

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

Maleen Hanst¹, Carmen G. Köhler¹, Axel Seifert¹, and Linda Schlemmer¹

¹Deutscher Wetterdienst, Frankfurter Straße 135, 63067 Offenbach am Main, Germany

Correspondence: Maleen Hanst (maleen.hanst@dwd.de)

Abstract. ~~Contrails~~ Persistent contrails and contrail-induced cirrus clouds are considered the most significant non-CO₂ contributors to aviation's climate impact ~~and occur primarily~~. These clouds primarily form in ice-supersaturated regions (ISSRs). ~~Reliable prediction of~~, defined by relative humidity over ice (RH_{ice}) exceeding 100 %. Reliable prediction of 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~~, 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.

Copyright statement. TEXT

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), 2–14 % (Lee, 2018) caused by CO₂ and non-CO₂ effects.

While the uncertainty range for the climate impact of CO₂ emissions is relatively small, there is significant variability associated with non-CO₂ effects arising from emissions such as NO_x, H₂O, and, notably, the formation of persistent con-

trails and contrail-induced cirrus clouds (Matthes et al., 2017; Klöwer et al., 2021; Lührs et al., 2021; Lee et al., 2023).
 25 These ~~phenomena, collectively referred to as~~ aircraft-induced clouds ~~;~~ present a complex challenge for climate assessment
 (Teoh et al., 2024). While Kärcher (2018) ~~estimate that their contribution accounts~~ estimates that they account for more than
 half of aviation's total radiative forcing, ~~Bickel et al. (2020) contend that the~~ Bickel et al. (2025) contend that their net warming
 effect might be less than that of CO₂, primarily because it may be partially offset by a decrease in natural cirrus cloud coverage
~~;(Bickel et al., 2020).~~

30 Given the variety of findings and the potential trade-off between CO₂ and non-CO₂ impacts, effective strategies to miti-
 gate the climate impact of aviation must ~~consider~~ address both types of effects. ~~Among these strategies, climate-optimized~~
~~flight routing has gained~~ One such strategy that has gained increasing attention in recent years ~~;~~ as it seeks to minimize is
climate-optimized flight routing, which aims to reduce 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 ap-
 35 proach 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 pre-
 diction (NWP) models ~~---~~ in which relative humidity over ice (RH_{ice}) is a key parameter for evaluating contrail formation ac-
 40 cording 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~~ ice-supersaturated conditions (RH_{ice} > 100 %), which are
 essential for persistent contrail development. ~~Yet, despite its-~~

Beyond contrail modeling, ice-supersaturated regions play a critical role in ~~contrail prediction, the development and persistence~~
 45 of cirrus clouds, which are key regulators of the water vapor budget in the upper troposphere and lower stratosphere (Kärcher et al., 2023)
~~.~~ Improving the representation of supersaturation is therefore vital not only for contrail modeling but also for capturing the
broader impacts of cirrus cloud dynamics on atmospheric moisture and radiative balance (Dekoutsidis et al., 2023; Borella et al., 2025)
~~.~~ Yet, despite its relevance for climate-relevant processes, 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~~ (Kunz et al., 2014; Dyroff et al., 20
 50 .

~~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-~~

RH_{ice} ~~remains particularly difficult, even for state-of-the-art models. This is largely due to the limited availability of~~ prediction
 55 is particularly challenging due to limited upper tropospheric humidity ~~observations for data assimilation, a large variability of~~
~~humidity fields, and the~~ data, large humidity variability, and incomplete understanding of ice nucleation and cirrus cloud forma-
 tion ~~processes. Parameterizations of ice microphysical processes-~~ Improving cloud cover schemes and parameterizations of ice
microphysics are therefore an active area of research (Kärcher et al., 2022; Seifert et al., 2022; Spiehtinger et al., 2023; Achatz et al., 2024;

(Kärcher et al., 2022; Seifert et al., 2022; Spichtinger et al., 2023; Achatz et al., 2024; Grundner et al., 2024; Lüttmer et al., 2024)

60 . Additionally, predicting ice supersaturation ~~poses challenges due to resolution limitations~~ is complicated by resolution limits: NWP models ~~typically~~ represent mean atmospheric values and ~~may miss highly often miss~~ localized ice supersaturated regions (ISSRs), ~~particularly those associated with~~ especially those linked to unresolved mesoscale gravity waves (Wilhelm et al., 2018).

One way to circumvent these limitations is to ~~build a post-processing model which~~ develop machine learning methods to
65 derive RH_{ice} forecast corrections. The resulting correction model receives variables such as temperature, RH_{ice} , and others, and ~~outputs returns adjusted values of~~ RH_{ice} . Wang et al. (2025) focused their research on reanalysis data, deriving their post-processing 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~~
when validated against test data. Previous studies have also examined corrections to ERA5 reanalysis RH_{ice} , particularly in the
70 context of estimating the climate effects of aviation contrails (e.g., Teoh et al., 2022).

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~~ NWP models were 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 from~~ IFS (Integrated Forecasting System), GFS (Global Forecast System), and S-WRF (a Weather
75 Research and Forecasting model ~~configured configuration~~ by SATAVIA) ~~are evaluated and moderate scores in terms of were~~ evaluated using standard classification metrics, including the F_1 score and the Matthews Correlation Coefficient ~~were found. The study highlighted,~~ which reflect the models' ability to correctly identify ice-supersaturated conditions. Moderate scores were found, indicating room for improvement in ISSR prediction skill.

The study highlights that a correct ~~prediction of conditions which are not~~ conductive to contrail formation, mainly the
80 condition of identification of non-ISSR ~~is~~ is also crucial, as false negatives (thus, incorrect ISSR predictions) could potentially lead to unnecessary re-routing. For the S-WRF model, they ~~found find~~ 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 hence a false positive rate of ISSR 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~~
85 ~~the current status quo of aviation impacts. Conversely, the.~~ 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~~

These studies demonstrate the potential of machine learning models and state-of-the-art NWP systems to improve RH_{ice}
90 prediction, but they also reveal persistent limitations. In particular, the reliance on simplified cloud cover or microphysics schemes and deterministic forecasts restricts the ability of current models to capture the full variability and uncertainty associated with ice supersaturation.

95 ~~A key challenge in realistically representing RH_{ice} by using a two-moment cloud-ice microphysics parameterization scheme within a ten-member Ensemble Prediction System (EPS) in the global in NWP models lies in the treatment of subgrid-scale humidity variability and cloud formation processes. The IFS model addresses this through the Tompkins cloud cover scheme (Tiedtke, 1993; Tompkins et al., 2007; ECMWF, 2024), which employs a prognostic probability distribution function (PDF) of total water content to estimate cloud fraction. This statistical approach allows for a probabilistic representation of cloud cover and ice supersaturation but does not explicitly resolve the underlying ice microphysical processes.~~

~~In contrast, the ICON (ICOsahedral Nonhydrostatic) NWP model (Zängl et al., 2014) uses a physically based microphysics scheme. Within this approach, a key factor in a realistic representation of RH_{ice} is the scheme's ability to simulate the phase relaxation time – the timescale over which water vapor transitions to ice. 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, one-moment cloud ice microphysics is represented by a one-moment scheme in which parameterization, the specific ice mass is considered treated as 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, whereas ice particle number density is estimated from temperature. This approach tends to overestimate particle numbers at low temperatures, resulting in unrealistically short phase relaxation times and limiting the ability of the model to represent ice supersaturated conditions. To address these limitations, a two-moment cloud-ice scheme that includes ice microphysics scheme treats the ice particle number density as an additional prognostic variable (Köhler and Seifert, 2015). This allows ICON to better capture the phase relaxation time, and thereby the degree of ice supersaturation and the persistence of ice supersaturated regions. 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.~~

115 ~~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 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.~~

125 ~~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~~

Complementary to the model physics, ensemble forecasting is a powerful tool for capturing atmospheric variability and model uncertainty. This is particularly valuable for phenomena like ISSR, which are rare, spatially heterogeneous, and sensitive to small-scale processes. While ensemble means are commonly used to produce stable deterministic forecasts, they may obscure signals critical for ISSR detection. Instead, ensemble spread and extremes, such as the highest RH_{ice} values among members, may reveal localized supersaturation events and offer a probabilistic measure of forecast confidence.

In our study, we combine both approaches: We implement and evaluate a two-moment cloud ice microphysics scheme in the global ICON. Further, we explore its impact within a ten-member ensemble prediction system (EPS), ~~as captured by the standard deviation, may improve the reliability of ISSR classification~~ assessing how the ensemble can enhance ISSR identification beyond mean-state representation.

