\mathbf{s}),\ \mathbf{s} \in \mathbb{R}^2\}$ be the spatial Gaussian process modelling the transformed probability of a certain wind feature, indexed by the spatial coordinates $\mathbf{s}$ that correspond to the latitude and longitude associated with the (transformed) probability. Further, we denote the mean function by $\mathbb{E}\{X(\mathbf{s})\} = \mu(\mathbf{s})$ and the covariance function by $\mathrm{Cov}\{X(\mathbf{s}), X(\mathbf{s}')\} = C(\mathbf{s}, \mathbf{s}')$. Then, for a given set of station-based data $\mathbf{x} = \{X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n)\}^T$, the spatial prediction at a grid cell $\mathbf{s}_0$ is given as $\hat{X}(\mathbf{s}_0) = \mu(\mathbf{s}_0) + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n)\}^T$, $\Sigma_{12} = \{C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n)\}$ and $\Sigma_{22} = \{C(\mathbf{s}_i, \mathbf{s}_j)\}_{i,j=1}^n$. Additionally, one can obtain the prediction variance as $\mathrm{Var}\{\hat{X}(\mathbf{s}_0)\} = C(\mathbf{s}_0, \mathbf{s}_0) - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$, and the full predictive distribution as $X(\mathbf{s}_0) \mid \mathbf{x} \sim \mathcal{N}\big(\hat{X}(\mathbf{s}_0), \sqrt{\mathrm{Var}\{\hat{X}(\mathbf{s}_0)\}}\big)$~~

~~The choice and estimation of the mean function $\mu(\cdot)$ and the covariance function $C(\cdot, \cdot)$ are key elements of the Kriging implementation. While one can choose any parametric or nonparametric functional representation for $\mu(\cdot)$, the valid choice~~

for $C(\cdot,\cdot)$ is limited to the class of positive semidefinite functions. In practice, the covariance function is often assumed to be stationary, which implies that the covariance function depends on the spatial locations only through spatial lags, i.e., $C(\mathbf{s}_i,\mathbf{s}_j) = K(\mathbf{s}_i - \mathbf{s}_j)$ for some positive semidefinite function $K(\cdot)$. In our implementation, we have specified the mean function $\mu(\cdot) = c$, $c \in \mathbb{R}$ to be a constant valued function, and the covariance function $K(\cdot)$ to be the Matérn class of stationary covariance function (Matérn, 1986).

## B1  Assessing probability predictions of multiple classes

Consider a multi-class probability prediction $\mathbf{f} = (f_1,\ldots,f_k)$, where $f_1,\ldots,f_k \in [0,1]$ and $\sum_{i=1}^{k} f_i = 1$, for a nominal target variable $Y \in \{1,\ldots,k\}$ with $k \geq 3$ classes that are not ordered. In mathematical terms, $\mathbf{f}$ is called (auto-)calibrated if $\mathbb{P}(Y = i \mid \mathbf{f}) = f_i$ for all $i = 1,\ldots,k$, where $\mathbb{P}$ refers to the joint distribution of prediction and observation (Gneiting and Ranjan, 2013). The BS for a probability vector $\mathbf{f}$ and the realising class $i \in \{1,\ldots,k\}$ is given by $S(\mathbf{f},i) = \sum_{j=1}^{k}(f_j - \mathbb{1}\{i = j\})^2$ (Brier, 1950). Now, consider a probability prediction $f \in [0,1]$ for a dichotomous target variable $Y \in \{0,1\}$. In mathematical terms, $f$ is called calibrated if $\mathbb{P}(Y = 1 \mid f) = f$ (Gneiting and Ranjan, 2013). The (binary) BS of a probability prediction $f$ and realisation $y \in \{0,1\}$ reduces to $S(f,y) = (f-y)^2$. Both the binary and multivariate BS can be used to evaluate the improvement of a prediction method over a reference using the BSS. Given the mean BS of the predictions of interest $\bar{S}_f$ and that of the reference predictions $\bar{S}_{ref}$, the skill score is calculated via $SS_f = 1 - \bar{S}_f/\bar{S}_{ref}$. Positive skill indicates improvement over the reference with $100\%$ referring to a perfect prediction, $0\%$ to no improvement and a negative skill to a decrease in predictive performance with respect to the reference.

