https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Predicting Ice Supersaturation for Contrail Avoidance: Ensemble Forecasting using ICON with Two-Moment Ice Microphysics

Maleen Hanst[1], Carmen G. Köhler[1], Axel Seifert[1], and Linda Schlemmer[1]
[1]Deutscher Wetterdienst, Frankfurter Straße 135, 63067 Offenbach am Main, Germany
Correspondence: Maleen Hanst (maleen.hanst@dwd.de)

Abstract. Contrails and contrail-induced cirrus clouds are considered the most significant non-CO_2 contributors to aviation's climate impact and occur primarily in ice-supersaturated regions (ISSRs). Reliable prediction of relative humidity over ice (RH_{ice}) in the upper troposphere and lower stratosphere allows mitigating their formation by re-routing flights. We implemented a two-moment cloud ice microphysics parameterization within a ten-member Ensemble Prediction System (EPS) using the global ICON (ICOsahedral Nonhydrostatic) model. RH_{ice} predictions were evaluated against radiosonde and aircraft observations from the Northern Hemisphere during 2024–2025. Treating ISSR prediction ($RH_{ice} > 100\%$) as a binary classification problem, we find that the probability of detection (POD) of ISSRs increases to 0.6 for the two-moment scheme (ICON 2-Mom), compared to 0.4 for the operational ICON with a one-moment ice microphysics scheme, while maintaining a low false positive rate (FPR < 0.1). Further evaluation of the ICON 2-Mom EPS using Receiver Operating Characteristic (ROC) analysis shows a POD of 0.8 for a decision model that requires at least three ensemble members to predict ISSR, with an FPR of 0.13. Additionally, we incorporate ensemble spread information to develop a meta-model that further reduces the FPR. Since June 2024, more than 100 flights have been rerouted based on ICON 2-Mom EPS predictions in a contrail avoidance trial, demonstrating the practical value of improved ISSR forecasts for climate-conscious aviation. This study highlights the significant potential of ensemble-based modeling for predicting ISSRs and RH_{ice}, supporting environmentally optimized flight planning and advancing applications in weather and climate science.

1 Introduction

The impact of aviation on climate change is a growing concern, especially as the number of aircraft increases (Yamashita et al., 2016; Grewe et al., 2021). Air traffic is estimated to contribute to global warming by approximately 3.5% (Lee et al., 2023), with an uncertainty range of 2% - 14% (Lee, 2018), caused by CO_2 and non-CO_2 effects.

While the uncertainty range for the climate impact of CO_2 emissions is relatively small, there is significant variability associated with non-CO_2 effects arising from emissions such as NO_x, H_2O, and, notably, the formation of persistent contrails and contrail-induced cirrus clouds (Matthes et al., 2017; Klöwer et al., 2021; Lührs et al., 2021; Lee et al., 2023). These phenomena,

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

collectively referred to as aircraft-induced clouds, present a complex challenge for climate assessment. While Kärcher (2018) estimate that their contribution accounts for more than half of aviation's total radiative forcing, Bickel et al. (2020) contend that the net warming effect might be less than that of CO_2, primarily because it may be partially offset by a decrease in natural cirrus cloud coverage. Given the variety of findings and the potential trade-off between CO_2 and non-CO_2 impacts, effective strategies to mitigate the climate impact of aviation must consider both types of effects. Among these strategies, climate-optimized flight routing has gained attention in recent years, as it seeks to minimize aviation-induced warming by accounting for a comprehensive range of atmospheric impacts (Schumann et al., 2011; Grewe et al., 2017a, b; Matthes et al., 2017; Simorgh et al., 2022). This approach is built upon climate response models such as the Contrail Cirrus Prediction (CoCiP) model (Schumann, 2012), its Python adaptation PyContrails (Shapiro et al., 2023), or algorithmic Climate Change Functions (aCCF) (Dietmüller et al., 2022; Matthes et al., 2023), which provide the necessary computational framework.

Climate response models rely on four-dimensional meteorological fields – typically derived from numerical weather prediction (NWP) models – in which relative humidity over ice (RH_{ice}) is a key parameter for evaluating contrail formation according to the Schmidt–Appleman criterion (Schmidt, 1941; Appleman, 1953; Schumann, 1996). To provide climate response models with physically consistent and representative atmospheric inputs, it is therefore crucial that NWP models accurately capture RH_{ice}, especially under conditions of ice supersaturation ($RH_{ice} > 100\%$), which are essential for persistent contrail development. Yet, despite its critical role in contrail prediction, RH_{ice} remains one of the most uncertain variables in NWP models, with ongoing difficulties in capturing its variability, dynamics, and interactions with cloud microphysics.

Errors and uncertainties in Numerical Weather Prediction (NWP) models stem from various factors, including sparse and noisy observational data for initial conditions, as well as inherent limitations in model physics and numerical methods. Among these challenges, accurate prediction of relative humidity over ice (RH_{ice}) remains particularly difficult, even for state-of-the-art models. This is largely due to the limited availability of upper tropospheric humidity observations for data assimilation, a large variability of humidity fields, and the incomplete understanding of ice nucleation and cirrus cloud formation processes. Parameterizations of ice microphysical processes are therefore an active area of research (Kärcher et al., 2022; Seifert et al., 2022; Spichtinger et al., 2023; Achatz et al., 2024; Lüttmer et al., 2024). Additionally, predicting ice supersaturation poses challenges due to resolution limitations: NWP models typically represent mean atmospheric values and may miss highly localized ice supersaturated regions (ISSRs), particularly those associated with unresolved mesoscale gravity waves (Wilhelm et al., 2018).

One way to circumvent these limitations is to build a postprocessing model which receives variables such as temperature, RH_{ice}, and others, and outputs RH_{ice}. Wang et al. (2025) focused their research on reanalysis data, deriving their postprocessing model inputs from ERA5 (ECMWF Reanalysis v5) data, and trained their model using humidity measurements from the In-service Aircraft for a Global Observing System (IAGOS), showing RH_{ice} mean absolute error improvements in test data.

The use of high-resolution NWP models is another approach to dealing with uncertainties in predicting RH_{ice}. In a recent study by Thompson et al. (2024), several leading high-resolution NWP models have been validated with respect to RH_{ice} using radiosonde and IAGOS data, and the results are discussed in the context of contrail avoidance flight routing. RH_{ice} predictions of IFS (Integrated Forecasting System), GFS (Global Forecast System), and S-WRF (Weather Research and Forecasting model

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

configured by SATAVIA) are evaluated and moderate scores in terms of the F_1 score and the Matthews Correlation Coefficient were found. The study highlighted that a correct prediction of conditions which are *not* conductive to contrail formation, mainly the condition of non-ISSR, is also crucial, as false negatives (thus, incorrect ISSR predictions) could potentially lead to unnecessary re-routing. For the S-WRF model, they found a true positive rate for the non-ISSR condition of 90.7 % and a true positive rate for the ISSR condition of 45.9 %. Hence, for ISSR they observe a false positive rate of 9.3% and a false negative rate of 54.1%. The authors point out that the relatively high false negative rate of ISSR indicates missed opportunities for contrail avoidance, which is not ideal, but also has no consequences other than the current status quo of aviation impacts. Conversely, the low false positive rate of ISSR suggests that there may be only few worst-case scenarios where aircraft are diverted to an incorrectly predicted non-ISSR due to an incorrectly predicted ISSR, resulting in both additional CO_2 emissions and possible contrail formation.

In our study, we investigate the ability to predict RH_{ice} by using a two-moment cloud ice microphysics parameterization scheme within a ten-member Ensemble Prediction System (EPS) in the global ICON (ICOsahedral Nonhydrostatic) NWP model (Zängl et al., 2014). ICON is used by the German Meteorological Service (DWD) and is developed through the ICON partnership, which includes the Deutsches Klimarechenzentrum, Max Planck Institute for Meteorology, Karlsruhe Institute of Technology, Center for Climate Systems Modeling, and DWD. In the operational configuration of the ICON model, cloud ice microphysics is represented by a one-moment scheme in which ice mass is considered a prognostic variable. However, this approach cannot capture higher levels of supersaturation. To improve predictions of RH_{ice} in the upper troposphere and lower stratosphere (UTLS), we adopt a two-moment cloud ice scheme that includes ice particle number density as an additional prognostic variable. We have implemented a simplified and slightly adapted version of Köhler and Seifert (2015), which allows a more realistic representation of ice microphysical processes than the operational model, while remaining consistent in the warm phase.

Building on ICON with the two-moment ice microphysics scheme, we set up an ensemble prediction system analogous to the operational global ICON. To balance the benefits of ensemble forecasting with the constraints of computational resources, we selected ten of the 40 ensemble members used in the operational configuration.

Ensemble forecasting offers a powerful framework for assessing both the predictability of atmospheric phenomena and the uncertainties inherent in numerical weather prediction (NWP) models (Epstein, 1969; Lewis, 2005). It is state-of-the-art to describe the initial conditions of an NWP model using probabilistic distributions (Du et al., 2018) and to perform ensemble-based data assimilation, not only to obtain initial conditions for ensemble forecasts, but also for deterministic forecasts (Snyder and Zhang, 2003; Hunt et al., 2007). Furthermore, NWP model imperfections can be addressed by multi-model, multi-physics and stochastic physics approaches (Berner et al., 2011) integrated into the ensemble forecast generation process.

Ensemble forecasts inherently provide access to uncertainty estimates by generating a probability distribution for each grid point. Although taking the ensemble mean is a common method for deriving a more stable deterministic forecast, ISSR prediction may benefit from alternative uses of ensemble information. Since NWP models generally represent mean conditions, extreme RH_{ice} values within the ensemble may indicate potentially extreme ISSR events. Additionally, the spread of RHice values, as captured by the standard deviation, may improve the reliability of ISSR classification.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Nevertheless, the practical application of ensemble forecasts remains challenging. While they provide access to uncertainty and thus represent a more complete and realistic forecasting framework compared to deterministic point forecasts, they still face the typical challenges associated with forecast application. Not only is the process of NWP forecasting inherently imperfect at each step (data collection, data assimilation, model physics and model numerics) but the interpretation of the resulting forecast output, whether deterministic or probabilistic, also leaves room for methodological variability. Different post-processing techniques and user perspectives can lead to significant differences in how forecasts are applied in real-world scenarios (Du et al., 2018), especially in the case of ensemble forecasts.

The main contribution of this study is to carefully analyze and interpret the EPS predictions of RH_{ice} based on ICON with the two-moment ice microphysics scheme (ICON 2-Mom EPS) through verification with radiosonde observations. For this purpose, several ensemble metrics are considered and meta-models based on the EPS are discussed, showing a great potential of ensemble-based forecasts of RH_{ice} compared to deterministic forecasts.

The ICON 2-Mom EPS has been established as a dedicated forecasting system at DWD to provide continuous meteorological data for research on contrail avoidance flights. This setup was developed within the D-KULT project (demonstrator climate and environmentally friendly air transport), which aims to demonstrate the feasibility of climate-optimized flight trajectories with a focus on reducing contrail formation in European airspace. It aims to optimize flight paths using climate response models that account for both CO_2 and non-CO_2 effects, while balancing emissions, noise, operating costs and real-world constraints such as airspace regulations and airport capacity. One of the components is the integration of the ICON 2-Mom EPS forecast to identify potential persistent contrail regions for contrail avoidance flight planning. In real-world trials, more than 100 flights have already been rerouted using information from these forecasts, demonstrating the practical application of climate-aware flight paths.

The outline of this paper is as follows: In Section 2, we describe the details of the dedicated ICON forecasting system, in particular, the two-moment cloud ice microphysics scheme and the ensemble setup. In Section 3, an overview over the in-situ observation measurement data used for verification is given. The verification methodology and results are presented and analyzed in Section 4, where we start by evaluating the deterministic model with the new two-moment ice microphysics scheme and then move on to analyze ensemble metrics of the EPS setup. This is followed by a discussion of the results in Section 5 and a conclusion in Section 6.

2 Model

In the following, the dedicated ICON forecasting system is described, which has been specifically established for climate-optimized flying and differs from the operational setup of the ICON model at DWD. The two main changes are the use of a two-moment cloud ice microphysics scheme and the reduction of ensemble members, both of which are described in detail below.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

2.1 Two-moment Cloud Ice Microphysics parameterization in ICON

Accurate prediction of the potential for persistent contrail formation requires a realistic representation of relative humidity over ice in NWP models. A key factor in this is the model's ability to simulate the phase relaxation time – the timescale over which water vapour transitions to ice. In the operational configuration of ICON, a one-moment cloud ice microphysics scheme is used, in which the specific ice mass is treated as a prognostic variable and the ice particle number density is estimated from temperature. This approach tends to overestimate particle numbers at low temperatures, resulting in unrealistically short relaxation times and limiting the ability of the model to represent ice supersaturated conditions.

To address these limitations, a two-moment ice microphysics scheme has been implemented in ICON, in which the ice particle number density is treated as an additional prognostic variable. The implementation is based on previous versions, as described in Köhler and Seifert (2015), and has been adapted to maintain consistency with the operational one-moment scheme of the ICON model, using a single ice mode. In the updated version, heterogeneous ice nucleation is parameterized based on laboratory measurements from the Karlsruhe Institute of Technology (KIT) (Ullrich et al., 2017), while homogeneous nucleation follows the approach of (Kärcher, 2018). This two-moment scheme provides a more physical and realistic representation of ice microphysics, especially under conditions relevant to contrail formation. More details on the scheme and its implementation can be found in the Appendix A. The difference in model behavior is illustrated in Fig. 1(a), which shows the global structures of relative humidity over ice with the operational ICON (top) and the ICON with the two-moment cloud ice microphysics scheme (bottom). While the cloud structures remain similar, the degree of ice supersaturation increases significantly with the new two-moment cloud ice scheme. The extent to which this is realistic is elaborated in Section 4 in comparison with observational data.