~~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 ~~As part of this study is to carefully analyze and interpret the EPS predictions of RH_{ice} based on ICON with the, the ten-member ICON ensemble with the new~~ 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 the German Meteorological Service (DWD). It provides~~ continuous meteorological data ~~for to support~~ research on contrail avoidance flights. This setup was developed within the D-KULT project (~~demonstrator climate and environmentally friendly air transport~~ Demonstrator for 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~~ The project seeks 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-, and operational~~ constraints such as airspace regulations and airport capacity. One ~~of the components~~ key component is the integration of ~~the ICON-2-Mom-EPS forecast forecasts from the ICON ensemble~~ 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~~ applicability of climate-aware flight ~~paths~~ operations.

~~The outline of this paper-~~

The structure of this work is as follows: In Section 2, we ~~describe the details of the dedicated ICON forecasting system; in particular, the introduce the new~~ two-moment cloud ice microphysics scheme and the ensemble setup. In Section 3, ~~an overview over~~ parameterization, the ensemble generation, and the details of the model setup. Section 3 provides an overview

of the in-situ ~~observation-measurement~~ observational data used for verification ~~is given~~. The verification methodology ~~and is~~ described in Section 4, and the results are presented ~~and analyzed~~ in Section 5, where we ~~start by evaluating the deterministic~~ evaluate the deterministic ICON 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 ~~scheme~~, particularly in comparison to the operational ICON model with the one-moment scheme. Building on the deterministic model verification, we then assess the ten-member ensemble setup. We conclude with a discussion in Section 6 and ~~a conclusion~~ final remarks in Section 8.

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.

2.1 Two-Moment Cloud Ice Microphysics Parameterization in ICON

Accurate prediction of the potential for persistent contrail formation requires a realistic representation of RH_{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 The ICON model, developed by the ICON partnership including Deutsches Klimarechenzentrum, Max Planck Institute for Meteorology, Karlsruhe Institute of Technology, the Center for Climate Systems Modeling, and the German Meteorological Service (DWD), is used operationally by DWD. In its global operational configuration, cloud ice microphysics ~~scheme is used, in which the~~ is represented by a one-moment scheme, where 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. prognostic and ice particle density is diagnosed from temperature (see Appendix A for details).

To address these limitations, a

2.1 Two-Moment Cloud Ice Microphysics Parameterization in ICON

The 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 cloud ice scheme in ICON is an extension of 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 one-moment cloud ice scheme. It adds a prognostic equation for cloud ice number density and includes explicit ice nucleation processes. 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.

195 Köhler and Seifert (2015, hereafter KS15) 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 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. ??(a), which shows the global structures of relative humidity over ice with the operational ICON (top) and the ICON with the used in this study is a simplified and updated version of KS15. The two-mode representation in KS15 is omitted for computational efficiency, as are the timestep refinements for homogeneous nucleation. In a two-moment cloud ice microphysics scheme (bottom) scheme, sources and sinks of ice particles must be explicitly parameterized. The three primary sources of ice particles are detrainment of ice from deep convective clouds, homogeneous nucleation, and heterogeneous nucleation.

Deep Moist Convection

205 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 $D_{i,conv} = 200 \mu\text{m}$ is assumed, corresponding to a mean mass of $m_{i,conv} = 10^{-9}$ 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. 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 5 in comparison with observational data

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, it neglects subgrid-scale variability, which may lead to an underestimation of nucleation events. The impact on cloud ice number concentration is 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.

Heterogeneous Ice Nucleation

220 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 $N_{dust} = 1 \times 10^3 \text{ m}^{-3}$ is assumed in the upper troposphere above $p_0 = 200$ hPa. Below that pressure height the profile increases following

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

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 $1 \mu\text{m}$ and a standard deviation of 2.5. The number of nucleated ice particles is then diagnosed as:

$$N_i^* = N_{\text{dust}} \{1 - \exp[-\bar{S}_{\text{dust}} n_S(T, S_i)]\} \quad (2)$$

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.

2.2 Ensemble Prediction System

2.2 Ensemble Generation

~~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 which are known to be sensitive. Thereby, different components of the system are perturbed, including gravity waves, convection, microphysics, the cloud scheme, turbulence and land surface. For example for convection, well-known parameters such as the entrainment rate or the excess of moisture or temperature used in the ascent of a test parcel are targeted. 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~~ with time-dependent perturbations varying sinusoidally within their range. The randomisation is accomplished by a phase shift of the sinusoidal wave depending on the ensemble member ID (for more details see Chapter 13.2 in Reinert et al., 2025). This approach introduces variability among ensemble members while preserving the consistency of individual forecast trajectories. The combined perturbation strategy ensures a

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:

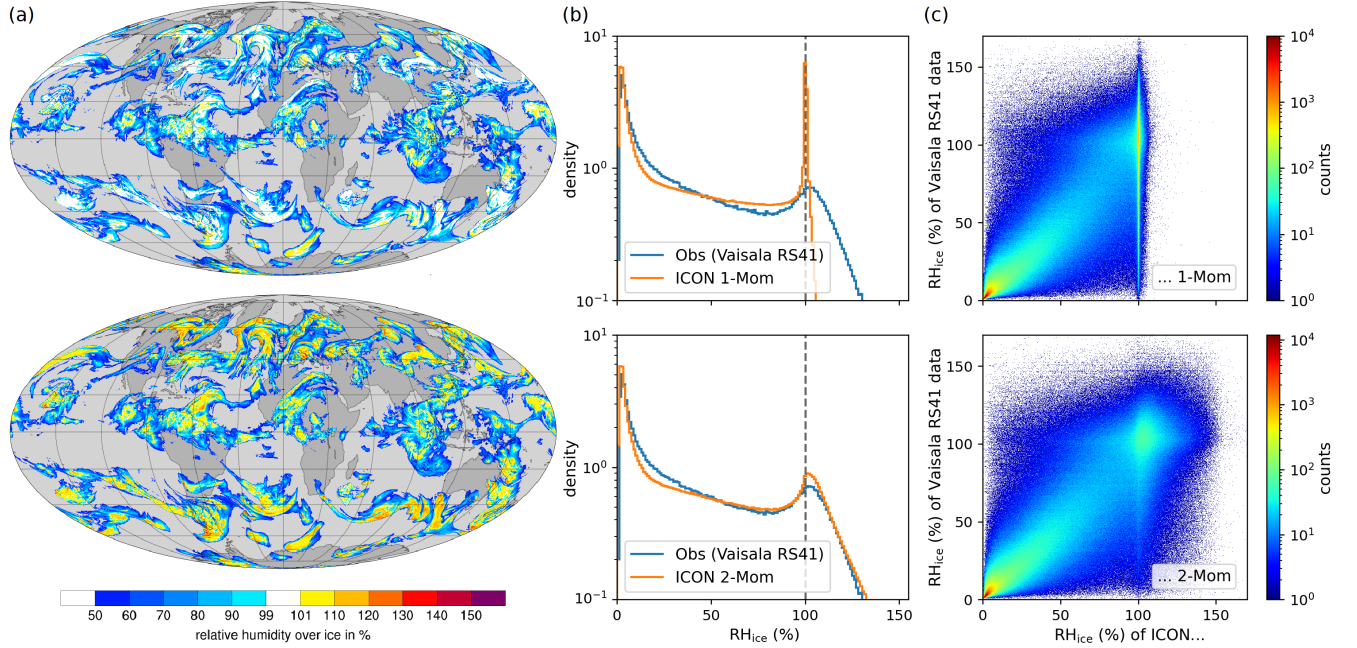


Figure 1. Relative humidity over ice (RH_{ice}) of the operational ICON with one-moment ice microphysics scheme (top row) and of ICON with two-moment ice microphysics scheme (bottom row): (a) Global forecast-only data of RH_{ice} near the tropopause (~ 10.2 km); (b) normalized histograms of RH_{ice} of Vaisala RS41 radiosonde data and ICON; (c) 2D histograms of RH_{ice} of spatio-temporally matched points between Vaisala RS41 radiosonde data and ICON forecasts with a lead time of 12 hours; heights between 8500–12500 gpm, corresponding to most common commercial flight altitudes.

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

As a third source of uncertainty, the sea-surface temperatures over oceans are perturbed in the initial conditions.