The one-against-all approach reduces the multi-class prediction problem to a set of $k$ dichotomous problems. For each class $i \in \{1,\ldots,k\}$, the probability $f_i$ is a prediction for $\tilde{Y} = \mathbb{1}\{Y = i\}$, where $\mathbb{1}$ denotes the indicator function. Note that evaluating the predictions $f_i$ for $\tilde{Y}$ for each class is not equivalent to checking the multi-class calibration criterion, as the joint distribution of $f_i$ and not that of $\mathbf{f}$ is considered in the one-against-all approach. The all-pairs approach reduces the multi-class prediction problem to a set of $k(k-1)/2$ dichotomous problems. For each pair of classes $(i,j)$, where $i,j \in \{1,\ldots,k\}$ and $i > j$, we consider only samples with $Y \in \{i,j\}$. Then, the conditional probability $\tilde{f}_{i,j}$ is a prediction for $\tilde{Y}$, where

$$\tilde{f}_{i,j} = \frac{f_i}{f_i + f_j} \quad \text{and} \quad \tilde{Y} = \begin{cases} 1 & \text{for} \quad Y = i, \\ 0 & \text{for} \quad Y = j. \end{cases}$$

## B2  Kriging

Let $\{X(\mathbf{s}), \mathbf{s} \in \mathbb{R}^2\}$ be the spatial Gaussian process modelling the transformed probability of a certain wind feature, indexed by the spatial coordinates $\mathbf{s}$ that correspond to the latitude and longitude associated with the (transformed) probability. Further, we denote the mean function by $\mathbb{E}\{X(\mathbf{s})\} = \mu(\mathbf{s})$ and the covariance function by $\text{Cov}\{X(\mathbf{s}),X(\mathbf{s}')\} = C(\mathbf{s},\mathbf{s}')$. Then, for a given set of station-based data $\boldsymbol{x} = \{X(\mathbf{s}_1),\ldots,X(\mathbf{s}_n)\}^T$, the spatial prediction at a grid cell $\mathbf{s}_0$ is given as $\hat{X}(\mathbf{s}_0) = \mu(\mathbf{s}_0) + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$,

where $\boldsymbol{\mu} = \{\mu(\mathbf{s}_1), \ldots, \mu(\mathbf{s}_n)\}^T$, $\Sigma_{12} = \{C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n)\}$ and $\Sigma_{22} = \{C(\mathbf{s}_i, \mathbf{s}_j)\}_{i,j=1}^n$. Additionally, one can obtain the prediction variance as $\mathrm{Var}\{\hat{X}(\mathbf{s}_0)\} = C(\mathbf{s}_0, \mathbf{s}_0) - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$, and the full predictive distribution as $X(\mathbf{s}_0) \mid \boldsymbol{x} \sim \mathcal{N}(\hat{X}(\mathbf{s}_0), \sqrt{\mathrm{Var}\{\hat{X}(\mathbf{s}_0)\}})$.

The choice and estimation of the mean function $\mu(\cdot)$ and the covariance function $C(\cdot, \cdot)$ are key elements of the Kriging implementation. While one can choose any parametric or non-parametric functional representation for $\mu(\cdot)$, the valid choice for $C(\cdot, \cdot)$ is limited to the class of positive semidefinite functions. In practice, the covariance function is often assumed to be stationary, which implies that the covariance function depends on the spatial locations only through spatial lags, i.e., $C(\mathbf{s}_i, \mathbf{s}_j) = K(\mathbf{s}_i - \mathbf{s}_j)$ for some positive semidefinite function $K(\cdot)$. In our implementation, we have specified the mean function $\mu(\cdot) = c$, $c \in \mathbb{R}$ to be a constant valued function, and the covariance function $K(\cdot)$ to be the Matérn class of stationary covariance function (Matérn, 1986).

## Appendix C: Storm Severity Index

The SSI was originally developed to estimate windstorm-related damage to buildings and infrastructure (Klawa and Ulbrich, 2003). With this aim, daily wind gust maxima ($v_{g,max}$) for DWD stations were first scaled with its $98^{th}$ percentile ($v_{g,98,s}$) to take local conditions into account. Next, the exceedences above $v_{g,98,s}$ are cubed to account for the wind destructiveness and are weighted with the population density as a proxy for the insured values. Later developments introduced formulations for grid-based data, i.e., reanalysis data and climate models, definition of affected areas ("windstorm footprints") and feature tracking (e.g., Pinto et al., 2007; Leckebusch et al., 2008). Following Pantillon et al. (2018), we use the SSI formulation for station observations over Germany considering only the meteorological impact (no population weighting):

$$\mathrm{SSI} = \sum_{\text{station s}} \left\{ \left( \frac{v_{g,max,s}}{v_{g,98,s}} - 1 \right)^3 \right\}_{v_{g,max} > v_{g,98}} \tag{C1}$$

Daily maximum wind gusts over Germany are available from the DWD surface network (Deutscher Wetterdienst). The SSI is calculated for each single day and displayed in Table 2 for the 12 selected cases. If a storm affects the region for several days, the maximum SSI of the daily values is selected. As the aggregated SSI value is dependent on the number of stations, possible changes in the surface network need to be taken into account for when analysing longer time series. For the period of our case studies (2015–2020), the number of stations remained stable, and thus a comparison of the SSI values is largely fair. Still, the obtained SSI values only serve as a comparison for the selected case studies in this paper and should not be compared numerically to SSI values from other works.