2.2 Ensemble Prediction System

For the D-KULT project, a dedicated ten-member global ensemble prediction system based on the operational ICON model (Reinert et al., 2025) has been established. Although the full operational system consists of 40 ensemble members, the reduced configuration of 10 members was found to be sufficient to meet the project requirements for forecasting key variables such as relative humidity over ice. The ensemble generation is based on the Local Ensemble Transform Kalman Filter (LETKF) method (Ott et al., 2004; Hunt et al., 2007), which perturbs the initial conditions of all members simultaneously in a member-dimensional space. The initial state of each ensemble member is computed by combining its background state – a short-range forecast – with a weighted correction derived from the differences between observations and model background. These weights are computed via a gain matrix that incorporates both observation error and background error covariances, ensuring that each member assimilates observation information in a distinct but dynamically consistent way.

In addition to initial condition perturbations, the system includes stochastic perturbations of selected physical parameterizations. For the global ensemble system, these physical parameters are randomly perturbed for each ensemble member at the start of the forecast and remain fixed throughout the forecast integration. This approach introduces variability among ensemble members while preserving the consistency of individual forecast trajectories. The combined perturbation strategy ensures a

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

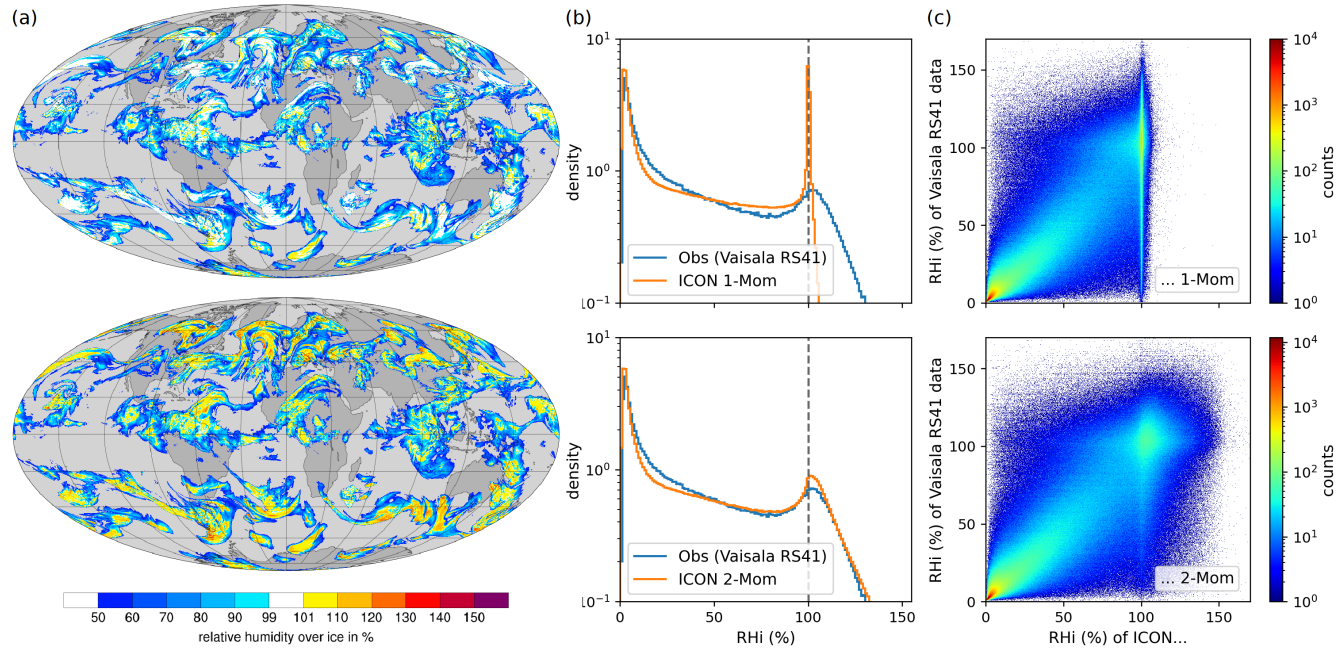

Figure 1. Relative humidity over ice (RH_{ice}) of the operational ICON with one-moment ice microphysics scheme (top row) and of ICON with two-moment ice microphysics scheme as implemented in the dedicated system (bottom row): (a) Global forecast-only data of RH_{ice} near the tropopause; (b) normalized histograms of RH_{ice} of Vaisala RS41 radiosonde data and ICON; (c) scatter plot of RH_{ice} of Vaisala RS41 radiosonde data versus ICON forecasts with a lead time of 12 hours; pressure between 150-300 hPa, corresponding to most common commercial flight altitudes.

realistic representation of forecast uncertainty, which is crucial for assessing the sensitivity of contrail formation potential to meteorological variability.

2.3 Model Setup

The dedicated system for the D-KULT project is based on ICON version 2.6.6. Detailed information on the adapted code can be found in the Appendix A. The system runs on the ICON R3B06 grid, which has a horizontal spacing of about 26 km and a vertical spacing of about 200 m at the most common commercial flight altitudes of 8.5-12.5 geopotential kilometers. It starts from the operational analysis, which is based on the one-moment ice microphysics scheme, so that we require a spin-up time of at least 6 hours in our evaluations below to build up ice supersaturation. The model is run four times a day, initialized at 00, 06, 12, and 18 UTC with a forecast lead time of 60 hours, producing hourly forecasts. The system consists of ten ensemble members, whose generation is based on the first ten members of the operational ensemble prediction system. This is a reasonable approach as discussed in the Appendix B.

The model outlined forms the basis for the evaluations performed in this study and will be referred to as *ICON 2-Mom EPS* in the remainder of this study. Since the dedicated system does not consist of an additional deterministic model run, we

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

use individual members of the ensemble as approximates to a deterministic model setup for our evaluation, denoted by *ICON 2-Mom*.

3 Observation Data

This study emphasizes in situ measurements for verification, with the primary analysis based on radiosonde data. Additionally, data from the In-Service Aircraft for a Global Observing System (IAGOS; see https://www.iagos.org/) were considered.

3.1 Vaisala RS41 Radiosonde Data

We restricted our radiosonde verification to Vaisala Radiosonde RS41 data, as this type of radiosonde is best scored for humidity measurements in the UTLS (Dirksen et al., 2022; Borg et al., 2023). The temperature sensor utilizes stable, linear resistive platinum technology that yields a measurement accuracy of ± 0.2°C. The humidity sensor integrates humidity and temperature sensing elements and is based on capacitive polymer technology with an accuracy of $\pm 3\%$ RH. Height, pressure, and wind speed and direction data are derived from GPS measurements. For more details on techniques and precision compare Vaisala (2013). We limited our verification to the Northern Hemisphere, where 105 radiosonde stations frequently yield Vaisala RS41 data. In Figure 2(a), the radiosonde locations are shown. Most of them produce daily data from two ascents (around 0 UTC and 12 UTC), which is stored in so-called TEMP BUFR files: TEMP reports include a standardized set of meteorological data, such as temperature, air pressure, wind speed and direction, and humidity at various atmospheric levels. The files are in the BUFR (Binary Universal Form for the Representation of meteorological data) format which is a standardized binary format used by the World Meteorological Organization (WMO) to encode and transmit various types of weather observations, including radiosonde data.

The stored Vaisala RS41 radiosonde height resolution is approximately 1 gpm, with a measurement accuracy of ± 10 gpm. In the TEMP files, the dew point temperature is stored from which we derive RH_{ice} as described in Appendix C. In Fig. 2(b), example radiosonde height profiles of temperature and RH_{ice} are shown together with the corresponding ICON 2-Mom EPS data.

3.2 IAGOS Near-Real-Time Data

In addition to radiosonde data, we use in-situ aircraft data for our verification. The In-service Aircraft for a Global Observing System (IAGOS) is a European research infrastructure that uses commercial aircraft to collect atmospheric data. IAGOS-CORE contains several measurement instruments, e.g., for ozone, carbone monoxide, humidity, and cloud particles, and optionally for nitrogen oxides, greenhouse gases, and more (https://iagos.aeris-data.fr/instrumentation/). Again, the humidity measurement technology used here combines humidity and temperature sensing elements. In more detail, it consists of a capacitive relative humidity sensor (Humicap-H, Vaisala, Finland) and a platinum resistance sensor (PT100) for the measurement of the temperature at the humidity sensing surface. The time resolution of the temperature measurements is 4 s with an accuracy of ± 0.5 K, while the time resolution of the humidity measurements ranges from 1s at 300 K to 120 s at 200 K, with an accu-

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

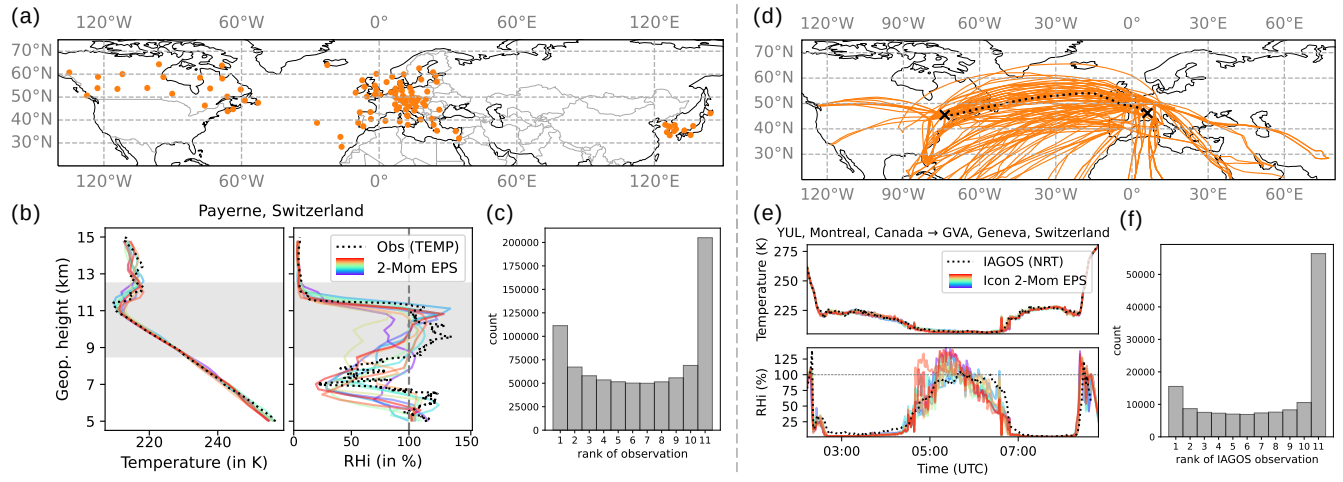

Figure 2. Radiosonde (left) and IAGOS (right) observation data. (a) Locations of 105 stations equipped with Vaisala RS41 radiosondes in the Northern Hemisphere. (b) Example height profiles of temperature and RH_{ice} from Vaisala RS41 (TEMP) observations and ICON 2-Mom EPS forecasts with a lead time of 12 h. (c) Rank histogram: For each spatio-temporal point (comprising ICON 2-Mom EPS values and the corresponding radiosonde measurement), the observed value is ranked among the ten ensemble members, and the resulting ranks are displayed in a histogram. (d) IAGOS flight routes of 188 flights from December 2024, limited to the Northern Hemisphere. (e) Spatio-temporal comparison of flight data and ICON 2-Mom EPS: Time series of temperature and RH_{ice} from one example flight. (f) Rank histogram for IAGOS flight data, analogous to (c).

racy of $\pm 6\%$ (for more details, see www.iagos.org/iagos-core-instruments/h2o/). There are several levels of data processing, from which we have used near-real-time (NRT) data, where humidity measurements are subject to automated quality control, usually within 72 hours (https://iagos.aeris-data.fr/levels/). Only data with validity flag "good" were used (https://iagos.aeris-data.fr/data-quality/) for 625 flights between August 2024 and January 2025. Fig. 2(d) shows the flight routes for December 2024. For a highlighted example flight route, the temperature and RH_{ice} time series are shown together with the corresponding ICON 2-Mom EPS data (Fig. 2(e)). Similar to the radiosonde verification, the analysis is confined to the Northern Hemisphere.

4 Verification Analysis

We evaluated the dedicated ICON system in two steps to successively unravel the improvements in predicting RH_{ice} resulting from the adapted two-moment ice microphysics scheme (ICON 2-Mom) and the ensemble setup (ICON 2-Mom EPS). The methodology used to verify the deterministic model also serves as the basis for the verification of the ensemble prediction system.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

4.1 Verification of Deterministic Model

In the following subsections, we start with the verification of the deterministic model ICON 2-Mom and, in particular, compare it to the operational ICON with the one-moment cloud ice microphysics scheme (denoted by *ICON 1-Mom*). When validating an NWP model with observational data, climatological comparisons on the one hand and spatio-temporal comparisons (e.g., with metrics such as the RMSE) on the other can span the evaluation horizon. We start with a brief look at both, before moving on to consider categorical scores.

4.1.1 Relative Frequency Distribution of RH_{ice}

To enable a climatological comparison between model and observations, we analyze normalized histograms of RH_{ice} within the 8.5–12.5 km geopotential height range. Figure 1(b) displays the observed RH_{ice} relative frequency distributions (densities), shown alongside the corresponding model-based distributions from the operational ICON 1-Mom (top) and ICON 2-Mom (bottom) configurations.

Pronounced differences emerge in the density tail, which reflects ice supersaturation. The operational system exhibits a sharp peak near 100%, followed by a rapid decline, with maximum RH_{ice} values reaching only $\approx$103%. In contrast, the two-moment scheme more accurately captures the tail structure, slightly overshooting at low supersaturation but successfully reproducing the upper range, including RH_{ice} values up to 135%. A few higher values were excluded from the plot due to axis truncation, ensuring comparability without distortion from rare outliers.