2.3 Model Setup

260 The dedicated ~~system for the D-KULT project is~~ ICON forecasting system which is implemented and evaluated in this study ~~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 ~~200–300~~ 200–300 m at the most common commercial flight altitudes of 8.5–~~12.5~~ 12.5 geopotential kilometers. It starts from the operational

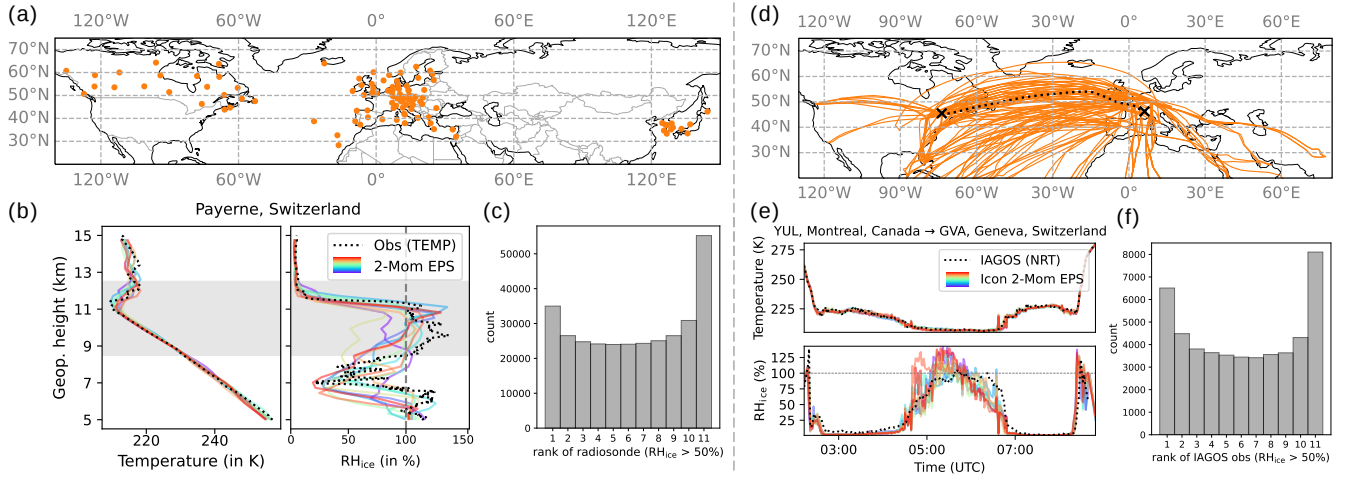


Figure 2. Radiosonde (left) and IAGOS (right) observation data. (a) Locations of 105 stations equipped with Vaisala RS41 radiosondes in the Northern Hemisphere. (b) Example height profiles of temperature and RH_{ice} from Vaisala RS41 (TEMP) observations and ICON 2-Mom EPS forecasts with a lead time 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. The rank histogram includes only samples where the observed RH_{ice} exceeds 50 %. (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).

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
 265 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* *ICON 2-Mom EPS* in the remainder of this study. Since the dedicated *ICON forecasting* system does not consist of an
 270 additional deterministic model run, we use individual members of the ensemble as approximates to a deterministic model setup for our evaluation, denoted by *ICON-2-Mom-ICON 2-Mom* in the following. Similarly, the operational *ICON with the one-moment ice microphysics scheme* is denoted by *ICON 1-Mom*.

Figure 1(a) illustrates the difference in model behavior between *ICON 1-Mom* and *ICON 2-Mom*, showing global patterns of relative humidity over ice for both schemes. While overall cloud structures remain comparable, the two-moment cloud ice
 275 scheme produces a markedly higher degree of ice supersaturation. The realism of this behavior is examined in the remainder of this study through comparison with observational data.

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.

280 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~~)([Dirksen et al., 2022; Borg et al., 2023; WMO, 2024](#)). The temperature ~~sensor utilizes stable, linear resistive platinum technology that yields a measurement is measured with an~~ accuracy of ± 0.2 °C ~~. The humidity sensor integrates humidity and temperature sensing elements and is based on capacitive polymer~~
285 ~~technology and the humidity~~ with an accuracy of ± 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 (~~[Radiosonde observations are typically conducted twice daily, with balloon 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 (UTC. The resulting data are stored in standardized binary files known as~~
290 ~~Binary Universal Form for the Representation of meteorological data) format which is a standardized binary format used (BUFR), a format developed~~ by the World Meteorological Organization (WMO) to encode and transmit various types of weather observations, ~~including radiosonde data~~ ~~. These files contain TEMP reports, which include a structured set of~~
295 ~~atmospheric measurements such as temperature, pressure, humidity, and wind speed and direction at multiple vertical levels. TEMP BUFR files serve as the standardized source of radiosonde data used in this study.~~

The ~~stored~~ Vaisala RS41 ~~radiosonde height resolution is~~ [radiosondes used in this study record vertical profiles with a height resolution of](#) approximately 1 ~~gpm, with geopotential meter (gpm) and~~ a measurement accuracy of ± 10 gpm. ~~In the TEMP files, the ± 10 gpm. Within the standardized TEMP BUFR files, dew point temperature is stored from which we derive~~
300 ~~provided, from which~~ [RH_{ice} as described is derived following the method outlined in Appendix C. In Fig. ??Figure 2\(b\) ; example radiosonde height illustrates example vertical profiles of temperature and RH_{ice} are shown together with obtained from radiosonde measurements, shown alongside](#) 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 Ob-
305 serving 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~~ [carbon](#) 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 1 s at 300 K to 120 s at 200 K, with an accuracy of ± 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. ??(d) shows the flight routes for December 2024. For an highlighted example flight route, the temperature and RH_{ice} time series are shown together with the corresponding ICON 2-Mom EPS data (Fig. ??(e)). Similar to the radiosonde verification, the analysis is confined to the Northern Hemisphere.

4 Verification Analysis Methods

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.

4.1 Spatio-Temporal Matching of Model and Observation Data

4.2 Verification of Deterministic Model ICON 2-Mom

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 ??(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 ≈ 103 %. In contrast, the two-moment scheme more accurately captures the tail structure, slightly overshooting at low supersaturation but successfully reproducing

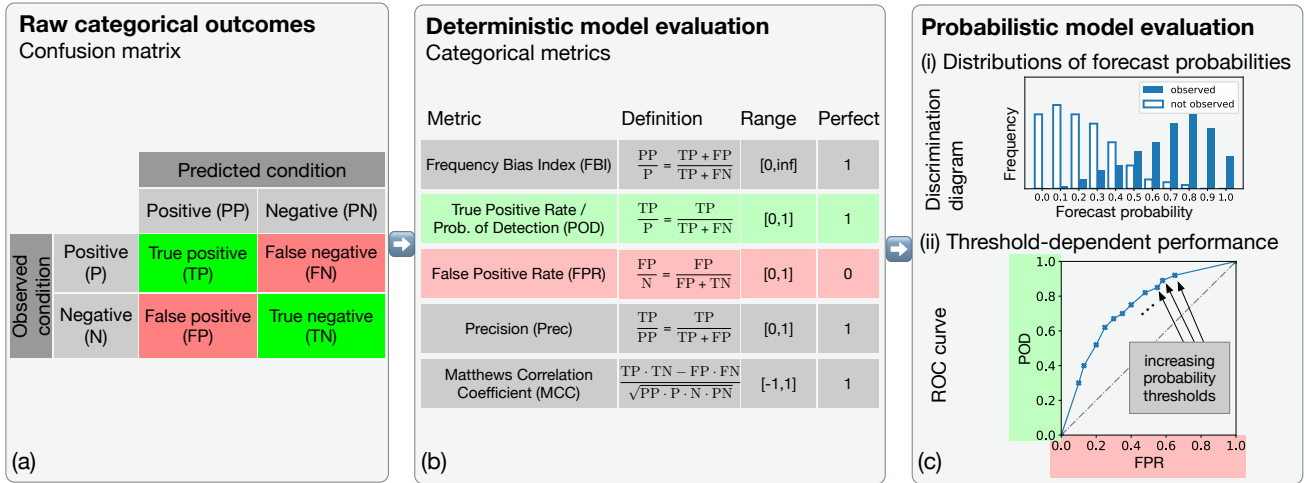


Figure 3. Radiosonde (left) and IAGOS (right) observation data Overview of categorical verification methods used in this study. (a) **Locations** Confusion Matrix: Provides a structured summary of 105 stations equipped how model predictions align with Vaisala RS41 radiosondes actual observations in the Northern Hemisphere a binary classification setting. Each prediction is categorized as a true positive (bTP) Example height profiles of temperature and RH_{ice} from Vaisala RS41, false positive (TEMPFP) observations and ICON 2-Mom EPS forecasts with a lead-time of 12 h-, false negative (eFN) Rank histogram: For each spatio-temporal point-, or true negative (comprising ICON 2-Mom EPS values and the corresponding radiosonde measurement TN), depending on its agreement with the observed value is ranked among outcome. This matrix forms the ten ensemble members, and the resulting ranks are displayed foundation for computing categorical performance metrics such as listed in a histogram(b). (db) **IAGOS flight routes** Categorical metrics: Frequency bias index (FBI), probability of 188 flights from December 2024 detection (POD), limited to false positive rate (FPR), precision, and the Northern Hemisphere Matthews correlation coefficient (MCC) offer distinct insights into model behavior as described in Section 4.2. (ec) **Spatio-temporal comparison** Categorical evaluation of flight data and ICON 2-Mom EPS the ensemble prediction system: Time-series of temperature and RH_{ice} from one example flight. (fi) **Rank histogram** the discrimination diagram shows two distributions of forecast probabilities; one for IAGOS flight data the case where the event was observed in the measurements, analogous and one where it was not observed, highlighting the model's ability to separate events by probability; (eii) the Receiver Operating Characteristic (ROC) curve illustrates the trade-off between the POD and FPR across different classification models based on the ensemble's probabilistic event forecast.