**Figure D1.** Screenshot of the interactive visualisation and labelling tool as part of RAMEFI (Eisenstein et al., 2022a) using the Python package `bokeh` (Bokeh Development Team, 2021). The user can switch between timesteps and loaded winter storms. Top row shows a map of stations and their parameters, here $\tilde{v}$, where the user can select an area by mouse using a lasso or clicking on single stations. Labels for the introduced features can be set for the selected stations. A table includes all data points and parameters. Histograms are shown for several parameters for the whole region (light blue) and currently selected stations (dark blue).
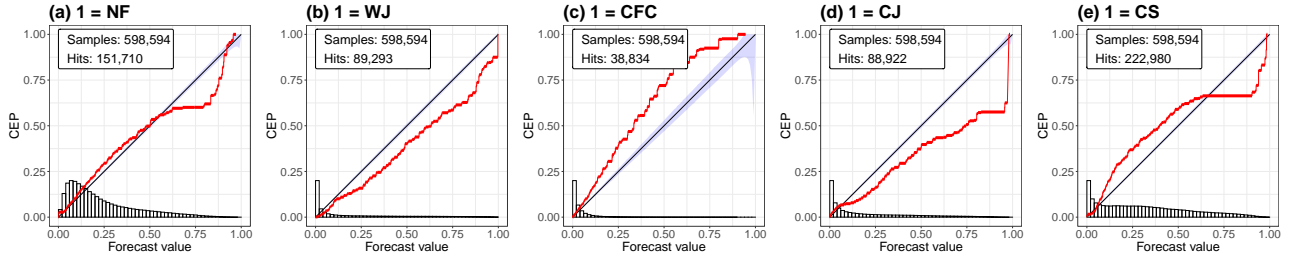