4.1.2 Spatio-temporal Comparison

The ICON grid employed features a horizontal resolution of approximately 26 km and a vertical resolution of 200 m within the 8500–12 500 gpm altitude range. Radiosonde data from a given station are mostly horizontally fixed and provide dense vertical coverage. To generate matched ICON–radiosonde data pairs, the ICON grid cell center closest to each radiosonde station was first identified. Subsequently, radiosonde observations were linearly interpolated to the ICON levels, as the model provides mean values across vertical layers with considerably lower resolution than the radiosonde data. Over the 14-month verification period, this approach yielded approximately 820 000 spatio-temporal matching points from more than 63 000 radiosonde profiles. Figure 2(b) shows example radiosonde profiles of temperature and RH_{ice} from one station, compared with ICON ensemble values from the nearest grid cell center.

IAGOS data represent aircraft-based observations and thus capture horizontal trajectories spanning several hours. Matched ICON–IAGOS data pairs were generated by identifying all ICON grid cell centers that were nearest to at least one point along each flight path. Each selected ICON cell was then paired with its closest flight data point, and the model data were vertically interpolated to match the altitude of that observation. An ICON spin up time of a minimum of 6 hours was required. Over the four-month verification period, this procedure yielded approximately 200 000 spatio-temporal matching points from 625 flights. Figure 2(e) shows an example time series of temperature and RH_{ice} from an intercontinental flight, together with the corresponding ICON ensemble values from the nearest model grid cell.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

In the main part of this study, we limit our verification to radiosonde data; the use of IAGOS data is explicitly indicated whenever applicable. As an initial step toward evaluating the spatio-temporal matching points, we examined a simple scatter plot. While ICON 2-Mom reproduces the supersaturation range comparably to the observations, notable scatter remains around the one-to-one line (see Fig. 1(c), showing ICON 1-Mom, top, and ICON 2-Mom, bottom). However, there is no absolute need for perfect agreement between modelled and observed RH_{ice} values. In our context, it is sufficient for the model to realistically capture the occurrence and extent of ice supersaturation. Crucially, the model should be able to distinguish between ISSR events and non-events, both of which have important operational implications for flight planning and routing. To evaluate this capability, we focus on categorical scores below such as the probability of detection and the false positive rate for conditions in which RH_{ice} exceeds selected thresholds.

4.1.3 Categorical Scores of ICON 2-Mom

Instead of analyzing the full continuous range of RH_{ice}, the values can be partitioned based on a specified threshold. This results in a binary classification, distinguishing between two events:

$$RH_{ice} \leq \text{threshold} \quad \text{or} \quad RH_{ice} > \text{threshold}.$$

In this study, we are particularly interested in ice supersaturation ($RH_{ice} > 100\%$) and events of higher ice supersaturation ($RH_{ice} \gg 100\%$). The spatio-temporal matching points between model output and observational data are categorized in a confusion matrix, which serves as the foundation for computing categorical scores (see Table 1). In the remainder of this study, we consider events of the type

$$\{RH_{ice} > \text{threshold}\}_{\text{threshold} \in \{100\%,105\%,110\%,120\%\}}.$$

RH_{ice} > threshold	positive prediction	negative prediction
positive observation	true positive (TP)	false negative (FN)
negative observation	false positive (FP)	true negative (TN)

Table 1. Confusion matrix: Categorization of predicted events (positive predictions) and predicted non-events (negative predictions) in relation to the actual observed situation.

As a starting point for evaluating categorical performance, we consider the Frequency Bias Index (FBI), defined as the ratio of the number of predicted events to the number of observed events:

$$\text{FBI} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{FN}}.$$

The results are shown in Figure 3(a). For the ISSR event (blue curves), the FBI is slightly above 1 for ICON 2-Mom, indicating a modest overprediction, whereas ICON 1-Mom exhibits lower values around 0.75, reflecting underprediction. In both configurations, the FBI remains relatively constant across the examined altitude range of 8.5–12.5 km geopotential height. At higher

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

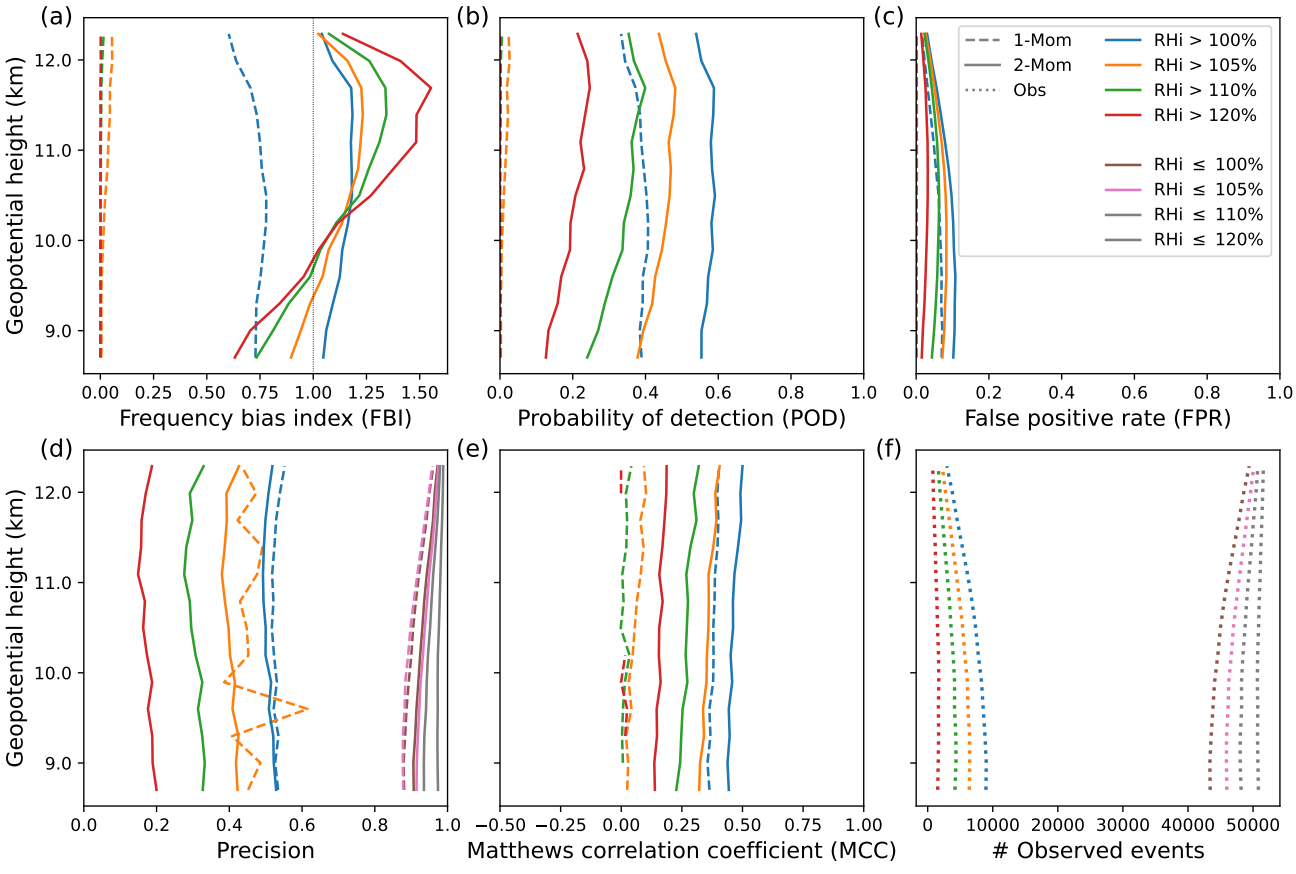

Figure 3. Categorical scores of operational ICON (1-Mom) and ICON 2-Mom versus Vaisala Radiosonde RS41 measurement data. All data from the Northern Hemisphere and in most frequent flight altitudes of 8.5 - 12.5 km geopotential height; verification period of 11.5 months: June 15th, 2024 - May 31th, 2025; ICON initial times 12 UTC and 00 UTC; forecast lead time 12 h; linear interpolation of observation measurements with respect to ICON (on average, there are 13 model levels in the considered altitudes); yielding $\sim$ 680 000 samples in total, with ice supersaturation in $\sim$ 13 % of cases. Scores of events of ice supersaturation: (a) Frequency bias index: ratio of model events to actual observed events; (b) Probability of detection: proportion of actual observed events that are correctly identified by the model; (c) False positive rate: proportion of actual observed non-events that are incorrectly classified by the model as positives; (d) Precision: proportion of positive predictions that are correct. (e) Matthews correlation coefficient: considering all four entries of the confusion matrix (TP, FP, FN, TN) together (missing values are due to vanishing denominators); (f) Vaisala RS41 radiosonde observation measurements.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

RH_{ice} thresholds, the FBI for ICON 2-Mom is slightly below 1 for lower heights but rises to a maximum of approximately 1.5 near 12 km for the event $RH_{ice} > 120\%$. In contrast, ICON 1-Mom yields an FBI of zero across the entire height range, indicating a failure to detect high supersaturation events. These results demonstrate that the two-moment scheme not only predicts ice supersaturation more frequently than the one-moment scheme but also tends to slightly overestimate observed event frequency. Meanwhile, the one-moment scheme consistently underestimates event occurrence.

Moving on to the Probability of Detection (POD), which is the proportion of observed events correctly identified by the model:

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

we find that, for ISSR events, the POD increases from about 0.4 for ICON 1-Mom to about 0.6 for ICON 2-Mom and is almost constant over the altitude range in both cases. For events defined by higher RH_{ice} thresholds, the two-moment scheme retains some detection capability, with the POD gradually decreasing to about 0.15–0.2 for $RH_{ice} > 120\%$. In contrast, as also indicated by the FBI, the one-moment scheme fails to detect RH_{ice} values above 105%, yielding POD values near zero across the altitude range.

To complement the probability of detection, also known as sensitivity, we additionally consider the False Positive Rate (FPR = 1 - specificity), which quantifies the proportion of actually observed non-events that are incorrectly classified by the model as positive events:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}.$$

The false positive rate is relatively low in all cases, reaching a maximum of slightly above 0.1 for the two-moment scheme and $RH_{ice} > 100\%$ (Fig. 3(c)).

POD and FPR are both computed relative to the ground truth: the former with respect to the number of observed events, and the latter with respect to the number of observed non-events. It may also be informative to examine the proportion of predicted events that are actually correct, quantified by the precision:

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

For the ISSR event, both schemes yield similar precision values between 0.5 and 0.55 across the entire altitude range (see Fig. 3(d)). For events with higher RH_{ice} thresholds, the precision of ICON 2-Mom decreases successively, reaching values as low as 0.2 for $RH_{ice} > 120\%$. In contrast, ICON 1-Mom yields very few or even no positive predictions in these regimes, making precision largely undefined; accordingly, it is omitted for these cases.

In the context of flight planning, accurate prediction of non-ISSR conditions is also critical, as false negatives in this category can lead to unnecessary re-routing and, consequently, avoidable increases in CO_2 emissions. When considering the complementary events ($RH_{ice} \leq$ threshold) as "positive" events, the model exhibits high precision, with average values exceeding 0.9 across all threshold levels. Combined with the low false positive rate observed for $RH_{ice} >$ threshold events, this high precision further supports the conclusion that ICON 2-Mom is quite reliable in detecting non-ISSR conditions.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Another way to account for the class imbalance in our dataset (13% ISSR events) and the practical relevance of both event categories is to employ the Matthews Correlation Coefficient (MCC) as a balanced performance metric. Unlike single-aspect measures, the MCC incorporates all four elements of the confusion matrix simultaneously into a single scalar value, making it particularly suitable for evaluating classification performance under skewed data distributions:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}. \quad (305)$$

The MCC ranges from -1 to +1, where +1 indicates perfect discrimination between events and non-events, 0 reflects random predictive skill, and -1 represents complete misclassification. The results of our analysis are shown in Figure 3(e). For ISSR/non-ISSR classification (blue curves), ICON 2-Mom achieves an average MCC of 0.47 across all altitudes, while ICON 1-Mom yields lower values between 0.38 and 0.39. At higher $\mathrm{RH_{ice}}$ thresholds, the MCC for ICON 2-Mom decreases successively, reaching a minimum of approximately 0.16. In contrast, the MCC for ICON 1-Mom approaches zero (or is undefined where the numerator vanishes), indicating no predictive skill. (310)

In summary, for ISSR events, ICON 2-Mom achieves a moderate MCC of nearly 0.5 and a 50% higher POD compared to the operational ICON 1-Mom, while maintaining a relatively low false positive rate below 0.1 at most altitudes. Nevertheless, a POD of 0.6 suggests that further improvements are possible, and we continue to explore potential gains from the ensemble setup introduced in Section 2.2. (315)

4.2 Verification of Ensemble Prediction System ICON 2-Mom EPS

Before examining categorical verification metrics for our ensemble configuration (ICON 2-Mom EPS), we begin this section with a general assessment of the full (continuous) ensemble output and the ensemble spread as a measure of uncertainty. The first question we address is whether the ensemble spread adequately captures the variability observed in the data. Although the ensemble captures some of the variability present in the observations, it remains underdispersive, as indicated by the U-shaped rank histogram (Fig. 2(c)). This underdispersion can, in part, be attributed to the inherent spatial averaging over model grid cells, which tends to smooth out extremes. From a physical modeling perspective, key contributing factors may include the absence of subgrid-scale gravity waves in the model configuration and the use of prescribed aerosol fields from climatology, both of which limit variability in ice nucleation conditions. We also observe a more pronounced negative bias within the rank histogram, indicating that the model tends to underestimate $\mathrm{RH_{ice}}$ more often than it overestimates $\mathrm{RH_{ice}}$. Thus, further post-processing of the EPS model forecasts may be useful for predicting $\mathrm{RH_{ice}}$ and, in particular, for identifying ISSR or higher ice supersaturation. (320, 325)