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.

4.1.1 Spatio-Temporal Comparison

The ICON grid used in our model setup has a horizontal resolution of ~~around~~ approximately 26 km and a vertical resolution of ~~200–200–300~~ m within the altitude range of ~~8500–12 500 gpm~~. ~~8500–12500 gpm~~. ~~ICON simulations with hourly forecasts up to a lead time of 60 hours were started in 6 hour intervals.~~

345 Radiosonde Data

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. ~~No horizontal interpolation was applied. However, the impact is~~
350 ~~expected to be minimal, as typical horizontal scales of ISSRs are on the order of 140 km (Spichtinger and Leschner, 2016).~~

~~For temporal matching, the start time of the accent was used as a reference, and we select the corresponding ICON simulation whose initial time is closest to the observation time minus the required lead time. Since the simulation provides hourly forecasts, this approach ensures temporal matching to the nearest hour. The exact lead time is explicitly stated in all evaluations and never below the required spin-up time of 6 hours.~~

355 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

IAGOS data represent aircraft-based observations and thus capture horizontal trajectories spanning several hours. Matched
360 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. For~~
~~temporal matching, the minimum lead time was fixed at 6 hours to account for the required ICON spin-up. Since flights span several hours, different ICON simulations were used, each selected based on the initial time closest to the observation time~~
365 ~~minus the 6-hour lead time. As ICON simulations are initialized in 6-hour intervals, this approach may result in a maximum temporal mismatch of ± 3 hours.~~

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 a sample~~ time series of temperature and RH_{ice} from an intercontinental flight, together with the corresponding ICON ensemble values~~from the nearest model grid cell.~~

370 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. ??(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
 375 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.

4.2 Categorical Metrics

380 4.2.1 **Categorical Scores of ICON 2-Mom**

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
 385 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)
 390 together (missing values are due to vanishing denominators); (f) Vaisala RS41 radiosonde observation measurements.

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~~: complementary events:

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

~~In this study, we are particularly interested in ice supersaturation~~

395 In addition to the duration of ISSRs, pronounced ice supersaturation has been associated with the persistence of contrails (Teoh et al., 2022). While this link is relatively weak, relative humidity remains the dominant factor in contrail-cirrus evolution, governing both the total ice mass and total extinction (Unterstrasser and Gierens, 2010). Given its relevance, this study focuses on ice supersaturation events ($RH_{ice} > 100\%$) and ~~events of higher ice~~ on cases of pronounced supersaturation ($RH_{ice} \gg 100\%$). These are treated as the positive events in our categorical verification framework, which we particularly aim to
 400 distinguish from their complementary cases. The spatio-temporal matching points between model output and observational data ~~are categorized~~ can then be categorized with respect to the positive event. Positive predictions and negative predictions

are classified with respect to the observed condition as true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). The results are indicated in a confusion matrix (see Figure 3(a)), which serves as the foundation basis for computing categorical scores (see Table ??) metrics (see Figure 3(b)). Below we provide all metrics which we later use to evaluate the performance of ICON 2-Mom (EPS). $RH_{ice} > \text{threshold}$ positive-prediction-negative-prediction-observation-true-positive (TP)-false-negative (FN)-negative-observation-false-positive (FP)-true-negative (TN)-Confusion-matrix: Categorization of predicted events (positive predictions) and predicted non-events (negative predictions) in relation to the actual-observed situation. In the remainder of this study, we consider eventsof the type-

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

The frequency bias index (FBI) is defined as the ratio of the forecast frequency of an event to its observed frequency

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

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. It indicates whether the forecast system tends to overforecast ($FBI > 1$) or underforecast ($FBI < 1$) a given event.

The Probability of Detection (POD, also known as sensitivity) evaluates the forecast system's ability to correctly identify observed events. POD is defined as the ratio of correctly predicted events to the total number of observed events, given by

$$POD = \frac{TP}{TP + FN}.$$

The false positive rate (FPR, also defined as $1 - \text{specificity}$) quantifies the proportion of non-events that were incorrectly forecast as events. It is defined as

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

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 events: non-events. To complement these metrics, precision provides a forecast-centric perspective, highlighting the trustworthiness of predicted events and if defined by

$$FBI_{Prec} = \frac{TP + FP}{TP + FN} \frac{TP}{TP + FP}.$$

The results are shown in Figure ??(a)

The Matthews correlation coefficient (MCC) is a composite measure that accounts for all four components of the confusion matrix simultaneously. MCC is particularly well-suited for datasets with class imbalance (in our case we have about 13% ISSR events), as it reflects the quality of binary classifications regardless of event prevalence. It is defined as

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

430 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.

4.3 Categorical Verification of Probabilistic Model

Ensemble forecasts provide a distribution of values for any forecast quantity of interest. For binary events such as ice supersaturation, the forecast probability is defined as the fraction of ensemble members predicting the event.

435 Discrimination diagram

The discrimination diagram visualizes two conditional distributions of the forecast probabilities: one conditioned on the event being observed in the measurement data, and the other conditioned on the event not being observed.

To assess the discriminative capability of the EPS, we employ the discrimination diagram, which visualizes two conditional distributions of the forecast probabilities: one conditioned on the event being observed in the measurement data, and the other
440 conditioned on the event not being observed.

These distributions are represented as normalized histograms of the EPS forecast probabilities. A clear separation between the two distributions indicates strong discriminability, reflecting the ensemble's ability to assign higher probabilities to observed events and lower probabilities to observed non-events. This method provides a threshold-independent diagnostic of classification performance in a probabilistic forecasting framework. An example sketch of a discrimination diagram is provided in Figure
445 3(c)(i).

Receiver Operating Characteristic (ROC) Curve

The ROC curve is a powerful threshold-dependent verification tool to evaluate the performance of a binary classification model. Such a model typically predicts not just a binary label directly, but rather a scalar score (in our context, this score is the predicted event probability). A score is turned into an event prediction if it is above a certain threshold. The threshold itself becomes
450 part of the model; by varying the threshold, we effectively obtain a multitude of models, each with its own POD and FPR. The ROC shows the POD versus the FPR for all of these models at once. The top-left corner corresponds to a perfect classification model. An example sketch of a ROC curve is provided in Figure 3(c)(ii).

5 Verification Results

We evaluate the RH_{ice} predictions of ICON equipped with the new two-moment ice microphysics scheme in two steps. First,
455 we verify the deterministic model, ICON 2-Mom, which includes a comparison with ICON 1-Mom. Second, we evaluate the ensemble prediction system, ICON 2-Mom EPS.

Radiosonde data were used unless the use of IAGOS data is indicated. Only data within the 8.5–12.5 km geopotential height range were included to match commercial flight altitudes.

5.1 Verification of Deterministic Model ICON 2-Mom

460 5.1.1 Relative Frequency Distribution of RH_{ice}

Figure 1(b) displays the relative frequency distributions of the observed RH_{ice} compared to the corresponding model-based distributions from the operational ICON 1-Mom (top) and the new ICON 2-Mom (bottom) configurations. Pronounced differences emerge in the tail of the density distribution, which reflects ice supersaturation. ICON 1-Mom exhibits a sharp peak near 100 %, followed by a rapid decline, with maximum RH_{ice} values reaching only ≈ 103 %. In contrast, ICON 2-Mom 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.

5.1.2 Continuous Spatio-Temporal Comparison

We examined the 2D histograms of RH_{ice} of spatio-temporally matched points between Vaisala RS41 radiosonde data and ICON forecasts (Fig. 1(c)). While ICON 2-Mom reproduces the observed supersaturation range reasonably well – unlike ICON 1-Mom – noticeable scatter remains around the one-to-one line. However, perfect agreement between modeled and observed RH_{ice} values is not strictly required in our context. Crucially, the model must reliably distinguish between ISSR events and non-events, as both have significant operational implications for flight planning and routing. To assess this capability, we proceed below with a verification based on categorical performance metrics.