**Figure D2.** As Fig. 6 but based on COSMO-REA6 data.
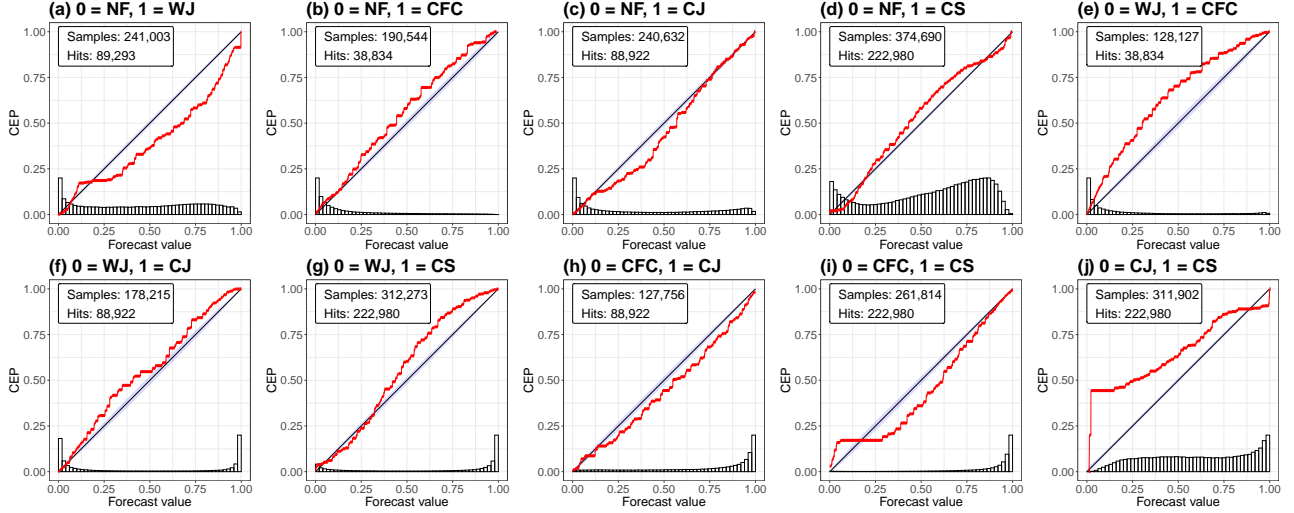
**Figure D3.** As Fig. 7 but based on COSMO-REA6 data.
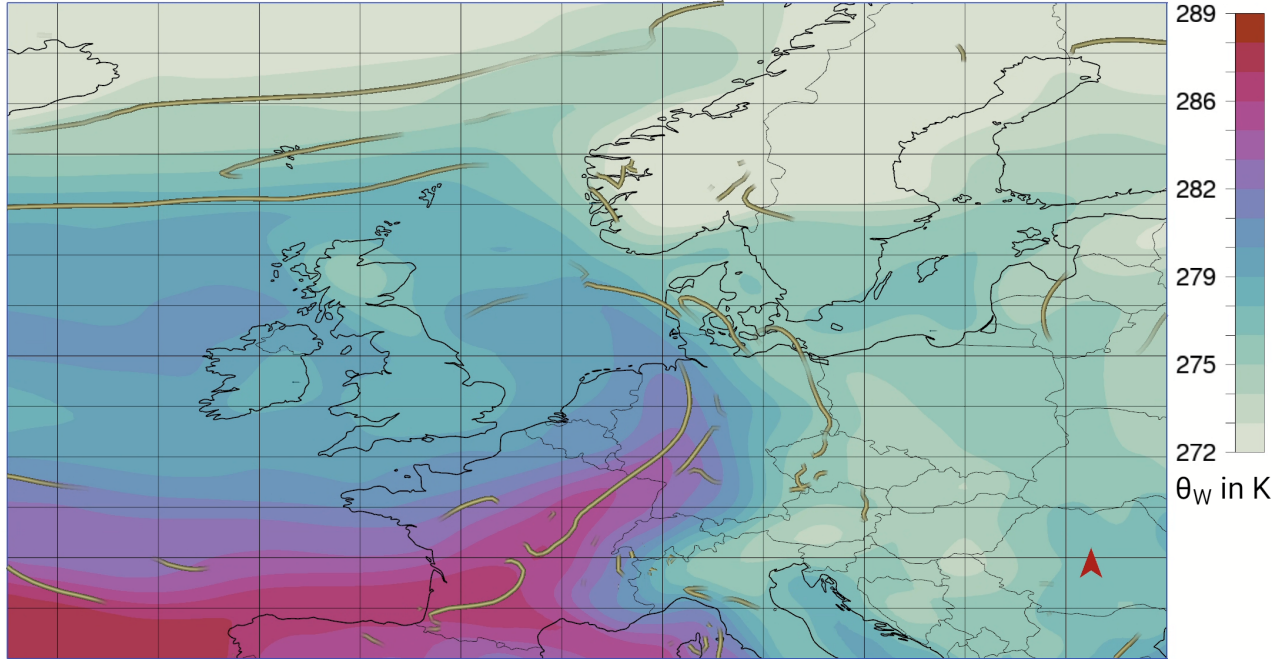
**Figure D4.** Detected fronts at $850\,\mathrm{hPa}$ (green tubes) and wet-bulb potential temperature $\theta_\mathrm{W}$ (shading) for storm *Burglind* on 03 January 2018, 06 UTC using the visualisation software Met.3D ~~(Rautenhaus et al., 2015)~~(Beckert et al., 2022; Rautenhaus et al., 2015). Data source: ERA5.
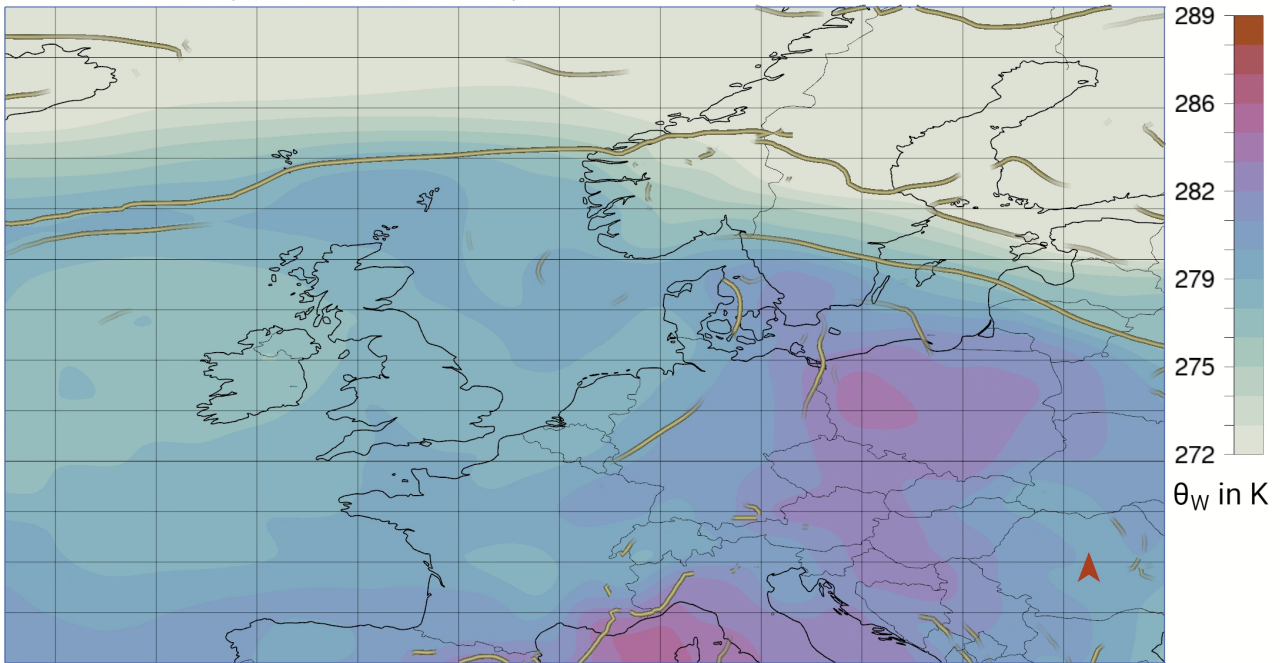
**Figure D5.** As Fig. D4 but for storm *Bennet* on 04 March 2019, 12 UTC.
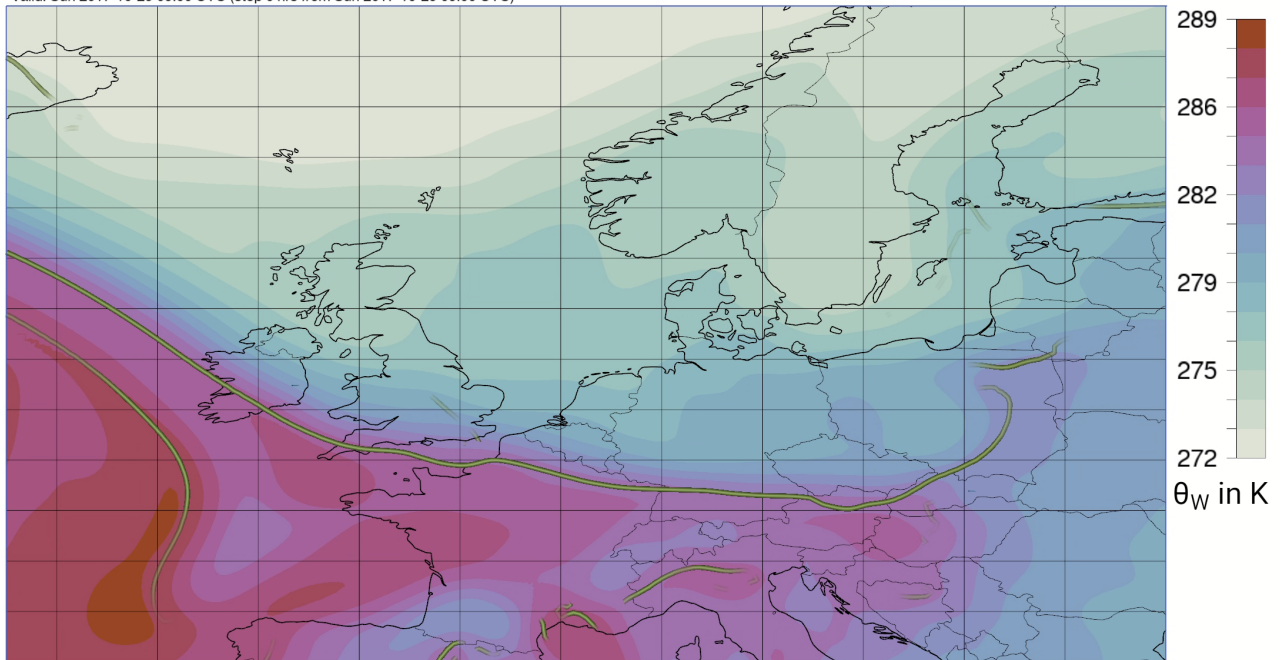
**Figure D6.** As Fig. D4 but for storm *Herwart* on 29 October 2017, 09 UTC,