4.2.1 Prediction of ISSR and Higher Ice Supersaturation

Again, in the context of flight routing, the most important property of the EPS is its ability to distinguish between ISSR and non-ISSR conditions (or higher supersaturation), as both have significant practical implications. Therefore, we again consider binary events such as ice supersaturation ($\mathrm{RH_{ice}} > 100\%$) and higher supersaturation ($\mathrm{RH_{ice}} \gg 100\%$): The ensemble inherently provides probabilistic forecasts for these events via the proportion of members with the corresponding event. We start by (330)

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

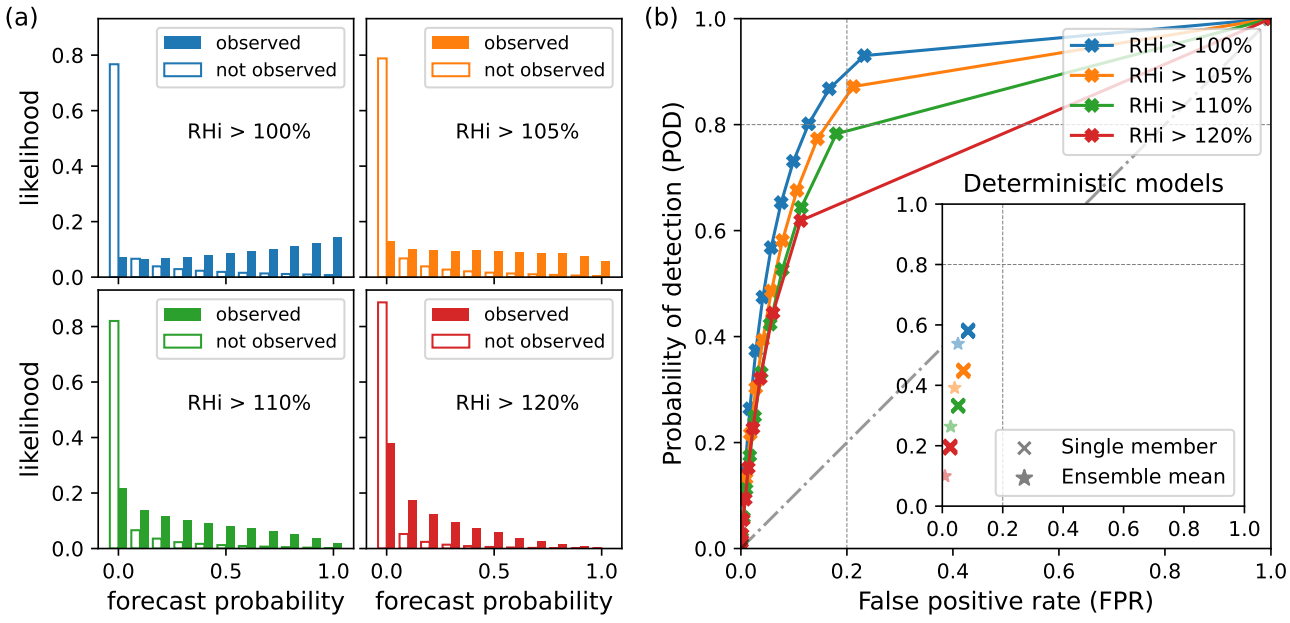

Figure 4. Ensemble metrics targeting the ability of the model to discriminate between events and non-events (e.g., ISSR and non-ISSR in blue); verification period 14 months: April 2024 - May 2025, leading to ∼ 820 000 samples. (a) Discrimination diagram: Conditional distributions of EPS forecast probabilities; conditioned on that the event was actually observed in the measurement data and conditioned on that it was not observed. (b) Receiver Operating Characteristics (ROC) curve: Probability of detection versus false positive rate of ice supersaturation events for varying "decision" models (pseudo-deterministic models received from the EPS by applying various probability threshold conversions).

considering two conditional distributions of the supersaturation forecast probabilities; the first conditional on the event actually being observed in the measurement data, and the second conditional on the event not being observed in the measurement data. In both cases, the corresponding relative frequencies of the EPS forecast probabilities are plotted in a histogram, the discrimination diagram (Fig. 4(a)). Little overlap between the two conditional distributions indicates good discriminability. More specifically, the "not observed" distribution has a dominant peak at zero, indicating that the ensemble members tend to agree when no ISSR or higher ice supersaturation is present. For increasing forecast probabilities, the "not observed" distribution decreases rapidly and is of the same order of magnitude as the "observed" distribution for values of 0.1 and 0.2, before dropping almost to zero for higher forecast probabilities. In contrast, the "observed" distribution is much more uniform, increasing only slightly from low to high prediction probabilities in the case of ISSR (blue). For higher RH_{ice} thresholds, its shape changes from a more uniform to a more pronounced left-sloping distribution, gradually overlapping more and more with the "not observed" distribution. This shows that the ability of the model to discriminate between events and non-events decreases significantly for events with higher ice supersaturation.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Focusing on the ISSR event, the corresponding discrimination diagram shows that the overlap between the two conditional distributions becomes small for forecast probabilities above approximately 0.3. This observation motivates the next step: identifying an appropriate threshold to convert forecast probabilities into binary predictions (0 or 1), thereby enabling a "yes"/"no" decision for the presence of ISSR or higher ice supersaturation. Such a threshold-based conversion yields what we term a "pseudo-deterministic" model. Throughout this study, we refer to these models as decision models, characterized by their underlying conversion thresholds. Specifically, the k-out-of-10 decision model defines the threshold as $k/10$. That is, if at least k out of the 10 ensemble members predict the event, the model outputs a positive prediction. Formally, for each forecast probability p produced by the original EPS, the deterministic forecast p_{conv} is given by:

$$\text{k-out-of-10 decision model}: p_{\text{conv}} = \begin{cases} 1, & \text{if } p \geq \frac{k}{10}, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

We also refer to this model simply as *decision model k*.

The challenge of finding a "good" decision model can be addressed using the Receiver Operating Characteristic (ROC) curve, which plots the POD versus the FPR of all potential decision models. The construction of the ROC curve for a binary event is as follows: For increasing probability thresholds, here 0.0, 0.1, 0.2, up to 1.0, the EPS forecast probabilities are converted to 0 or 1 depending on whether they are below or above the threshold as defined in (1). For the resulting pseudo-deterministic decision models, the POD and FPR can be calculated with respect to the observed data and plotted on a curve. For the ISSR event this results in the blue curve in Fig. 4(b). In general, the closer a point on the curve is to the left corner, the better, as this indicates high POD versus low FPR. In the case of ISSR, when probability thresholds of 0.2 or 0.3 are applied (resulting in decision model 2 or 3), the POD is greater than 0.8 while the FPR remains less than 0.17. However, depending on the false positive cost (which would result from a potential re-routing despite the non-ISSR condition) and the false negative cost (which would result from ISSR passing), a conversion threshold (aka a decision model) can be chosen to obtain an appropriate trade-off between POD and FPR. In the hypothetical (but unrealistic) case of identical costs, the *Youden Index* could be used to determine the point(s) on the ROC curve with the optimal trade-off between POD and FPR:

$$\text{Youden Index} = \text{POD} - \text{FPR},$$

by maximizing it across all possible conversion thresholds (decision models). The range of possible outcomes is from -1 to +1, where 1 indicates a perfect model performance, 0 corresponds to no better than random chance, and negative values reflect performance worse than random guessing. The results corresponding to the ROC curve in Fig. 4(b) are summarized in Table 2.

Comparing the scores of potential decision models based on the EPS, as shown in the ROC curve in Fig. 4(b), with the results of the deterministic ICON 2-Mom model (inset), reveals that the EPS can significantly increase the POD for ISSR detection from 0.6 to 0.8–0.9. This improvement comes with a moderate rise in the FPR from approximately 0.1 to 0.13–0.23, depending on the decision model employed. Notably, the Youden index also improves substantially (see Table 2).

To complement this view, and following the discussion in Section 4.1.3, the precision–recall curve offers an alternative perspective that focuses specifically on the model's performance for the positive class in the context of an unbalanced dataset.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

RH_{ice} threshold	max Youden Index EPS (Det)	decision model
100%	0.70 (0.50)	1 2
105%	0.66 (0.38)	1
110%	0.60 (0.28)	1
120%	0.51 (0.17)	1

Table 2. Youden Index (YI) for the ISSR event and for events with higher ice supersaturation; the maximum YI of the EPS based decision models is shown together with the YI of the deterministic ICON 2-Mom model (single members) in brackets. The second column shows the index of the decision model(s) which correspond to the maximum YI.

It plots the POD (also referred to as recall) against the precision, thereby emphasizing the accuracy of positive predictions when the positive event is relatively rare.

Similarly to the ROC curve, in Figure 5(a), the recall-precision point for each decision model based on the EPS is plotted on a curve. For further comparison, the values of the individual ensemble members are shown. The closer a point is to the upper right corner, the better the trade-off between recall and precision. Overall, the precision is only moderate and gets worse for higher ice supersaturation events. In Figure 5(b) we also show the F_1 score, which takes into account both precision and recall, making it a useful scalar measure for determining the balance between the two:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

The F_1 score ranges from 0 to 1. For the ISSR event, the maximum F_1 score is 0.61, obtained from decision models 3, 4, and 5. The corresponding F_1 scores of the single members (deterministic models) range from 0.54 to 0.55. For events with higher ice supersaturation, decision models 2 or 3 perform best. In all cases, the corresponding F_1 score increases by about 0.06 and 0.08 compared to the deterministic models represented by the single members.

As the precision-recall curve and the F_1 score are not symmetric with respect to what we define as the "positive" event, e.g. ISSR or non-ISSR, we perform a similar analysis by defining non-ISSR as the positive event. In the context of flight routing, the correct identification of non-ISSR is also critical, as false negatives of this event could lead to unnecessary re-routing, resulting in additional CO_2 emissions. For the non-ISSR event, the maximum F_1 score is 0.94 and is given for the decision models that require at least 1-6 members with *non*-ISSR. Note that the trivial model, which always predicts non-ISSR (corresponding to decision model 0), also has a high F_1 score of 0.93. In all four event cases, the scores of decision models 0–5 (note again the adapted definition of the decision models with respect to the non-ISSR event) are very similar. Compared to the deterministic model results, the highest-scoring decision models show an increase in F_1 of 0.01–0.02.

We conclude this subsection by shifting the focus from model performance on specifically defined positive events to a more holistic evaluation using the Matthews Correlation Coefficient for the EPS-based decision models. As discussed in Section

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

4.1.3, the MCC provides a balanced assessment of model skill for both event and non-event classifications, similar to the ROC curve, and is particularly informative in the context of imbalanced datasets. In the case of ISSR/non-ISSR classification, decision models 1–7 achieve higher MCC values than their deterministic counterparts (i.e., individual ensemble members), with decision models 3 and 4 reaching a maximum MCC of 0.55. By contrast, the MCC values for the deterministic models remain around 0.47 (see Fig. 5(c)).

RH_{ice} threshold	max MCC EPS (Det)	decision model	POD	FPR
100%	0.55 (0.47)	3 4	0.80 0.73	0.13 0.10
105%	0.46 (0.37)	2 3	0.77 0.68	0.14 0.11
110%	0.37 (0.28)	2	0.64	0.11
120%	0.25 (0.16)	2	0.62	0.11

Table 3. For each RH_{ice} threshold event, the maximum MCC value of the decision models based on the EPS is shown (rounded to the second decimal place), together with the indices of the corresponding decision model(s). The MCC of the deterministic model (single members) is given in brackets. The last two columns show the ROC values (POD versus FPR) of the decision model(s) with the maximum MCC.

Table 3 shows the maximum MCC for each RH_{ice} threshold event, along with the indices of the corresponding decision models. For comparison with the ROC results, the associated POD and FPR values of these models are also provided.

To summarise this subsection, we have evaluated a range of ensemble verification metrics to assess how well our EPS model, ICON 2-Mom EPS, can distinguish between ISSR and non-ISSR conditions (or higher ice supersaturation). These metrics emphasize different diagnostic aspects: POD and FPR in the ROC curve; precision and recall (POD) in the precision–recall curve as well as in the F_1 score; and all entries of the confusion matrix in the MCC. Across all metrics considered, we observe substantial improvements in the performance of decision models based on the ensemble setup compared to the deterministic model. Depending on user requirements, a particular metric or a trade-off among metrics can be used to select the most appropriate decision model for a given application. Although the specific ID of the best ISSR decision model depends on the metric used, it consistently falls below 5 in all cases. For the remainder of this study, we limit our evaluation to the ROC curve and its associated scores, POD and FPR.