475 5.1.3 Categorical Verification

In the remainder of this study, we consider events of the type

$RH_{ice} > \text{threshold}$,

with $\text{threshold} \in \{100\%, 105\%, 110\%, 120\%\}$.

Figure 4(a) compares the FBI between ICON 1-Mom and ICON 2-Mom for these events. 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 RH_{ice} thresholds, the FBI for ICON 2-Mom is slightly below 1 for at 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 – which consistently underestimates event occurrence – but also tends to slightly overestimate observed event frequency. Meanwhile, the one-moment scheme consistently underestimates event occurrence.

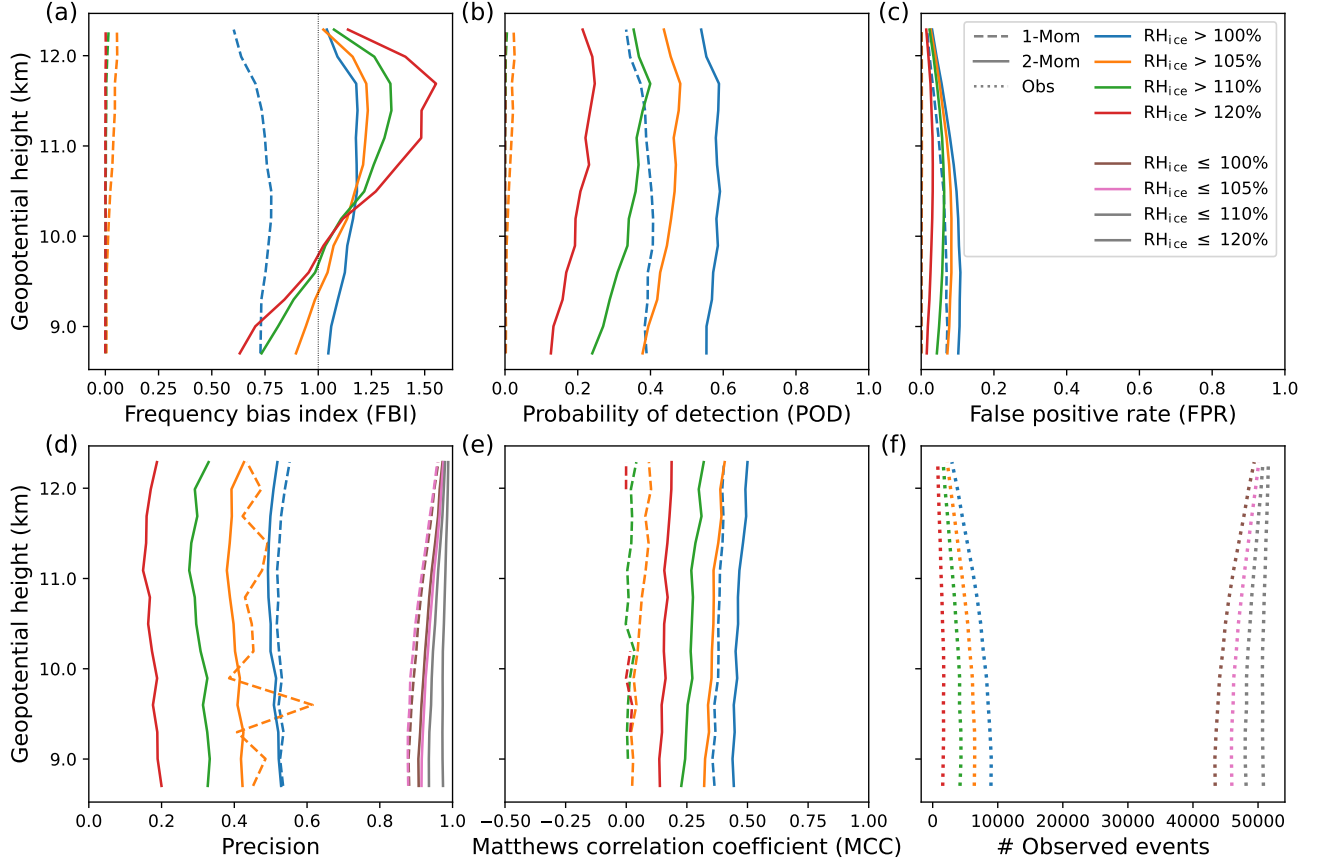


Figure 4. Categorical verification of ICON 1-Mom and ICON 2-Mom against Vaisala RS41 radiosonde measurements. The analysis covers data from the Northern Hemisphere within the most frequently flown altitude range of 8.5–12.5 km geopotential height, over a verification period of 11.5 months (June 15, 2024 – May 31, 2025). Forecasts are initialized at 00 and 12 UTC with a lead time of 12 h. Observational profiles are linearly interpolated to ICON model levels (~ 13 levels within the target altitude range), yielding approximately 680 000 samples, with ice supersaturation present in $\sim 13\%$ of cases. Panels show categorical scores for ice supersaturation events: (a) FBI; (b) POD; (c) FPR; (d) precision; (e) MCC; (f) Number of Vaisala RS41 radiosonde RH_{ice} event observations.

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

$$490 \quad \text{POD} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

we find that, The POD for ISSR events, the POD increases from about (RH_{ice} > 100 %) increases from approximately 0.4 for ICON 1-Mom to about around 0.6 for ICON 2-Mom and is almost constant over, remaining nearly constant across the altitude range in both cases configurations. For events defined by higher RH_{ice} thresholds, the two-moment scheme ICON 2-Mom retains some detection capability, with the POD values gradually decreasing to about 0.15–0.2 for RH_{ice} > 120 %. In contrast, as also indicated by the FBI, the one-moment scheme consistent with the FBI results, ICON 1-Mom fails to detect RH_{ice} values above 105 %, yielding POD values near zero across the altitude range throughout the vertical domain.

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:-

$$500 \quad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}.$$

The false positive rate is relatively low in The FPR remains relatively low across all cases, reaching a maximum of peaking slightly above 0.1 for the two-moment scheme and RH_{ice} > 100 ICON 2-Mom at RH_{ice} > 100 % (Fig. ??4(c)), indicating a limited tendency toward false alarms.

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:-

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

For the ISSR event, both Both schemes yield similar precision values between 0.5 and 0.55 for ISSR events across the entire altitude range (see Fig. ??Fig. 4(d)). For events with higher RH_{ice} thresholds, the precision of ICON 2-Mom decreases successively declines progressively, reaching values as low as 0.2 for RH_{ice} > 120 %. In contrast, ICON 1-Mom yields produces very few or even no positive predictions in these regimes, making rendering precision largely undefined; accordingly, it is omitted for these cases.

In the context of flight planning, accurate prediction of non-ISSR conditions is also equally critical, as false negatives in this category can lead to unnecessary re-routing and, consequently, avoidable increases in CO₂ emissions. When considering treating the complementary events (RH_{ice} ≤ threshold) as "positive" "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 underscores the reliability of ICON 2-Mom is quite reliable in detecting in correctly identifying non-ISSR conditions.

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:-

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

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 ?? The MCC shown in Fig. 4(e) summarizes overall classification performance. For ISSR/non-ISSR classification (blue curves), ICON 2-Mom achieves an average MCC of 0.47 across all altitudes, while, indicating moderate predictive skill. In comparison, ICON 1-Mom yields consistently lower values between 0.38 and 0.39 . At higher RH_{ice} thresholds, the MCC for of ICON 2-Mom decreases successively declines progressively, reaching a minimum of approximately 0.16 . In contrast, the MCC MCC values for ICON 1-Mom approaches zero (or is approach zero or become undefined where the numerator vanishes), indicating no predictive skill, reflecting a lack of predictive capability in these regimes.

In summary, for ISSR events, ICON 2-Mom achieves a moderate MCC of nearly 0.5 and a POD that is approximately 50% higher POD compared to than that of the operational ICON 1-Mom, while maintaining a relatively low false positive rate FPR below 0.1 at across most altitudes. Nevertheless Despite this improvement, a POD of 0.6 suggests that further improvements are possible, and indicates that a substantial fraction of events remains undetected. To address this, we continue to explore investigate potential gains from the ensemble setup introduced in Section 2.2.

5.2 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- We begin 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. ?? (evaluation of the ensemble's ability to represent RH_{ice} variability, using the rank histogram as a diagnostic tool. The rank histogram is constructed by ranking the observed value relative to the ten sorted ensemble forecasts and recording its position across all spatio-temporal matching samples.