*Author contributions.* LE developed the interactive tool, labelled the features, drafted the meteorological part of the study and wrote the original manuscript. BS applied the random forest, evaluated the results and drafted the mathematical part of the study. GAQ implemented the Kriging algorithm. PK designed the overall project, acquired the funding and coordinates the scientific work. PK and JGP jointly supervise the PhD of LE. All authors contributed with discussions and text revisions.

840    *Competing interests.* PK is member of the editorial board of *Weather and Climate Dynamics*. The peer-review process was guided by an independent editor. The authors have no other competing interests to declare.

*Acknowledgements.* The research leading to these results has been accomplished within the project C5 ”Dynamical feature-based ensemble postprocessing of wind gusts within European winter storms" of the Transregional Collaborative Research Center SFB/TRR 165 ”Waves to Weather” funded by the German Science Foundation (DFG). JGP thanks the AXA Research Fund for support. We thank Sebastian Trepte
845    and Olivier Mestre for providing the surface observations data set and Robert Redl for pre-processing the data. Further thanks go to Sebastian Lerch and Johannes Resin for thorough discussions and helpful comments as well as Andreas Beckert for the front visualisation using Met.3D (https://met3d.wavestoweather.de). LE thanks Suzanne L. Gray for hosting her at the University of Reading for an extended research stay in fall 2021 and many discussions. The authors also acknowledge Ambrogio Volonté and an anonymous reviewer for their valuable comments, which helped to improve this article.