4.2.2 Longer Forecast Lead Times

So far we have focused on ICON data with a forecast lead time of 12 hours. For many flights 12 hour forecasts are sufficient. However, in general, longer forecasts should be provided. Therefore, we considered lead time increments from 12 hours up to a maximum of 48 hours, which is the standard time horizon of weather forecasts for flight planning (see Figure 6). As the

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

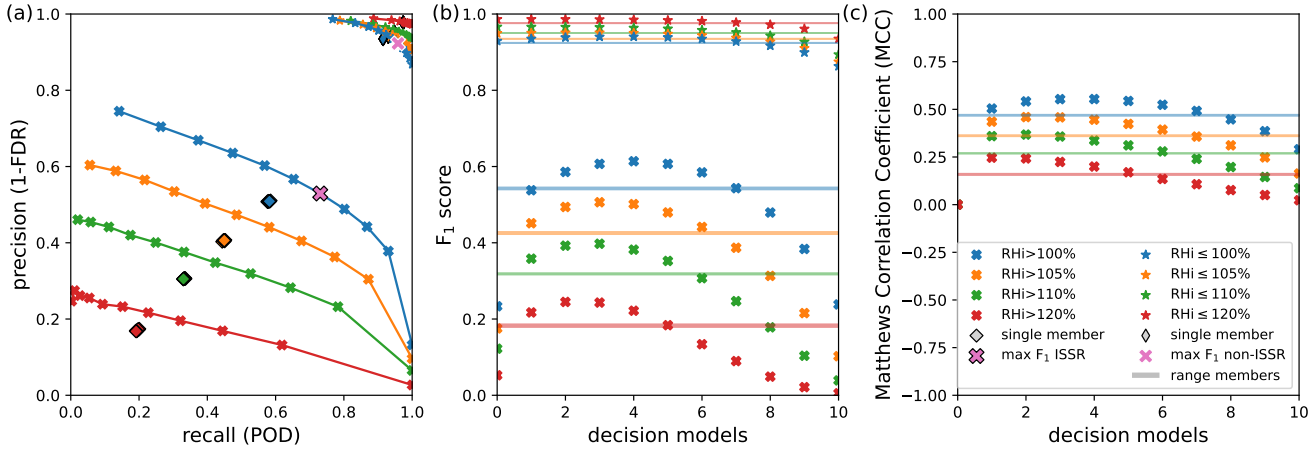

Figure 5. Scores that take into account the unbalanced dataset with respect to the ISSR event or higher ice supersaturation events in two different ways: The precision-recall curve and the F_1 score by focusing on the performance of the model with respect to what is defined as the 'positive' event, and the Matthews correlation coefficient by providing a balanced evaluation measure with respect to all four categories of the confusion matrix. (a) Precision-recall curve for the EPS: For increasing prediction probability conversion thresholds, the recall (POD) is plotted against the precision (1-FDR) of the corresponding decision model, both with respect to the 'positive' events $\{\mathrm{RH}_{ice} > \text{threshold}\}$ (bold crosses) or $\{\mathrm{RH}_{ice} \leq \text{threshold}\}$ (stars). In both cases, the scores from the decision model with the maximum F_1 score are highlighted in purple (compare (b) and note that, when F_1 is rounded to two decimal places, more decision models are optimal as discussed in Section 4.2.1). For the single ensemble members, recall is similarly plotted against precision for both types of events (diamond and thin diamond). A zoom showing the details of the top right corner is provided in the Appendix, Fig. D1 (b) F_1 scores, both for the positive events $\{\mathrm{RH}_{ice} > \text{threshold}\}$ and $\{\mathrm{RH}_{ice} \leq \text{threshold}\}$ and for 1) the EPS decision models and 2) for the single members, for which the range is shown as transparent lines. Note that the decision model index in the ISSR case is with respect to the required minimum number of ISSR events in the ensemble, while the decision model index in the non-ISSR case is with respect to the required minimum number of non-ISSR events in the ensemble. (c) Matthews Correlation Coefficient (MCC) for the EPS decision models as well as for the single members.

lead time increases, the ROC curves shift slightly to the right, indicating higher false positive rates. In contrast, no downward shift of the ROC curves is observed for high POD values of around 0.8 for the first 36 hours. The POD only starts to decrease after 36 hours. Overall, the degradation is not that severe, and at least up to 36 hours, the potential scores remain roughly in the range of POD > 0.8 and FPR < 0.2.

4.2.3 Incorporating the Ensemble Spread

The results of the ROC curve are statistical in nature, in our case from an 14-month verification period. As discussed, we aim to use them to achieve high scoring future forecasts of $\{\mathrm{RH}_{ice} > \text{threshold}\}$ through appropriate interpretation of the EPS (via decision models). Here, we further incorporate ensemble spread information in order to get more reliable scores in more specific situations. In general, the ensemble spread should be an indication of the confidence in a forecast. Therefore, in the

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

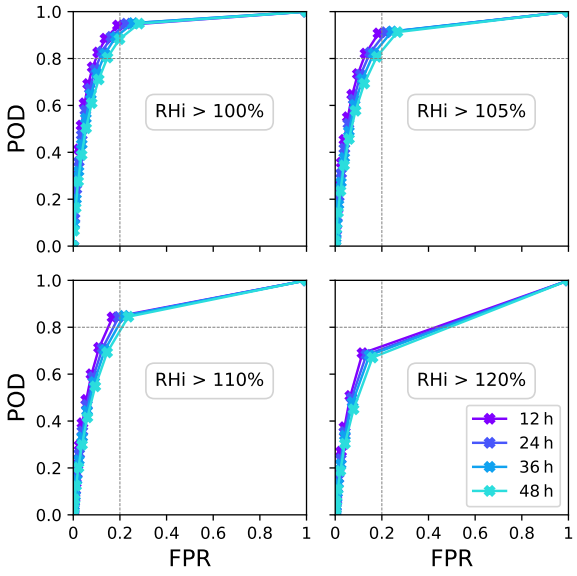

Figure 6. ROC curves for increasing forecast lead times and increasing RH_{ice} thresholds; time period five months: 1.1.2025 - 31.5.2025; ICON initial times 0 UTC and 12 UTC; Northern Hemisphere.

context of RH_{ice} forecasts, we further stratify the ROC curve in terms of the underlying ensemble spread at each grid point, particularly to achieve a lower FPR. Ensemble spread is typically measured by the standard deviation. The inset of Figure 7(a) shows a histogram of the standard deviation of $RH_{ice}/100\%$; more than 50% of the ensemble forecasts have a std below 0.1, with a peak near zero, and only a small proportion have std values greater than 0.2. The colored bins in the histogram serve as a legend for the ROC curves in the main Figure 7(a): The EPS forecasts are partitioned with respect to their std and the corresponding ROC curves are shown in the same color. The general trend is consistent with our expectation; the lower the std, the closer the corresponding ROC curve is to the upper left corner, and vice verse, the higher the std, the closer the ROC curve is to the diagonal, indicating that the model has low skill in these cases. In more detail, in more than half of the cases a significantly improved ROC is obtained with POD between 0.9 and 1 and FPR $\leq$0.1 for the ISSR condition with decision models 1-2. In case the std is greater than 0.1, the ROC curves tend more and more to the diagonal and at least five or six members should indicate ice supersaturation to achieve an FPR of $\leq$0.1 (indicated by the vertical magenta line). In these cases, only a lower POD can be obtained, between 0.8 and 0.3, depending on the underlying std.

As the shape of the ROC curves varies significantly along different std regimes, we were also interested in the std values of different RH_{ice} regimes, particularly when RH_{ice} is around or above 100%. In Fig. 7(b), summary statistics of std are shown for increasing 10% bins of RH_{ice}. Following an increase in std values, they fall before 100 % and reach another local minimum in the RH_{ice} regime of 100 %-110 % with a median around 0.1. The relative mean squared error (RMSE) shows a similar

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

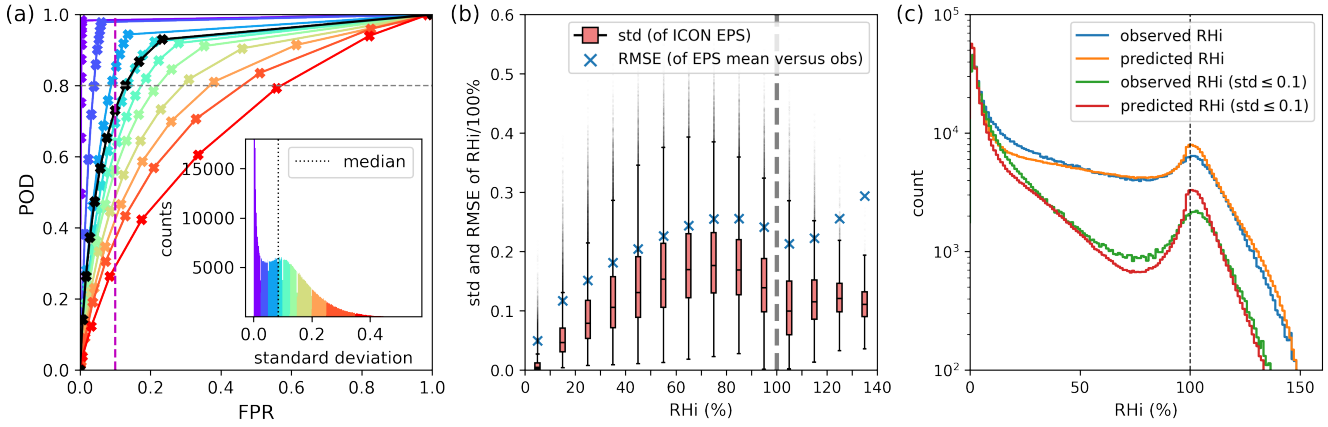

Figure 7. Event RH_{ice}>100%: Inclusion of the ensemble spread of RH_{ice}, measured by the standard deviation (std) of RH_{ice}/100%. (a) ROC stratification along the standard deviation; the inset shows the histogram of the standard deviation of RH_{ice}/100% and also serves as a legend for the ROCs on EPS subsets with associated std; the black ROC is the original one without std stratification. (b) Standard deviation and RMSE for 10% bins of the predicted RH_{ice} mean; the coral coloured boxes represent the interquartile range (IQR) (middle 50% of the std data) and the black horizontal line inside the boxes represents the median. The bottom of the box is Q1 (25th percentile) and the top is Q3 (75th percentile). The vertical lines extending from the boxes represent the variability of the data outside Q1 and Q3. They typically reach the minimum and maximum values within 1.5 × IQR. All data points outside 1.5 × IQR from Q1 or Q3 are plotted individually as outliers. Blue crosses indicate the RMSE between the ensemble mean and the observed data points. (c) Full histograms of observed and predicted RH_{ice} values and histograms conditioned on std≤0.1 are shown; in the observation case, the corresponding std values were defined by the corresponding spatio-temporally matching EPS values. In the EPS model case, the counts were divided by 10 to obtain a similar range of values to the observations.

qualitative behavior for $RH_{ice} < 120\%$. For higher RH_{ice} regimes, the RMSE increases to its maximum over the whole RH_{ice} value range.

We take another perspective in Fig. 7(c), where the full RH_{ice} histograms of the observations and the ensemble forecasts are shown, as well as both conditioned on std≤0.1; in the case of the observations this is done by assigning the std-value of the corresponding spatio-temporal EPS matching point. For low std-values (std≤0.1), the corresponding conditional RH_{ice} histograms show a large peak for low humidity values in the same range as the full unconditioned histograms. Another peak is observed for RH_{ice} values around 100%, which is approximately one order of magnitude lower than that of the unconditioned histograms. This difference persists in the supersaturation tail of the histograms, where the maximum RH_{ice} values reached in the conditional case are around 130%, based on the 820 000 verification points (where all counts below 100 were cut in this plot). When comparing the conditional histograms of the model and the observations, the observation histogram exhibits a slightly lower peak around 100%, similar to the difference observed in the full histograms. In conclusion, even when the model exhibits high confidence, as reflected by a low standard deviation, the histogram still displays intermediate supersaturation. This suggests that certain ISSRs can be well predicted.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

The results shown in Figures 7(b) and 7(c) are similar to the findings of Borella et al. (2024) who parameterized the subgrid-scale distribution of water vapor in the UTLS using IAGOS data. They identified mostly quadratic behavior of the standard deviation of RH_{ice} relative to the mean value of RH_{ice} itself, with a maximum peak between 70% and ∼110% depending on temperature, before exhibiting an upward trend for even higher RH_{ice} values. Their temperature analysis revealed that this peak becomes lower and moves to larger RH_{ice} values as the temperature decreases. Our ROC stratification approach does not consider temperature, but may do so in further studies.

The increased predictability in the regime around $RH_{ice} \approx 100\%$ can be explained by a more stable microphysical behavior in this near-thermodynamic equilibrium state, which is captured by the model. In this regime, mature cirrus clouds are dominant compared to young or short-lived cirrus clouds which often form in regions of high ice supersaturation, driven by upward motion from gravity waves or deep convection. These young clouds experience rapid crystal growth due to significant mesoscale temperature fluctuations (MTFs) caused by gravity waves, which create high spatio-temporal variability in supersaturation. The fluctuating vertical motions and ice crystal concentrations make forecasting cloud evolution difficult. As a result, young and short-lived cirrus clouds introduce significant uncertainty in predicting supersaturation, as the microphysical processes are highly dynamic and rapidly changing. In contrast, mature cirrus clouds, approaching thermodynamic equilibrium ($RH_{ice} \approx 100\%$), display weak supersaturation conditions, typically linked to slow, steady-state ascent. Under these conditions, ice crystals grow and gradually deplete ambient water vapor, creating a balanced system that enhances the predictability of ice crystal evolution and overall cloud dynamics.

In clear-sky regions, where clouds and associated microphysical processes are absent, the predictability of RH_{ice} is governed primarily by large-scale thermodynamic and dynamical processes. Supersaturation can persist in these regions due to the lack of ice nuclei. Observations show that clear-sky supersaturation is often associated with weak vertical motions and low temperatures in the upper troposphere, particularly in mid- and high-latitude regions (Kahn et al., 2009). However, MTFs caused by gravity waves can still occur, challenging predictability, particularly for models that do not resolve mesoscale temperature or humidity fluctuations. Overall, while the absence of cloud feedbacks simplifies the microphysical environment, potential variability in temperature, humidity, and vertical motion still introduces uncertainty, i.e., the predictability of RH_{ice} in clear skies depends on the given specific large- and mesoscale thermodynamic and dynamical processes.

4.2.4 Comparison with IAGOS Data

The RH_{ice} density of the IAGOS data, limited to the Northern Hemisphere for better comparison with our radiosonde verification, confirms the characteristic bimodal shape of the RH_{ice} density (see inset of Fig. 8). Compared to the ICON (and radiosonde) data, the first peak in the IAGOS density appears shifted to the right, suggesting fewer near-zero RH_{ice} values in the IAGOS dataset than in the ICON data. The peak around $RH_{ice} \approx 100\%$ is shifted to the left and is less pronounced in the IAGOS data. It also does not reach the same high RH_{ice} values as ICON. Nevertheless, at least up to the event $RH_{ice} > 120\%$, the shape of the ROC curves closely resembles that of the radiosonde data (see Fig 8).