Fig. 2(c) shows the resulting histogram for the subset of samples where the observed RH_{ice} is above 50% . We consider this restricted rank histogram because ICON tends to underestimate very low humidity values, which are not the subject of this study but would obscure the relevant behavior (also reflected by the RH_{ice} histogram in Fig. 1(b, bottom)). This underdispersion can, in part, be attributed to the inherent The histogram exhibits a U-shape, indicating underdispersion, i.e., the ensemble fails to capture the full variability present in the observations. This behavior is partly due to spatial averaging over model grid cells, which tends to smooth out extremes. From a physical modeling perspective, key contributing factors may include the absence However, counteracting this, so-called upscaling effects of the model tend to display small-scale physical behavior on the

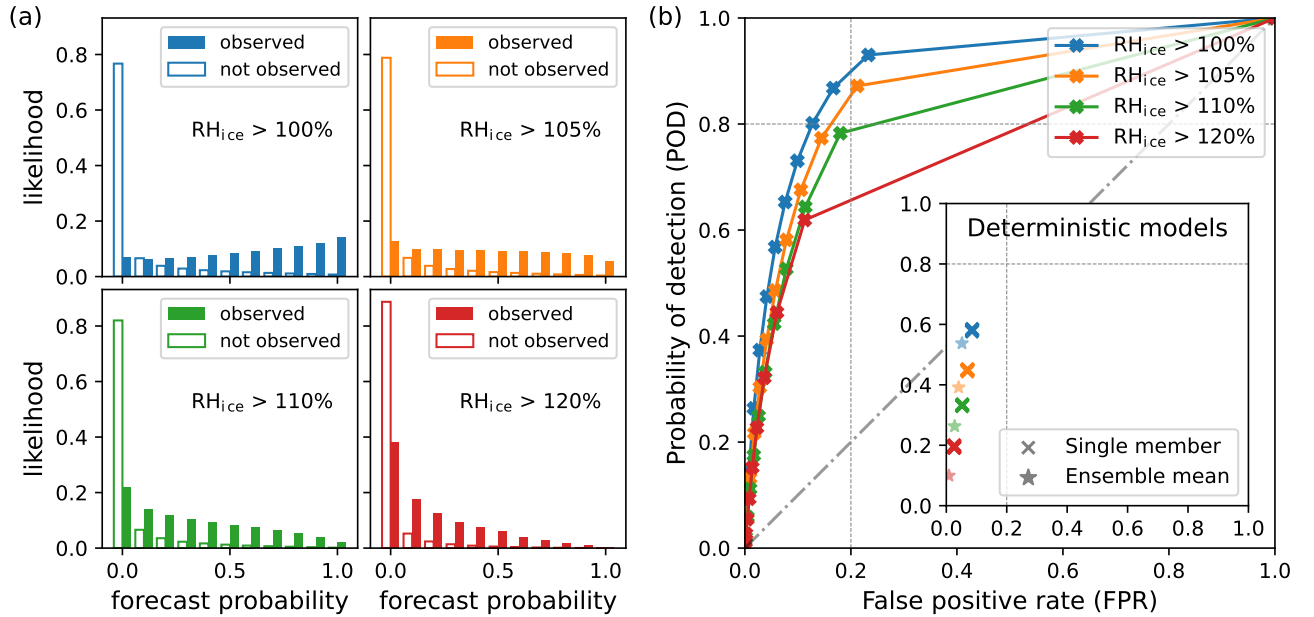


Figure 5. Ensemble verification metrics targeting illustrating the ability discriminatory skill of the model to discriminate ICON 2-Mom EPS in distinguishing between events and non-events (e.g., ISSR and non-ISSR, shown in blue). The verification period spans 14 months (April 2024 – May 2025), leading to ~820 000 yielding approximately 820 000 samples. (a) Discrimination diagram: Conditional distributions of EPS forecast probabilities; conditioned on, given that the event was actually observed and not observed in the measurement data and conditioned on that it was not observed. (b) Receiver Operating Characteristics Characteristic (ROC) curve: Probability of detection POD versus false positive rate of FPR for ice supersaturation events for varying RH_{ice} , evaluated across a range of threshold-based decision models (pseudo-deterministic models received derived from the EPS. These pseudo-deterministic models are constructed by applying various varying probability threshold conversions) thresholds to the ensemble output.

model scale. Thus, insufficient parameter perturbations may be another reason, together with the lack of subgrid-scale gravity waves in the model configuration and the use of climatologically prescribed aerosol fields from climatology, both of which limit constrain variability in ice nucleation conditions. We also observe a more pronounced negative bias within

Moreover, the rank histogram, indicating that the model tends to underestimate reveals a slight negative bias, with observed RH_{ice} more often than it overestimates values more often exceeding the ensemble forecast range than falling below it. This suggests a systematic underestimation of RH_{ice} . Thus, further post-processing of the EPS model forecasts may be useful for predicting RH_{ice} and, in particular, for identifying ISSR or higher ice supersaturation by the model, at least in parts of the $RH_{ice} > 50\%$ regime. We found that this mainly occurs at ice supersaturated conditions. However, the rank histogram does not provide any information about magnitudes. Thus, we further analyze the ensemble's ability to classify ISSR and non-ISSR conditions below.

5.2.1 Prediction of ISSR and Higher Ice Supersaturation/non-ISSR Discrimination Ability

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 ($RH_{ice} > 100\%$) and higher supersaturation ($RH_{ice} \gg 100\%$). The ensemble inherently provides probabilistic forecasts for these events via the proportion of members with the corresponding event. We start by considering two, we consider the discrimination diagram introduced in Section 4.3. Figure 5(a) shows the 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. ??(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 forecast probabilities for observed and non-observed events (events are defined as $RH_{ice} > 100\%$, and higher thresholds). For ISSR events, the "not observed" distribution peaks sharply at zero and declines rapidly, indicating strong agreement among ensemble members when no 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. "observed" distribution is relatively uniform, suggesting that the ensemble assigns a broad range of probabilities to actual events. As the threshold for supersaturation increases, the "observed" distribution becomes more left-skewed and increasingly overlaps with the "not observed" distribution, indicating a decline in discriminative skill for more extreme events.

Focusing on To conclude, for the ISSR event, the corresponding discrimination diagram shows that the diagram shows little overlap between the two conditional distributions becomes small for forecast probabilities above approximately forecast probability distributions below and above ~ 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, suggesting that a threshold-based conversion yields what we term a "pseudo-deterministic" model. Throughout this study, we conversion of forecast probabilities aimed at classifying ISSR versus non-ISSR may be appropriate.

5.2.2 Threshold-Dependent Performance

As ICON 2-Mom EPS consists of ten ensemble members, ISSR forecast probabilities can be 0, 0.1, 0.2, ..., 1. Thus, these values represent the relevant potential thresholds to turn the event forecast probability into an event prediction – yielding

595 classification models as introduced in section 4.3. We refer to these classification models as decision models, characterized
by their underlying conversion thresholds. Specifically, the k-out-of-10 decision model defines with the k-out-of-10 decision
model (or simply decision model k) defining 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:-

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

:-

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

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)
605 curve, which plots the POD versus the FPR of all potential decision models. The construction of To evaluate the performance
of these pseudo-deterministic decision models, we use 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 (??Section 4.3). 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
610 curve in Fig. ??ISSR event, the ROC curve (Fig. 5(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) shows strong discriminative skill
for thresholds of 0.2 or and 0.3 are applied (resulting in decision model (decision models 2 or and 3), the POD is greater than
with POD > 0.8 while the FPR remains less than and FPR < 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
615 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:-

Youden Index = POD - FPR,

by maximizing it across all possible conversion thresholds (decision models). The range of possible outcomes is from -1 to
620 +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. ??(b) are summarized in Table
??.

Youden Index (YI) for the ISSR event and for events with higher ice supersaturation;

$RH_{ice} = \max_{\text{Youden Index}} \text{Index decision model}$

threshold EPS(Det)

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.

Comparing the scores of potential decision models based on the EPS, as shown in the ROC curve in Fig. ?? A comparison between the scores of the EPS-based decision models and the deterministic ICON 2-Mom model (inset of Fig. 5(b), with the results of the deterministic ICON 2-Mom model (inset), reveals that the EPS can significantly increase the) shows a substantial improvement in the POD for ISSR detection from events, from approximately 0.6 in the deterministic case to 0.8-0.9. This improvement comes with a moderate rise 0.9 when using ensemble-based decision models. This gain in POD is accompanied by a moderate increase in the FPR from approximately , rising from ~0.1 to values between 0.13 -0.23 and 0.23, depending on the decision model employed. Notably, the Youden index also improves substantially (see Table ??) chosen threshold. These results highlight the added value of ensemble forecasts in enhancing event detection or classification.