## References

Bach, L., Schraff, C., Keller, J. D., and Hense, A.: Towards a probabilistic regional reanalysis system for Europe: evaluation of precipitation from experiments, Tellus A: Dynamic Meteorology and Oceanography, 68, 32 209, https://doi.org/10.3402/tellusa.v68.32209, 2016.

Beckert, A. A., Eisenstein, L., Oertel, A., Hewson, T., Craig, G. C., and Rautenhaus, M.: The three-dimensional structure of fronts in mid-latitude weather systems as represented by numerical weather prediction models, Weather and Climate Dynamics Discussions [preprint], 2022, 1–42, https://doi.org/10.5194/wcd-2022-36, in review, 2022.

Bivand, R. and Lewin-Koh, N.: maptools: Tools for Handling Spatial Objects, https://CRAN.R-project.org/package=maptools, r package version 1.1-2, 2021.

Bjerknes, J.: On the Structure of Moving Cyclones, Monthly Weather Review, 47, 95–99, https://doi.org/10.1175/1520-0493(1919)47<95:otsomc>2.0.co;2, 1919.

Bokeh Development Team: Bokeh: Python library for interactive visualization, https://bokeh.pydata.org/en/latest/, python package version 2.3.2, 2021.

Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S., and Steinke, S.: Towards a high-resolution regional reanalysis for the European CORDEX domain, Quarterly Journal of the Royal Meteorological Society, 141, 1–15, https://doi.org/https://doi.org/10.1002/qj.2486, 2015.

Breiman, L.: Classification and regression trees, Wadsworth Internat. Group, 1984.

Breiman, L.: Random forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Monthly Weather Review, 78, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2, 1950.

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

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

Catto, J. L. and Pfahl, S.: The importance of fronts for extreme precipitation, Journal of Geophysical Research: Atmospheres, 118, 10,791–10,801, https://doi.org/https://doi.org/10.1002/jgrd.50852, 2013.

Catto, J. L. and Raveh-Rubin, S.: Climatology and dynamics of the link between dry intrusions and cold fronts during winter. Part I: global climatology, Climate Dynamics, 53, 1873–1892, https://doi.org/10.1007/s00382-019-04745-w, 2019.

Clark, M. R.: A provisional climatology of cool-season convective lines in the UK, Atmospheric Research, 123, 180–196, https://doi.org/10.1016/j.atmosres.2012.09.018, 2013.

Clark, P. A. and Gray, S. L.: Sting jets in extratropical cyclones: a review, Quarterly Journal of the Royal Meteorological Society, 144, 943–969, https://doi.org/10.1002/qj.3267, 2018.

Deutscher Wetterdienst: Climate Data Center, https://cdc.dwd.de/portal/, [Accessed 27 April 2022].

Dimitriadis, T., Gneiting, T., and Jordan, A.: Stable reliability diagrams for probabilistic classifiers, Proceedings of the National Academy of Sciences, 118, e2016191 118, https://doi.org/10.1073/pnas.2016191118, 2021.

Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain: fields: Tools for spatial data, https://doi.org/10.5065/D6W957CT, r package version 12.5, 2017.

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

Eckhardt, S., Stohl, A., Wernli, H., James, P., Forster, C., and Spichtinger, N.: A 15-year climatology of warm conveyor belts, Journal of Climate, 17, 218–237, https://doi.org/10.1175/1520-0442(2004)017<0218:AYCOWC>2.0.CO;2, 2004.

890 Eisenstein, L., Pantillon, F., and Knippertz, P.: Dynamics of sting-jet storm Egon over continental Europe: Impact of surface properties and model resolution, Quarterly Journal of the Royal Meteorological Society, 146, 186–210, https://doi.org/10.1002/qj.3666, 2020.

Eisenstein, L., Schulz, B., Qadir, G. A., Pinto, J. G., and Knippertz, P.: RAMEFI (RAndom-forest based MEsoscale wind Feature Identification), https://doi.org/10.5281/zenodo.6541303, 2022a.

Eisenstein, L., Schulz, B., Qadir, G. A., Pinto, J. G., and Knippertz, P.: Objective identification of high-wind features within
895 extratropical cyclones using a probabilistic random forest. Part I: Method and illustrative case studies - Video Supplement, https://doi.org/10.5281/zenodo.6541277, 2022b.

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

900 Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T.: mvtnorm: Multivariate Normal and t Distributions, https://CRAN.R-project.org/package=mvtnorm, r package version 1.1-3, 2021.