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

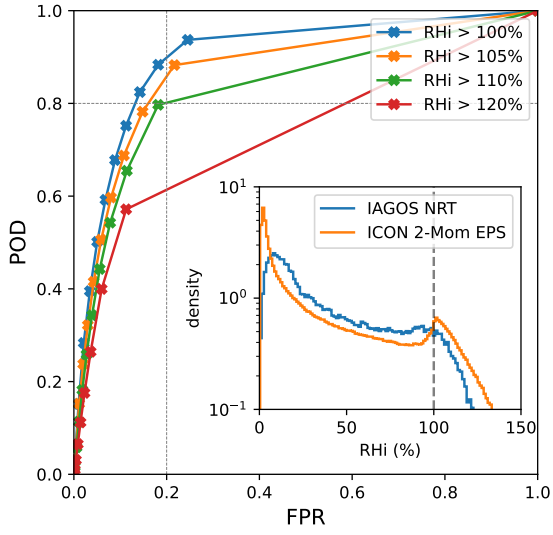

Figure 8. ROC curve of ICON 2-Mom EPS and IAGOS data, the inset figure shows the corresponding RH_{ice} densities. Evaluation performed with 625 flights from four months (August 2024, October 2024, December 2024, January 2025) on the Northern Hemisphere, leading to ∼200 000 spatio-temporal samples.

5 Discussion

Below we discuss the interrelationships of our results and their implications, particularly in the context of climate-optimized flight routing. We also consider the ROC curve of the ICON 1-Mom EPS to see what we can gain from an ensemble setup in case of the one-moment cloud ice microphysics scheme. We compare our results with those of a recent study and discuss promising approaches, such as neighborhood inclusion and, more generally, machine learning approaches, to build more sophisticated meta-models with improved scores.

5.1 Interrelationships of Results and Application Implications

For the RH_{ice}>100% event, the two-moment ice microphysics scheme introduced here significantly improves the POD compared to the operational one-moment scheme. The trade-off is a slightly higher FPR; as seen in the FBI, the scheme identifies slightly more events than are actually observed by radiosondes. However, the Matthews correlation coefficient, which is a more balanced measure for all four categories TP, FP, FN, TN, is also increased by ICON 2-Mom compared to ICON 1-Mom.

For events with larger RH_{ice} thresholds, the one-moment scheme breaks down almost completely, while the performance of the two-moment scheme deteriorates only moderately. The introduction of a 10-member ensemble setup with k-out-of-10 decision models allows fine-grained control of the balance between POD and FPR. The deterministic ICON 2-Mom model has a POD of about 0.6 and an FPR of about 0.1 for the RH_{ice}>100% event, while the ensemble setup covers a wide range

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

depending on the decision model. For example, if the goal is to detect as many ISSR events as possible, decision model 1 offers a POD of over 0.9 at an FPR of 0.25. At the other end of the range, decision model 9 has a POD of just under 0.3 but an FPR of almost 0. For most applications, the optimal decision model may lie somewhere in between. If adopted by aviation, the right balance would be found by quantifying the exact costs of false positives (unnecessary diversions) and false negatives (condensation trails). The ability to predict RH_{ice} events well above 100% may prove helpful in estimating costs.

If the ensemble spread is also taken into account, even finer control of POD or FPR is possible. Stratifying the ROC curve by the standard deviation of RH_{ice} reveals that situations where ensemble members are in strong agreement tend to yield good categorical scores (ROC curve near the upper-left corner), whereas situations with large ensemble standard deviations result in values that are only marginally better than random chance (ROC curve near the diagonal). Notably, this stratification requires only the ensemble data itself and can therefore be incorporated into the meta-model. For instance, if the primary objective is to keep the FPR below 0.1, decision model 1 suffices for low-spread data, whereas decision models 5 or 6 are more appropriate when dealing with high-spread conditions. This approach opens the possibility of constructing more refined models with improved scores by combining the basic decision models.

5.2 Ensemble Verification of ICON 1-Mom EPS

We also evaluated the ensemble data of the operational ICON 1-Mom EPS with respect to RH_{ice}. We wanted to compare the improvement of results such as the POD due to the ensemble setup when the microphysical scheme has not been adapted to a two-moment scheme. Therefore, we considered the ROC curve for the operational 40-member EPS as well as for 10-member subsets, compare Appendix Figure B1. By similarly defining decision models for ISSR, the POD can be increased to more than 0.8 with an FPR remaining below 0.2, which holds true for both the 40- and 10-member EPS. The full EPS yields a more fine-grained curve with slightly higher POD values in the top left corner than the 10-member EPS. Overall, the potential of an ensemble is highlighted in both cases, especially with respect to a possible increase in POD. However, the operational 1-Mom EPS still fails to predict events with higher RH_{ice} values (see inset in Fig. B1), as it relies on an NWP model with insufficient physical parameterization for larger RH_{ice} values. This finding again confirms that a high quality model is a fundamental part of the success of an EPS (Wang et al., 2018; Du et al., 2018).

Finally, we wanted to confirm that the specific selection of ten members from the original 40 had little or no effect on the scores due to the way the ensemble is generated. Therefore, we performed a 10-out-of-40 bootstrap and considered the mean and standard deviation of the corresponding points of the ROC curves of each subset EPS. The resulting standard deviation is negligibly small, encouraging us to transfer this finding to our ICON 2-Mom EPS, using the first ten members.

5.3 Model Resolution and Neighborshood Consideration

Several leading high-resolution NWP models have been validated with respect to RH_{ice} using radiosonde data in Thompson et al. (2024). The radiosonde data used were from 2022, covering ten months, and included data from radiosondes of lower or unknown quality than the Vaisala RS41 radiosondes. Model data were interpolated onto radiosonde data, which differs from our approach of interpolating radiosonde data onto model data. The most comparable results are the POD and FPR for

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

$RH_{ice} > 99.99\%$ events, where (POD, FPR) values of (0.46, 0.09) were obtained for the S-WRF model, (0.19, 0.02) for the GFS, and (0.50, 0.10) for the IFS. In all cases, the deterministic model was evaluated.

The study also introduced a 3D neighborhood verification, where the number of ISSR events of horizontal and vertical grid point neighborss affects the identification (definition) of true positives, false positives, false negatives and true negatives. Although in this study neighborhood incorporation is used for model comparison verification, it could also be used to define another meta-model - in this case not based on an EPS model, but on a deterministic NWP model. Of course, a similar definition could also be introduced based on an EPS model. However, although the concept of including neighbors into a model to identify ISSRs is worth exploring, the neighborhood verification presented in the study corresponds to two different models, where the one to be used is individually selected for each radiosonde observation, depending on whether ISSR was actually observed or not. This conditioning on the observation may improve the verification results, as the knowledge of the observation determines the decision of which model to use. For our purpose, which is to define a model for future predictions, it is not appropriate to condition this decision on the observation. But even when including only model neighbors values into a meta-model, the grid resolution we currently use (about 26 km horizontally and about 200 m vertically in the height range of interest) may be too low to adequately account for horizontal neighbors.

5.4 Prediction Improvement via Machine Learning

As evidenced by the ad hoc nature of decision models in both prior studies and this work, there is value in pursuing a more general approach to post-processing NWP data. While the k-out-of-10 decision models are based on intuitive thresholds, they are ultimately heuristic in nature – comparable to, for example, a binary deep neural network classifier trained and validated on model and radiosonde data. Due to the small amount of data ($\sim$ 820 000 samples), we chose to use the gradient boosting library CatBoost in classification mode. The results are shown in Appendix Fig. E1. The CatBoost model shows a slight improvement in the upper left region of interest compared to the k-out-of-10 decision models. In addition, the ROC curve is almost continuous and at high RH_{ice} gives access to POD values that are unattainable even for the 1-out-of-10 model, giving a greater degree of control over the desired balance between POD and FPR. Thus, the model reduces the need to run an EPS with many members (but more members slightly improve the predictions; see the 1-moment case in Fig. B1). Another advantage of the model is that more features than just RH_{ice} itself can easily be added as model inputs. Even extending the feature vector with physical quantities of neighboring cells is equally feasible. The results are very promising and more complex models are being investigated.

6 Conclusions

This study demonstrates the great potential of an EPS model for ISSR prediction, based on the ICON NWP model with an adapted two-moment ice microphysics scheme. The two-moment scheme more accurately captures the physical conditions associated with (higher) ice supersaturation, which many one-moment schemes struggle to represent or fail to identify. Prior

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

to evaluating the ensemble setup, the two-moment scheme underwent a careful verification process to confirm its suitability to represent ice supersaturation in NWP applications.

The EPS model itself has served as the foundation for further meta-model developments aimed at constructing deterministic models of ISSR/non-ISSR classification and higher ice supersaturation. These models are designed to provide flight planners with well-scored predictive tools that support actionable decision-making.

Simple k-out-of-N decision models spanned a wide range of POD-FPR values, with many of them achieving a significantly higher POD than the original deterministic NWP model while having only slightly worse FPRs. The k-out-of-N models were further used to define another meta-model by adaptively choosing k according to the ensemble spread, where situations with strong agreement of all members use a smaller k, and situations with disagreement use a larger k, in order to keep the FPR below a certain target level.

These approaches were statistical in nature, meaning that we used classical statistical methods and verification results to define a meta-model that functions as a newly developed forecast model. Building on this methodology, we trained a gradient boosting tree classifier representing a more advanced meta-model. Despite being trained in under a minute using default hyperparameters, the model outperformed the k-out-of-N models in the POD–FPR region of interest. Additional advantages of this model include an almost continuous ROC curve and its ability to integrate additional features in a straightforward manner.

While these investigations on the characteristics of the ICON 2-Mom EPS system were ongoing, a contrail avoidance trial based on the ensemble mean of this system rerouted more than 100 flights. The results presented in this study demonstrate that EPS-based meta-models bring us even closer to reliably identifying the potential for persistent contrail formation.

The results of this study can also be informative for the European Union's recently established Monitoring, Reporting and Verification (MRV) system, where climate response models are used to quantify the trade-off between contrails, CO_2 emissions and other greenhouse gases. Climate response models require up to 15 meteorological parameters, such as humidity, temperature, pressure and wind fields, of which RH_{ice} is of utmost importance for the contrail component, and it is RH_{ice} that is often poorly predicted by state-of-the-art operational NWP models. This study is a step towards improved prediction of ISSR and RH_{ice}.

Code and data availability. The verification code and data are available under Zenodo (https://doi.org/10.5281/zenodo.15881140).

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Appendix A: The Two-Moment Cloud Ice Scheme

The two-moment cloud ice scheme in ICON is an extension of the operational one-moment cloud ice scheme. It adds a prognostic equation for cloud ice number density and includes explicit ice nucleation processes. The original one-moment scheme is a legacy code developed by Günther Doms at DWD in the 1990s for the COSMO model, which was then known as the Lokalmodell (LM), and operated at a horizontal grid spacing of 7 km (Steppeler et al., 2003). In the 2000s, the same one-moment scheme was used in the operational global model GME, the predecessor of ICON (Majewski et al., 2002). A detailed description of the original one-moment cloud ice scheme is provided in Doms et al. (2021). It shares many similarities with the one-moment schemes by Lin et al. (1983) and Rutledge et al. (1986), both originally developed for mesoscale models.

Over the past 25 years, the operational one-moment cloud ice scheme has undergone many modifications, documented in Section 5.8 of the COSMO 6.0 documentation. Notable updates include warm-rain processes based on Seifert and Beheng (2001), snow particle geometry following Wilson and Ballard (1999), and snow size distributions derived from empirical relationships by Field et al. (2005). Ice crystal concentration is parameterized using the empirical formula by Cooper (1986).

In the two-moment scheme, the diagnostic ice particle number concentration is replaced by a prognostic equation. Examples of similar hybrid schemes include those by Reisner et al. (1998) and Thompson et al. (2004), though these originally used purely temperature-dependent ice initiation. Köhler and Seifert (2015) present a two-moment scheme that accounts for deposition nucleation based on ice supersaturation, and includes homogeneous freezing of sulfate aerosol droplets at low temperatures.

The version of the two-moment scheme used in this study is a simplified and updated version of Köhler and Seifert (2015, hereafter KS15). The two-mode representation in KS15 is omitted for computational efficiency, as are the timestep refinements for homogeneous nucleation.

In a two-moment scheme, sources and sinks of ice particles must be explicitly parameterized. The three primary sources of ice particles are homogeneous nucleation, heterogeneous nucleation, and detrainment of ice from deep convective clouds.

A1 Deep Moist Convection

ICON parameterizes moist convection using a bulk mass flux convection scheme (Tiedtke, 1989; Bechtold et al., 2008). For cloud ice detrainment from convection, a mean particle diameter of $D_{i,\mathrm{conv}} = 200\ \mu$m is assumed, corresponding to a mean mass of $m_{i,\mathrm{conv}} = 10^{-9}$ kg. A smaller mean mass would increase the number of detrained ice particles in the upper troposphere, leading to shorter phase relaxation times in convective anvils and reduced ice supersaturation. The assumed size also affects the effective radius of anvil clouds explicitly represented in the model.

A2 Homogeneous Ice Nucleation

For homogeneous ice nucleation, the parameterization by Kärcher et al. (2006) is used. It accounts for the presence of pre-existing ice particles and is applied using grid-scale vertical velocity and ice supersaturation. However, this neglects subgrid-scale variability, which may lead to an underestimation of nucleation events. The impact on cloud ice number concentration is

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

less straightforward. While nucleation events in nature occur on much smaller spatial scales, the model assumes that nucleated particles are evenly distributed across the grid box once the event is triggered.