To complement this view, and following the discussion in Section 5.1.3,

While the ROC curve provides a comprehensive view of classification performance across thresholds, it treats both classes equally and may obscure performance nuances in the presence of class imbalance. Therefore, we also evaluate the precision-recall curve offers an alternative perspective that (PR) curve, which focuses specifically on the model's performance for 's performance on the positive class in the context of an unbalanced dataset. 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 ?? Similarly to the construction of the ROC curve, the PR curve plots the recall (equivalent to POD) against precision. In Figure 6(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 each EPS-based decision model is represented as a point on the PR curve. The closer the points are to the top right corner, the better the trade-off between higher the recall and precision. Overall, the Although recall remains high even for intermediate thresholds, overall precision is only moderate and gets worse for higher ice deteriorates further for more extreme supersaturation events. In Figure ??(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. This reflects the increasing difficulty of making accurate positive predictions as the event definition becomes more stringent.

655 As the precision-recall curve and the F_1 score are not symmetric with respect to what we define as the "positive" event, We also conduct a complementary analysis by treating the complementary conditions (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 CO₂ 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
660 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 events, as this perspective is equally relevant for flight routing applications. The PR curve approaches the top right corner, reflecting both high POD and precision, and a zoomed-in view providing details
665 of this region is shown in Appendix D Fig. D1(a).

We

To conclude this subsection ~~by shifting the focus from model performance on specifically defined positive events,~~ we shift to a more holistic ~~evaluation using the Matthews Correlation Coefficient for the EPS-based decision models. As discussed in Section 5.1.3, the model assessment using the MCC, as introduced in Section 4.2. The~~ MCC provides a balanced ~~assessment of model skill for measure of classification skill across~~ both event and non-event ~~classifications, similar to the ROC curve, and is particularly informative categories, making it particularly valuable~~ in the context of imbalanced datasets. In the case of
670 of model skill for measure of classification skill across both event and non-event classifications, similar to the ROC curve, and is particularly informative categories, making it particularly valuable in the context of imbalanced datasets. In the case of

For the ISSR/non-ISSR classification, ~~decision models 1-7 achieve higher MCC values than EPS-based decision models 1-7 consistently outperform~~ their deterministic counterparts (i.e., individual ensemble members), with decision models 3 and 4 ~~reaching a maximum MCC of achieving the highest MCC values of approximately~~ 0.55. ~~By~~ In contrast, the ~~MCC values for the deterministic models remain~~ deterministic models yield MCC scores around 0.47 (see Fig. ??6(c)). These results reinforce the advantage of ensemble-based decision strategies in capturing both sides of the classification task more effectively.
675 the deterministic models remain deterministic models yield MCC scores around 0.47 (see Fig. ??6(c)). These results reinforce the advantage of ensemble-based decision strategies in capturing both sides of the classification task more effectively.

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

Table 1. 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 1 ~~shows~~ summarizes the maximum MCC ~~for each RH_{ice} threshold~~ event values achieved for each ice supersaturation threshold, along with the indices of the corresponding EPS-based decision models. For ~~comparison with the ROC results, the reference, MCC values of the deterministic (single-member) models are shown in brackets.~~ The final columns report the associated POD and FPR values ~~of these models are also provided~~, enabling direct comparison with ROC-based performance. In most cases, the decision models with highest MCC also show favorable POD–FPR combinations, underscoring their robustness across metrics. For the remainder of this study, we focus on ROC-based evaluation using its associated scores, POD and FPR, as a representative framework for assessing decision model performance.

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~~

5.2.3 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. 7). Compared to the ICON 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, up to $RH_{ice} > 120$ %, the shape of the ROC curves (see Fig. 7) derived from the IAGOS data closely resembles those derived from the radiosonde data (compare Fig. 5(b)). These findings strengthen our verification insights across different, independent observation systems.

5.2.4 Longer Forecast Lead Times

~~So far we have focused on ICON data with a forecast lead time of 12 hours. For~~ Although for many flights 12 hour forecasts are sufficient. ~~However, in general, longer forecasts should be provided. Therefore, we considered~~, we now consider 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 ??)~~. As, Figure 8) shows that – as the lead time increases, ~~–~~ the ROC curves shift slightly to the right, indicating higher ~~false positive rates~~ FPR. 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 and 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 $FPR < 0.2$.

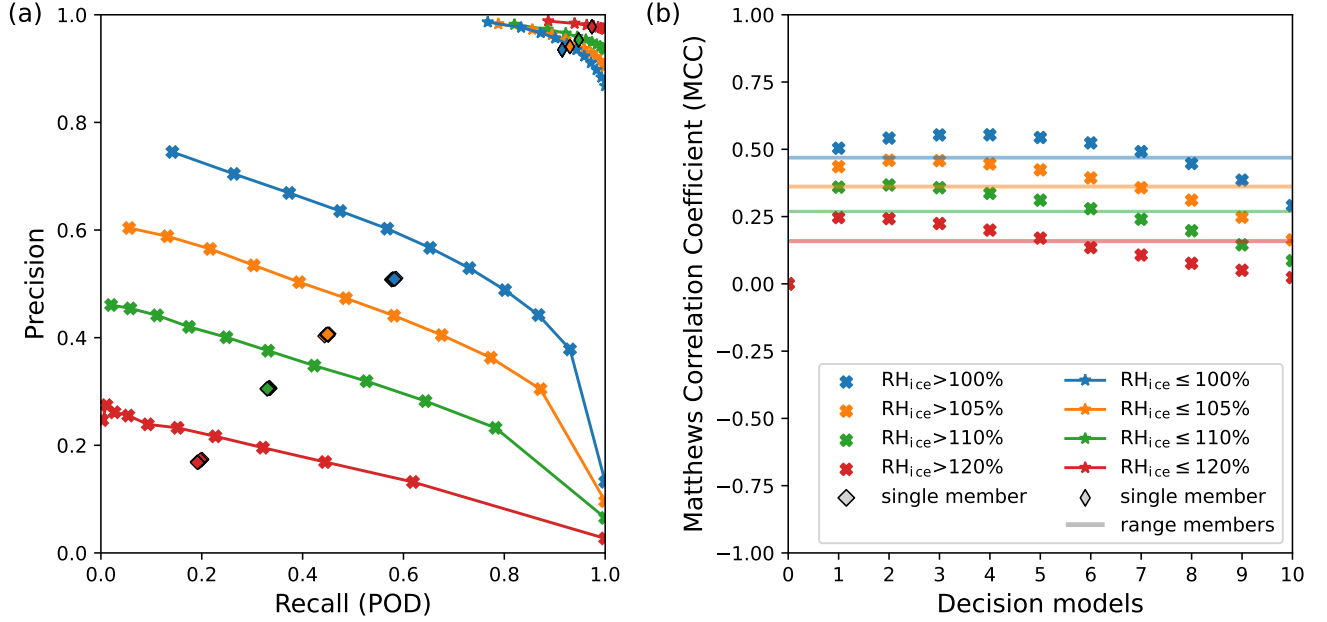


Figure 6. Scores that take into account the unbalanced dataset with respect to the ISSR event or higher ice supersaturation events in two different ways: The precision-recall (PR) curve and the F_1 -score by focusing on the performance of the model with respect to what is defined as the 'positive' event, and the Matthews correlation coefficient (MCC) by providing a balanced evaluation measure with respect to all four categories of the confusion matrix. (a) Precision-recall PR 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 ($RH_{ice} > \text{threshold}$) (bold crosses) or ($RH_{ice} \leq \text{threshold}$) (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 5.2.1). 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 ($RH_{ice} > \text{threshold}$) and ($RH_{ice} \leq \text{threshold}$) 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 (transparent lines).

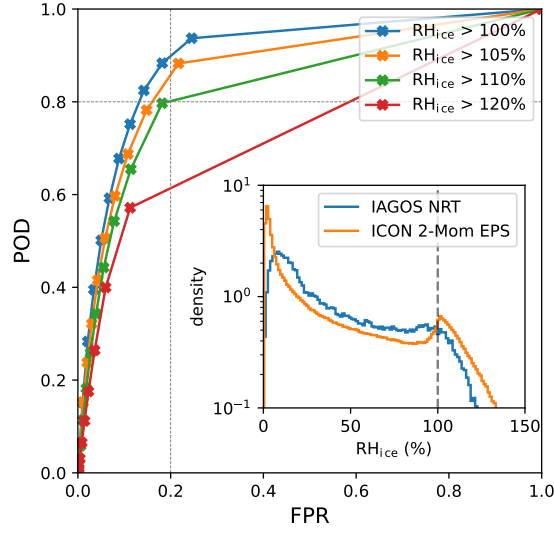


Figure 7. ROC curve of ICON 2-Mom EPS and IAGOS data, the inset figure shows the corresponding RH_{ice} densities. Evaluation performed with 625 flights from four months (August 2024, October 2024, December 2024, January 2025) on the Northern Hemisphere, leading to $\sim 200\,000$ spatio-temporal samples.