Gneiting, T. and Ranjan, R.: Combining predictive distributions, Electronic Journal of Statistics, 7, 1747–1782, https://doi.org/10.1214/13-EJS823, 2013.

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

Gray, S. L., Martínez-Alvarado, O., Ackerley, D., and Suri, D.: Development of a prototype real-time sting-jet precursor tool for forecasters, Weather, 76, 369–373, https://doi.org/https://doi.org/10.1002/wea.3889, 2021.

Greenwell, B. M.: pdp: An R package for constructing partial dependence plots, R Journal, 9, 421–436, https://doi.org/10.32614/rj-2017-016, 2017.

910 Guttorp, P. and Gneiting, T.: Studies in the history of probability and statistics XLIX: On the Matérn correlation family, Biometrika, 93, 989–995, https://doi.org/10.1093/biomet/93.4.989, 2006.

Hans-Ertel-Centre for Weather Research: COSMO Regional Reanalysis, https://reanalysis.meteo.uni-bonn.de, [Accessed 27 April 2022], 2019.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer, https://doi.org/10.1007/978-0-387-84858-7, 2009.

915 Hewson, T. D. and Neu, U.: Cyclones, windstorms and the IMILAST project, Tellus, Series A: Dynamic Meteorology and Oceanography, 67, 27 128, https://doi.org/10.3402/tellusa.v67.27128, 2015.

Hill, A. J., Herman, G. R., and S., S. R.: Forecasting Severe Weather with Random Forests, Monthly Weather Review, 148, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344., 2020.

Hu, G. and Franzke, C. L. E.: Evaluation of Daily Precipitation Extremes in Reanalysis and Gridded Observation-Based Data Sets Over
920 Germany, Geophysical Research Letters, 47, e2020GL089 624, https://doi.org/https://doi.org/10.1029/2020GL089624, e2020GL089624 10.1029/2020GL089624, 2020.

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

Lackmann, G.: Chapter 6: Fronts, in: Midlatitude Synoptic Meteorology: Dynamics, Analysis and Forecasting, pp. 131–166, American Meteorological Society, Boston, Mass., 2011.

Lagerquist, R., McGovern, A., and Smith, T.: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind, Weather and Forecasting, 32, 2175 – 2193, https://doi.org/10.1175/WAF-D-17-0038.1, 2017.

Leckebusch, G. C., Renggli, D., and Ulbrich, U.: Development and application of an objective storm severity measure for the Northeast Atlantic region, Meteorologische Zeitschrift, 17, 575–587, https://doi.org/10.1127/0941-2948/2008/0323, 2008.

Madonna, E., Wernli, H., Joos, H., and Martius, O.: Warm Conveyor Belts in the ERA-Interim Dataset (1979–2010). Part I: Climatology and Potential Vorticity Evolution, Journal of Climate, 27, 3–26, https://doi.org/10.1175/JCLI-D-12-00720.1, 2014.

Manning, C., Kendon, E. J., Fowler, H. J., Roberts, N. M., Berthou, S., Suri, D., and Roberts, M. J.: Extreme windstorms and sting jets in convection-permitting climate simulations over Europe, Climate Dynamics, 58, 2387–2404, https://doi.org/10.1007/s00382-021-06011-4, 2022.

Martínez-Alvarado, O., Baker, L. H., Gray, S. L., Methven, J., and Plant, R. S.: Distinguishing the Cold Conveyor Belt and Sting Jet Airstreams in an Intense Extratropical Cyclone, Monthly Weather Review, 142, 2571–2595, https://doi.org/10.1175/mwr-d-13-00348.1, 2014.

Matérn, B.: Spatial Variation, Berlin:Springer-Verlag, 2nd edn., 1986.

Matheron, G.: Principles of geostatistics, Economic geology, 58, 1246–1266, 1963.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the black box more transparent: Understanding the physical implications of machine learning, Bulletin of the American Meteorological Society, 100, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1, 2019.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I. V., Feser, F., Koszalka, I., Kreibich, H., Pantillon, F., Parolai, S., Pinto, J. G., Punge, H. J., Rivalta, E., Schröter, K., Strehlow, K., Weisse, R., and Wurpts, A.: Impact Forecasting to Support Emergency Management of Natural Hazards, Reviews of Geophysics, 58, e2020RG000 704, https://doi.org/https://doi.org/10.1029/2020RG000704, e2020RG000704 2020RG000704, 2020.

Original S code by Richard A. Becker and Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn.: maps: Draw Geographical Maps, https://CRAN.R-project.org/package=maps, r package version 3.3.0, 2018.

Pantillon, F., Lerch, S., Knippertz, P., and Corsmeier, U.: Forecasting wind gusts in winter storms using a calibrated convection-permitting ensemble, Quarterly Journal of the Royal Meteorological Society, 144, 1864–1881, https://doi.org/10.1002/qj.3380, 2018.