A3 Heterogeneous Ice Nucleation

Heterogeneous nucleation is represented using the INAS (Ice Nucleating Active Sites) approach of Ullrich et al. (2017), which includes parameterizations for deposition and immersion freezing on mineral dust and soot. Since prognostic aerosol fields are not available in ICON, but only in ICON-ART, a constant dust number concentration of $N_{\text{dust}} = 1 \times 10^3$ m^{-3} is assumed in the upper troposphere above $p_0 = 200$ hPa. Below that pressure height the profile increases following

$$N_{\text{dust}}(p) = N_{\text{dust,0}} \max\left\{\min\left[\exp\left(\gamma_{\text{dust}} \frac{p}{p_0}\right), 200\right], 1\right\} \tag{A1}$$

with $\gamma_{\text{dust}} = 1 \times 10^{-3}$. The dust surface area $\bar{S}_{\text{dust}}$ is calculated based on a lognormal particle size distribution with a mean diameter of 1 μm and a standard deviation of 2.5. The number of nucleated ice particles is then diagnosed as:

$$N_i^* = N_{\text{dust}} \left\{1 - \exp\left[-\bar{S}_{\text{dust}} n_S(T, S_i)\right]\right\} \tag{A2}$$

Here, n_S is the INAS density in m^{-2}, parameterized according to Eq. (7) for deposition and Eq. (5) for immersion freezing in Ullrich et al. (2017).

In numerical models, newly formed ice particles are typically diagnosed each timestep using $\Delta N_i = N_i^* - N_i^{\text{pre}}$, where N_i^{pre} is the number of pre-existing ice particles. However, this can overestimate heterogeneous nucleation since N_i^{pre} is reduced by sedimentation or aggregation, while N_{dust} remains constant. This effectively creates an unlimited reservoir of ice-nucleating particles. To avoid this artifact, a budget variable is introduced as described in KS15. A relaxation timescale of two hours is applied to simulate the recovery of nucleating particle availability due to atmospheric mixing.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Appendix B: Ensemble Setup of Operational ICON

Fig. B1 compares the ROC curves of the ICON 1-Mom EPS and the ICON 2-Mom EPS for $\mathrm{RH}_{ice} > 100\%$ events. The 1-Mom EPS is noteworthy as it is the operational ICON model. The 1-Mom 10-member data was obtained by resampling. Overall, the 2-Mom 10-member EPS performs as well as the 1-Mom 40-member EPS, while the 1-Mom 10-member EPS performs slightly but significantly worse. For larger $\mathrm{RH}_{\mathrm{ice}}$ thresholds, the 1-Mom EPS breaks down, as shown in the inset.

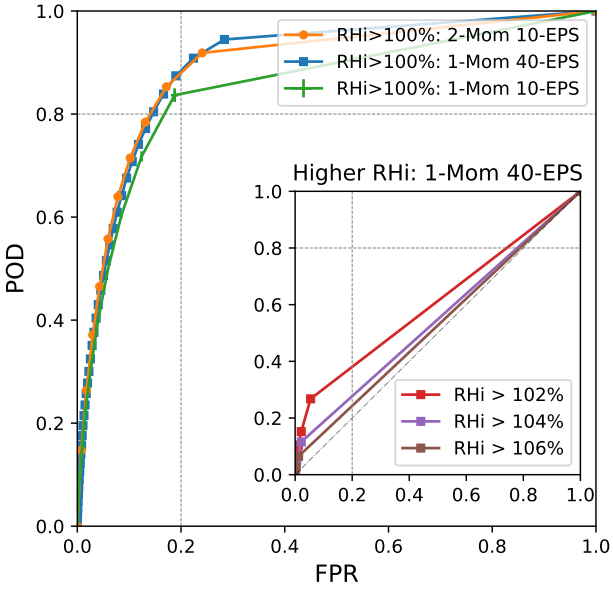

Figure B1. ROC curves for the ICON 2-Mom EPS (orange), the operational ICON 1-Mom EPS with 40 members (blue) and for the corresponding ICON 1-Mom EPS subsets with 10 members (green). For the latter, we randomly selected 1000 10-member EPS subsets, calculated the ROC curve for each and plotted the mean and standard deviation of the corresponding points on the curve. The inset figure shows ROC curves for the ICON 1-Mom 40-member EPS for higher $\mathrm{RH}_{\mathrm{ice}}$ thresholds up to 106%. Evaluation for three months (August 2024, October 2024, January 2025); ICON initial times 0 and 12 UTC; ICON forecast lead time 12h; Northern Hemisphere.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Appendix C: Calculation of RH_{ice}

C1 Computation of RH_{ice} for radiosonde data

In the TEMP BUFR files, as disseminated through the Global Telecommunication System (GTS), the dew point temperature (T_d) is provided, which allows us to compute the water vapour partial pressure (e) using the formula from Hardy (1998), ensuring consistency with the processing applied by radiosonde manufacturers, such as Vaisala. We further calculate the saturation vapour pressure over ice (e_i) consistently with the formula used in ICON which is given by

$$e_i = b_1 \frac{\exp(b_{2i}(T - b_3))}{T - b_{4i}} \tag{C1}$$

with coefficients

$b_1 = 610.78$, $b_{2i} = 21.87$, $b_3 = 273.16$, $b_{4i} = 7.66$.

and referred to as the Magnus-Tetens-Murray approximation (Magnus, 1844; Tetens, 1930; Murray, 1967). Therewith, we receive

$$\mathrm{RH}_{ice} = \frac{e}{e_i} 100\%. \tag{C2}$$

C2 Computation of RH_{ice} for ICON Data

First we calculate the water vapour partial pressure e by

$$e = r_v \, T \, \rho \, qv,$$

where the temperature T (in K), the density of moist air ρ (in kg/m^3), and the specific water vapour content qv (in kg/kg) are output variables of ICON, and $r_v = 461.51$ is the gas constant for water vapour. Finally, we calculate e_i again with C1 and RH_{ice} with C2.

Note that recently, as of May 2025, the coefficients in the C1 formula for the saturation vapour pressure over ice in the operational ICON model have been updated. We still use the old version of the coefficients given in C1 in our dedicated system and therefore in our verification analysis. However, at -37°C, the error is only about 2%.

C3 Computation of RH_{ice} for IAGOS Data

In the IAGOS NRT dataset, RH_{ice} is already included and has been calculated using the formula from Sonntag (1994), which is very similar to the Hardy formula.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Appendix D: Details of Precision-Recall Curve for non-ISSR

In Fig. D1 a zoom of the top right of Fig. 5 is provided, where the details of the precision-recall curve for the non-ISSR event and for the events $\{\mathrm{RH}_{ice} \leq \text{threshold}\}$ with threshold in $\{105\%, 110\%, 120\%\}$ can be seen. Note that *decision model* k here refers to the decision model which requires at least k events with $\{\mathrm{RH}_{ice} \leq \text{threshold}\}$.

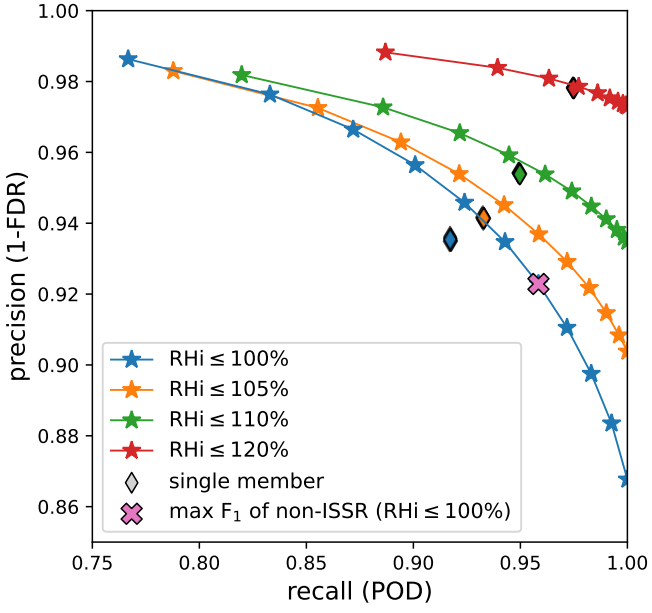

Figure D1. Precision-recall curve of events $\{\mathrm{RH}_{ice} \leq \text{threshold}\}$ with threshold in $\{100\%, 105\%, 110\%, 120\%\}$ (zoom of top right of Figure 5). Stars on the lines indicate the scores corresponding to the decision models based on the EPS. The pink cross highlights decision model 4 for which the maximum F_1 score is obtained. Thin diamonds inidcate the scores of the single ensemble members.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Appendix E: Binary Classification Models: CatBoost

CatBoost is a machine learning library based on gradient boosting on decision trees, where input features are either real values or categorical values. Prediction can happen either as regression or classification. For the task at hand, CatBoost was used in classification mode, with the cross-entropy loss J used for training:

$$J(\boldsymbol{y},\boldsymbol{p}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

where N is the total number of samples (spatio-temporal matching points of model and observation), y_i is 1 if an event was observed, otherwise 0, and p_i is the prediction probability of the model. The samples were divided into 75% training and validation data and 25% test data. The test data were taken from different months than the training/validation data to minimise the effect of potential correlations in the data. Figure E1 shows the performance of the model on the test data, compared to the EPS-based decision models of this study applied to the test data period.

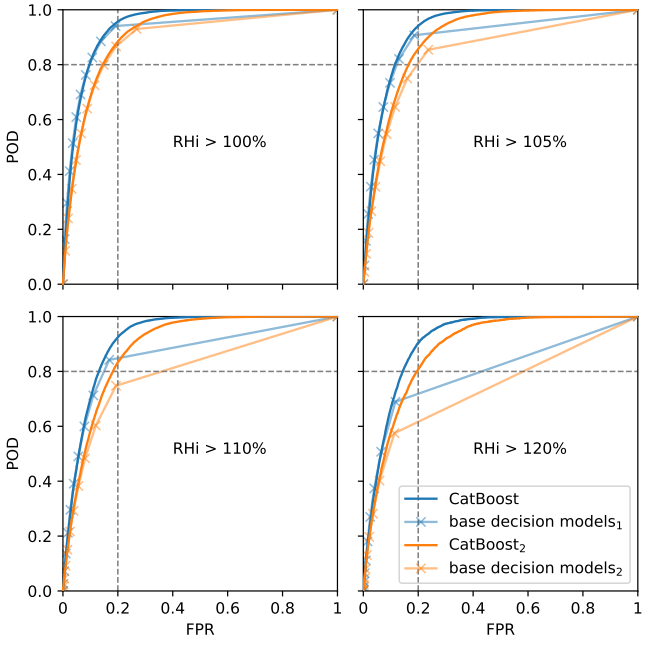

Figure E1. Comparison of ROC curves of EPS-based CatBoost and EPS-based decision model. CatBoost input features were the RH_{ice} values of all ten members and the mean temperature. Solid blue ROC curves: Training and validation period from April to December 2024; test data from January to March 2025; ROC calculated for the test data period. Light blue ROC curves: ROC for the EPS-based decision model, evaluated over the test data period. Solid and light orange curves indicate the same setting but with a different training and validation data period (July 2024 to March 2025) and a different test data period (April-June 2024). Except for a larger tree depth of 10, all CatBoost settings were kept at default, and training took about 30 seconds per RH_{ice} threshold.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Author contributions. LS and AS supervised and conceived the study, with contributions from MH and CK. AS implemented the adapted two-moment ice microphysics scheme in ICON, based on a former COSMO/GME implementation from CK. AS set up the dedicated ICON system at DWD. MH implemented the verification methods and performed the analysis. MH wrote the paper, with contributions from CK and AS. All authors discussed the results and contributed to the final version of the paper.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. MH and CK are funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under the German Aerospace Research Programme (LuFo) (FKZ: 20M2111F).

The authors thank the D-KULT project partners, especially Klaus Gierens, for helpful discussions. The authors also thank Prof. Ulrich Schumann for insightful discussions and sharing knowledge on the topic. Ruud Dirksen is acknowledged for providing valuable insights on radiosonde measurements.

Further thanks go to John Walter Acevedo Valencia for introducing MH to IAGOS data and to Susanne Rohs and Damien Boulanger for their generous support on IAGOS data and for providing the IAGOS NRT data. IAGOS data were created with support from the European Commission, national agencies in Germany (BMBF), France (MESR), and the UK (NERC), and the IAGOS member institutions (http://www.iagos.org/partners). The participating airlines (Lufthansa, Air France, Austrian, China Airlines, Hawaiian Airlines, Air Canada, Iberia, Eurowings Discover, Cathay Pacific, Air Namibia, Sabena) supported IAGOS by carrying the measurement equipment free of charge since 1994. The data are available at http://www.iagos.fr thanks to additional support from AERIS.

The authors thank Björn Beckmann for fruitful discussions on detailing the dedicated system, Thomas Hanisch and Tobias Göcke for helping to set it up, and Sven Ulbrich for curating its data. Christoph Gebhardt, Chiara Marsilli and Jochen Förstner are acknowledged for providing details on the ensemble generation at DWD. MH especially thanks Felix Reinhardt for fruitful discussions on the verification implementation and insightful discussions on the results.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

References

Achatz, U., Alexander, M. J., Becker, E., Chun, H.-Y., Dörnbrack, A., Holt, L., Plougonven, R., Polichtchouk, I., Sato, K., Sheshadri, A., et al.: Atmospheric gravity waves: Processes and parameterization, Journal of the Atmospheric Sciences, 81, 237–262, 2024.

Appleman, H.: The formation of exhaust condensation trails by jet aircraft, Bulletin of the American Meteorological Society, 34, 14–20, 1953.