Parton, G., Dore, A., and Vaughan, G.: A climatology of mid-tropospheric mesoscale strong wind events as observed by the MST radar, Aberystwyth, Meteorological Applications, 17, n/a–n/a, https://doi.org/10.1002/met.203, 2010.

Peterson, R. A.: Finding Optimal Normalizing Transformations via bestNormalize, The R Journal, 13, 310–329, https://doi.org/10.32614/RJ-2021-041, 2021.

Pinto, J. G., Fröhlich, E. L., Leckebusch, G. C., and Ulbrich, U.: Changing European storm loss potentials under modified climate conditions according to ensemble simulations of the ECHAM5/MPI-OM1 GCM, Natural Hazards and Earth System Sciences, 7, 165–175, https://doi.org/10.5194/nhess-7-165-2007, 2007.

R Core Team: R: A language and environment for statistical computing, https://www.r-project.org/, 2021.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, The MIT Press, https://doi.org/10.7551/mitpress/3206.001.0001, 2005.

Rautenhaus, M., Grams, C. M., Schäfler, A., and Westermann, R.: Three-dimensional visualization of ensemble weather forecasts – Part 2: Forecasting warm conveyor belt situations for aircraft-based field campaigns, Geoscientific Model Development, 8, 2355–2377, https://doi.org/10.5194/gmd-8-2355-2015, 2015.

Raveh-Rubin, S.: Dry intrusions: Lagrangian climatology and dynamical impact on the planetary boundary layer, Journal of Climate, 30, 6661–6682, https://doi.org/10.1175/JCLI-D-16-0782.1, 2017.

Raveh-Rubin, S. and Catto, J. L.: Climatology and dynamics of the link between dry intrusions and cold fronts during winter, Part II: Front-centred perspective, Climate Dynamics, 53, 1893–1909, https://doi.org/10.1007/s00382-019-04793-2, 2019.

Sanders, F.: On subjective probability forecasting, Journal of Applied Meteorology, 2, 191–201, https://doi.org/10.1175/1520-0450(1963)002<0191:OSPF>2.0.CO;2, 1963.

Sanders, F. and Gyakum, J. R.: Synoptic-Dynamic Climatology of the "Bomb", Monthly Weather Review, 108, 1589–1606, 1980.

Schultz, D. M., Keyser, D., and Bosart, L. F.: The Effect of Large-Scale Flow on Low-Level Frontal Structure and Evolution in Midlatitude Cyclones, Monthly Weather Review, 126, 1767–1791, https://doi.org/10.1175/1520-0493(1998)126<1767:teolsf>2.0.co;2, 1998.

Schulz, B. and Lerch, S.: Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison, Monthly Weather Review, 150, 235 – 257, https://doi.org/10.1175/MWR-D-21-0150.1, 2022.

Shapiro, M. A. and Keyser, D.: Fronts, Jet Streams and the Tropopause, in: Extratropical Cyclones: The Erik Palmen memorial volume, edited by Newton, C. W. and Holopainen, E. O., pp. 167–191, American Meteorological Society, 1990.

Smart, D. J. and Browning, K. A.: Attribution of strong winds to a cold conveyor belt and sting jet, Quarterly Journal of the Royal Meteorological Society, 140, 595–610, https://doi.org/10.1002/qj.2162, 2014.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, Monthly Weather Review, 144, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1, 2016.

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Bouallègue, Z. B., Bhend, J., Dabernig, M., Cruz, L. D., Hieta, L., Mestre, O., Moret, L., Plenković, I. O., Schmeits, M., Taillardat, M., den Bergh, J. V., Schaeybroeck, B. V., Whan, K., and Ylhaisi, J.: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, Bulletin of the American Meteorological Society, 102, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1, 2021.

Volonté, A., Clark, P. A., and Gray, S. L.: The role of mesoscale instabilities in the sting-jet dynamics of windstorm Tini, Quarterly Journal of the Royal Meteorological Society, 144, 877–899, https://doi.org/10.1002/qj.3264, 2018.

Wernli, H. and Davies, H. C.: A Lagrangian-based analysis of extratropical cyclones. I: The method and some applications, Quarterly Journal of the Royal Meteorological Society, 123, 467–489, https://doi.org/10.1256/smsqj.53810, 1997.

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

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Wright, M. N. and Ziegler, A.: Ranger: A fast implementation of random forests for high dimensional data in C++ and R, Journal of Statistical Software, 77, 1–17, https://doi.org/10.18637/jss.v077.i01, 2017.

Zadrozny, B. and Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 694–699, Association for Computing Machinery (ACM), New York, New York, USA, https://doi.org/10.1145/775047.775151, 2002.