Bechtold, P., Köhler, M., Jung, T., Doblas-Reyes, F., Leutbecher, M., Rodwell, M. J., Vitart, F., and Balsamo, G.: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales, 134, 1337–1351, 2008.

Berner, J., Ha, S.-Y., Hacker, J., Fournier, A., and Snyder, C.: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations, Monthly Weather Review, 139, 1972–1995, 2011.

Bickel, M., Ponater, M., Bock, L., Burkhardt, U., and Reineke, S.: Estimating the effective radiative forcing of contrail cirrus, Journal of Climate, 33, 1991–2005, 2020.

Borella, A., Vignon, É., Boucher, O., and Rohs, S.: An empirical parameterization of the subgrid-scale distribution of water vapor in the UTLS for atmospheric general circulation models, Journal of Geophysical Research: Atmospheres, 129, e2024JD040 981, 2024.

Borg, L. A., Dirksen, R. J., and Knuteson, R. O.: Land-based cal/val campaigns, in: Field Measurements for Passive Environmental Remote Sensing, pp. 219–233, Elsevier, 2023.

Cooper, W. A.: Ice Initiation in Natural Clouds, pp. 29–32, American Meteorological Society, Boston, MA, https://doi.org/10.1007/978-1-935704-17-1_4, 1986.

Dietmüller, S., Matthes, S., Dahlmann, K., Yamashita, H., Simorgh, A., Soler, M., Linke, F., Lührs, B., Meuser, M. M., Weder, C., et al.: A python library for computing individual and merged non-CO 2 algorithmic climate change functions: CLIMaCCF V1. 0, Geoscientific Model Development Discussions, 2022, 1–33, 2022.

Dirksen, R. J., Haefele, A., Vogt, F. P., Sommer, M., von Rohden, C., Martucci, G., Gonzague, R., Felix, C., Modolo, L., Vömel, H., Simeonov, T., Oelsner, P., Edwards, D., Oakley, T., Gardiner, T., and Ansari, M. I.: Report of WMO's 2022 Upper-Air Instrument Intercomparison Campaign, in: Instruments and Observing Methods Report No. 143, pp. 1–400, WMO, 2022.

Doms, G., Förstner, J., Heise, E., Herzog, H.-J., Mironov, D., Raschendorfer, M., Reinhardt, T., Ritter, B., Schrodin, R., Schulz, J.-P., and Vogel, G.: A Description of the Nonhydrostatic Regional COSMO Model: Part II Physical Parameterizations (v 6.0), Deutscher Wetterdienst, https://doi.org/10.5676/DWD_pub/nwv/cosmo-doc_6.00_II, 2021.

Du, J., Berner, J., Buizza, R., Charron, M., Houtekamer, P. L., Hou, D., Jankov, I., Mu, M., Wang, X., Wei, M., et al.: Ensemble methods for meteorological predictions, 2018.

Epstein, E. S.: Stochastic dynamic prediction, Tellus, 21, 739–759, 1969.

Field, P. R., Hogan, R. J., Brown, P. R., Illingworth, A. J., Choularton, T. W., and Cotton, R. J.: Parametrization of ice-particle size distributions for mid-latitude stratiform cloud, 131, 1997–2017, 2005.

Grewe, V., Dahlmann, K., Flink, J., Frömming, C., Ghosh, R., Gierens, K., Heller, R., Hendricks, J., Jöckel, P., Kaufmann, S., et al.: Mitigating the climate impact from aviation: Achievements and results of the DLR WeCare project, Aerospace, 4, 34, 2017a.

Grewe, V., Matthes, S., Frömming, C., Brinkop, S., Jöckel, P., Gierens, K., Champougny, T., Fuglestvedt, J., Haslerud, A., Irvine, E., et al.: Feasibility of climate-optimized air traffic routing for trans-Atlantic flights, Environmental Research Letters, 12, 034 003, 2017b.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Grewe, V., Gangoli Rao, A., Grönstedt, T., Xisto, C., Linke, F., Melkert, J., Middel, J., Ohlenforst, B., Blakey, S., Christie, S., et al.: Evaluating the climate impact of aviation emission scenarios towards the Paris agreement including COVID-19 effects, Nature communications, 12, 3841, 2021.

Hardy, B.: ITS-90 formulations for vapor pressure, frostpoint temperature, dewpoint temperature, and enhancement factors in the range–100 to+ 100 C, in: The proceedings of the third international symposium on Humidity & Moisture, Teddington, London, England, pp. 1–8, 1998.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D: Nonlinear Phenomena, 230, 112–126, 2007.

Kahn, B. H., Gettelman, A., Fetzer, E. J., Eldering, A., and Liang, C. K.: Cloudy and clear-sky relative humidity in the upper troposphere observed by the A-train, Journal of Geophysical Research: Atmospheres, 114, 2009.

Kärcher, B.: Formation and radiative forcing of contrail cirrus, Nature communications, 9, 1824, 2018.

Kärcher, B., Hendricks, J., and Lohmann, U.: Physically based parameterization of cirrus cloud formation for use in global atmospheric models, J. Geophys. Res., 111, 2006.

Kärcher, B., DeMott, P., Jensen, E., and Harrington, J.: Studies on the competition between homogeneous and heterogeneous ice nucleation in cirrus formation, Journal of Geophysical Research: Atmospheres, 127, e2021JD035 805, 2022.

Klöwer, M., Allen, M., Lee, D., Proud, S., Gallagher, L., and Skowron, A.: Quantifying aviation's contribution to global warming, Environmental Research Letters, 16, 104 027, 2021.

Köhler, C. G. and Seifert, A.: Identifying sensitivities for cirrus modelling using a two-moment two-mode bulk microphysics scheme, Tellus B: Chemical and Physical Meteorology, 67, 24 494, 2015.

Lee, D. S.: The current state of scientific understanding of the non-CO2 effects of aviation on climate, UK Department for Transport, Manchester Metropolitan University, 2018.

Lee, D. S., Allen, M. R., Cumpsty, N., Owen, B., Shine, K. P., and Skowron, A.: Uncertainties in mitigating aviation non-CO 2 emissions for climate and air quality using hydrocarbon fuels, Environmental Science: Atmospheres, 3, 1693–1740, 2023.

Lewis, J. M.: Roots of ensemble forecasting, Monthly weather review, 133, 1865–1885, 2005.

Lin, Y.-L., Farley, R. D., and Orville, H. D.: Bulk parameterization of the snow field in a cloud model, J. Appl. Meteorol., 22, 1065–1092, 1983.

Lührs, B., Linke, F., Matthes, S., Grewe, V., and Yin, F.: Climate impact mitigation potential of European air traffic in a weather situation with strong contrail formation, Aerospace, 8, 50, 2021.

Lüttmer, T., Spichtinger, P., and Seifert, A.: Investigating ice formation pathways using a novel two-moment multi-class cloud microphysics scheme, EGUsphere, 2024, 1–36, 2024.

Magnus, G.: Versuche über die Spannkräfte des Wasserdampfs, Annalen der Physik, 137, 225–247, 1844.

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M., Hanisch, T., Paul, G., Wergen, W., and Baumgardner, J.: The operational global icosahedral–hexagonal gridpoint model GME: Description and high-resolution tests, Monthly Weather Review, 130, 319–338, 2002.

Matthes, S., Grewe, V., Dahlmann, K., Frömming, C., Irvine, E., Lim, L., Linke, F., Lührs, B., Owen, B., Shine, K., et al.: A concept for multi-criteria environmental assessment of aircraft trajectories, Aerospace, 4, 42, 2017.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Matthes, S., Dietmüller, S., Dahlmann, K., Frömming, C., Peter, P., Yamashita, H., Grewe, V., Yin, F., and Castino, F.: Updated algorithmic climate change functions (aCCF) V1. 0A: evaluation with the climate-response model AirClim V2. 0, Geoscientific Model Development Discussions, 2023, 1–28, 2023.

Murray, F. W.: On the computation of saturation vapor pressure, Journal of Applied Meteorology and Climatology, 6, 203–204, 1967.

Ott, E., Hunt, B. R., Szunyogh, I., Zimin, A. V., Kostelich, E. J., Corazza, M., Kalnay, E., Patil, D., and Yorke, J. A.: A local ensemble Kalman filter for atmospheric data assimilation, Tellus A: Dynamic Meteorology and Oceanography, 56, 415–428, 2004.

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C., Gebhardt, C., Marsigli, C., Förstner, J., Z$^{\text{ï}}$angl, G., Schlemmer, L., Blahak, U., and Welzbacher, C.: DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System, Tech. rep., Deutscher Wetterdienst, 2025.

Reisner, J., Rasmussen, R. M., and Bruintjes, R. T.: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model, 124, 1071–1107, 1998.

Rutledge, S. A., Hegg, D. A., and Hobbs, P. V.: A numerical model for sulfur and nitrogen scavenging in narrow cold-frontal rainbands: 1. Model description and discussion of microphysical fields, J. Geophys. Res., 91, 14 385–14 402, 1986.

Schmidt, E.: Die Entstehung von Eisnebel aus den Auspuffgasen von Flugmotoren, Schriften der Deutschen Akademie der Luftfahrtforschung,, 5, 1–15, 1941.

Schumann, U.: On conditions for contrail formation from aircraft exhausts, Meteorologische Zeitschrift, 5, 4–23, http://dx.doi.org/10.1127/metz/5/1996/4, 1996.

Schumann, U.: A contrail cirrus prediction model, Geosci. Model Dev., 5, 543–480, 2012.

Schumann, U., Graf, K., and Mannstein, H.: Potential to reduce the climate impact of aviation by flight level changes, in: 3rd AIAA atmospheric space environments conference, p. 3376, 2011.

Seifert, A. and Beheng, K. D.: A double-moment parameterization for simulating autoconversion, accretion and selfcollection, 59-60, 265–281, 2001.

Seifert, A., Bachmann, V., Filipitsch, F., Förstner, J., Grams, C., Hoshyaripour, G. A., Quinting, J., Rohde, A., Vogel, H., Wagner, A., et al.: Aerosol-cloud-radiation interaction during Saharan dust episodes: the dusty cirrus puzzle, Atmospheric Chemistry and Physics Discussions, 2022, 1–35, 2022.

Shapiro, M., Engberg, Z., Teoh, R., Stettler, M., and Dean, T.: pycontrails: Python library for modeling aviation climate impacts, Zenodo [code], 10, 2023.

Simorgh, A., Soler, M., González-Arribas, D., Matthes, S., Grewe, V., Dietmüller, S., Baumann, S., Yamashita, H., Yin, F., Castino, F., et al.: A comprehensive survey on climate optimal aircraft trajectory planning, Aerospace, 9, 146, 2022.

Snyder, C. and Zhang, F.: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter, Monthly Weather Review, 131, 1663–1677, 2003.

Sonntag, D.: Advancements in the field of hygrometry, Meteorologische Zeitschrift (Berlin);(Germany), 3, 1994.

Spichtinger, P., Marschalik, P., and Baumgartner, M.: Impact of formulations of the homogeneous nucleation rate on ice nucleation events in cirrus, Atmospheric Chemistry and Physics, 23, 2035–2060, 2023.

Steppeler, J., Doms, G., Schättler, U., Bitzer, H., Gassmann, A., Damrath, U., and Gregoric, G.: Meso-gamma scale forecasts using the nonhydrostatic model LM, Meteorology and atmospheric Physics, 82, 75–96, 2003.

Tetens, O.: Uber einige meteorologische Begriffe, Z. geophys, 6, 297–309, 1930.

https://doi.org/10.5194/egusphere-2025-3312
Preprint. Discussion started: 28 July 2025

Thompson, G., Rasmussen, R. M., and Manning, K.: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis, Monthly Weather Review, 132, 519–542, 2004.

Thompson, G., Scholzen, C., O'Donoghue, S., Haughton, M., Jones, R. L., Durant, A., and Farrington, C.: On the fidelity of high-resolution numerical weather forecasts of contrail-favorable conditions, Atmospheric Research, 311, 107 663, 2024.

Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Mon. Weather Rev., 117, 1779–1800, 1989.

Ullrich, R., Hoose, C., Möhler, O., Niemand, M., Wagner, R., Höhler, K., Hiranuma, N., Saathoff, H., and Leisner, T.: A new ice nucleation active site parameterization for desert dust and soot, Journal of the Atmospheric Sciences, 74, 699–717, 2017.

Vaisala: Vaisala Radiosonde RS41 Measurement Performance, in: https://www.vaisala.com/sites/default/files/documents/White%20paper%20RS41%20Performance%20B211356EN-A.pdf, pp. 1–28, VAISALA, 2013.

Wang, J., Chen, J., Du, J., Zhang, Y., Xia, Y., and Deng, G.: Sensitivity of ensemble forecast verification to model bias, Monthly Weather Review, 146, 781–796, 2018.

Wang, Z., Bugliaro, L., Gierens, K., Hegglin, M. I., Rohs, S., Petzold, A., Kaufmann, S., and Voigt, C.: Machine learning for improvement of upper-tropospheric relative humidity in ERA5 weather model data, Atmospheric Chemistry and Physics, 25, 2845–2861, 2025.

Wilhelm, J., Akylas, T., Bölöni, G., Wei, J., Ribstein, B., Klein, R., and Achatz, U.: Interactions between mesoscale and submesoscale gravity waves and their efficient representation in mesoscale-resolving models, Journal of the Atmospheric Sciences, 75, 2257–2280, 2018.

Wilson, D. R. and Ballard, S. P.: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model, 125, 1607–1636, 1999.

Yamashita, H., Grewe, V., Jöckel, P., Linke, F., Schaefer, M., and Sasaki, D.: Air traffic simulation in chemistry-climate model EMAC 2.41: AirTraf 1.0, Geoscientific Model Development, 9, 3363–3392, 2016.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M.: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core, Quarterly Journal of the Royal Meteorological Society, 141, 563–579, https://doi.org/10.1002/qj.2378, 2014.