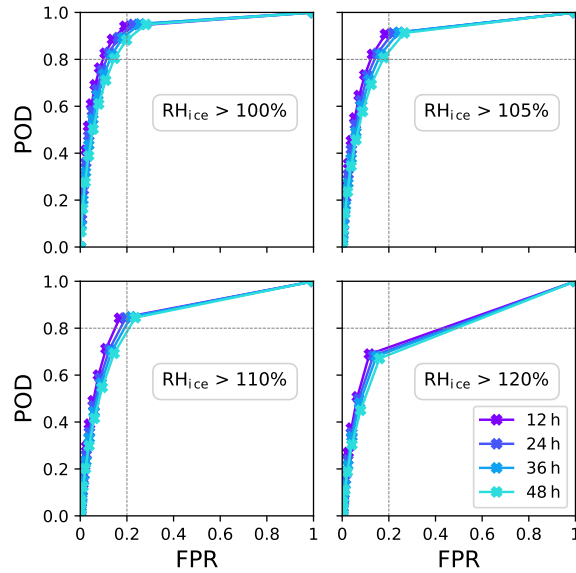


Figure 8. ROC curves for increasing forecast lead times and increasing RH_{ice} thresholds; time period five months: 1.1.2025–~~31.5~~–31.5.2025; ICON initial times 0 UTC and 12 UTC; Northern Hemisphere.

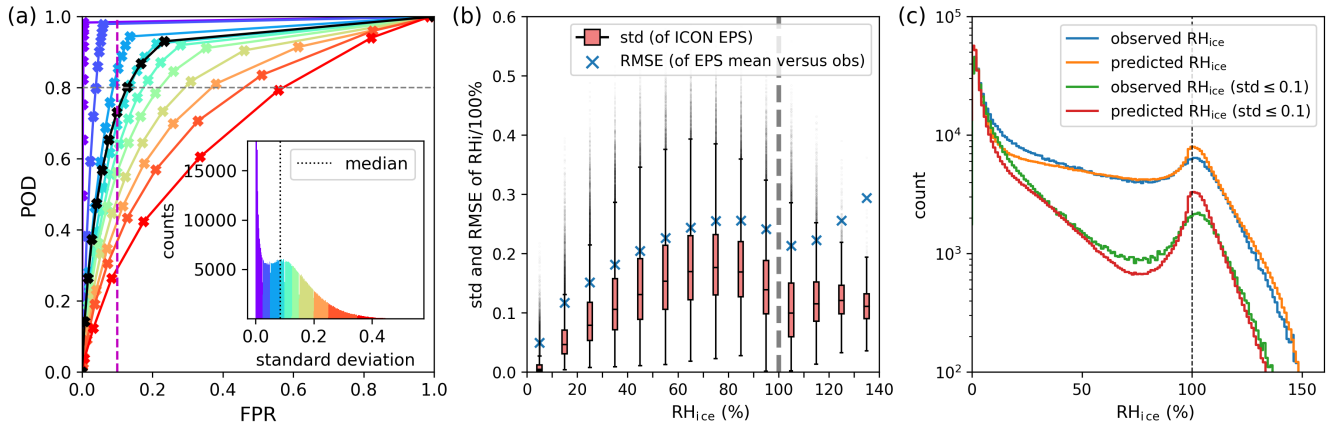


Figure 9. Event $RH_{ice} > 100\%$: Inclusion of the ensemble spread of RH_{ice} , measured by the standard deviation (std) of $RH_{ice}/100\%$. (a) ROC stratification along the curves on sample subsets grouped and color-coded by their standard deviation (std) values. The inset shows the histogram of the standard deviation std of $RH_{ice}/100\%$ and, which also serves as a legend for the ROCs on ROC curves corresponding to EPS subsets with associated std. The black ROC is represents the original one without std stratification curve based on all samples. (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 $1.5 \times IQR$. All data points outside $1.5 \times IQR$ 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 $std \leq 0.1$ 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.

710 5.2.5 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 $\{RH_{ice} > \text{threshold}\}$ through appropriate interpretation of the EPS (via decision models). Here, we We further incorporate ensemble spread information in order to get more reliable scores in more specific situations. In general, the The ensemble spread should be an indication of the confidence in a forecast and is typically measured by the standard deviation (std). Therefore, in the context of RH_{ice} -ISSR forecasts, we further stratify differentiate the ROC curve in terms of the underlying ensemble spread based on the underlying std at each grid point, particularly to achieve a lower FPR. Ensemble spread is typically measured by the standard deviation.

The inset of Figure ??9(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 ??9(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 versa, 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,~~ particular, a significantly improved ROC shape is obtained in more than half of the cases ~~a significantly improved ROC is~~ obtained with POD between – with POD of 0.9 and 1 and ~1 and FPR ≤ 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 ≤ 0.1 ~~(indicated by the vertical magenta line)~~. In these cases ~~– depending on the specific std –~~ only a lower POD of 0.3–0.8 can be obtained, between 0.8 and 0.3, depending on the underlying std.

~~As the shape of the ROC curves varies significantly along~~ Given the significant variation in ROC curve shapes across different std regimes, we ~~were also interested in analyze~~ the std values of different RH_{ice} regimes, particularly when RH_{ice} is around or above 100 %. In Fig. ~~??~~9(b), summary statistics of std are shown for increasing 10 % bins of RH_{ice}. Following an increase in std values, ~~they the bins~~ fall before 100 % and reach another local minimum in the RH_{ice} regime of 100–110–100–110 % with a median around 0.1. The relative mean squared error (RMSE) shows a similar qualitative behavior ~~for up to~~ for up to RH_{ice} < 120 %. For higher RH_{ice} regimes, the RMSE increases to its maximum over the whole ~~RH_{ice} value range~~ range of values.

~~We take another perspective in Fig. ??~~ In Fig. 9(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~~ spatio-temporally matched EPS 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~~ regime 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~~ std, the histogram still displays intermediate supersaturation. This suggests that certain ISSRs can be well predicted.

The ~~results shown in Figures ??(b) and ??(c) are similar to the findings of Borella et al. (2024) who parameterized the subgrid-scale distribution of water vapour 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} ≈ 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 mo-

tion 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 vapour, 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.

5.2.6 Comparison with IAGOS Data

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 $\sim 200\,000$ spatio-temporal samples.

6 Discussion

The

6.1 Observed Standard Deviation of RH_{ice}

The results shown in Figures 9(b) and 9(c) share notable similarities with the findings of Borella et al. (2024), who parameterized the subgrid-scale distribution of water vapor in the UTLS using IAGOS data. They observed a predominantly quadratic relationship between the standard deviation of 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 and its mean, with a peak occurring between 70 % and approximately 110 %, depending on temperature. Beyond this range, the standard deviation exhibited an upward trend at even higher RH_{ice} density (see inset of Fig. ??). 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. Their temperature-dependent analysis further revealed that this peak decreases in magnitude and shifts toward higher 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. ??). temperature decreases. While our approach to grouping ROC curves by ensemble spread does not currently account for temperature, incorporating it may be a valuable direction for future work.

7 Summary and Discussion

6.1 Comparing Microphysics-Based and Statistical Approaches to Ice Supersaturation

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. The results of our study demonstrate that the two-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.

6.2 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 implemented in ICON provides a microphysically based alternative to prognostic cloud cover schemes – such as the Tompkins scheme used in the IFS model (Tompkins et al., 2007) – that infer ice supersaturation from subgrid-scale humidity distributions. The Tompkins approach offers some advantages for operational weather forecasting due to its computational efficiency and its ability to represent subgrid-scale humidity variability. This can be advantageous for realistic cloud fraction estimates on coarse grids. However, this scheme does not explicitly prognose specific ice mass or ice particle number density, and phase relaxation time is effectively zero because the scheme assumes instantaneous in-cloud equilibrium. Indeed, the current cloud scheme of IFS assumes ice supersaturation only in the cloud-free portion of the grid box (ECMWF, 2024), which can lead to an underestimation or smoothing of ice supersaturation under certain conditions. In contrast, ICON 2-Mom compared to ICON-1-Mom, prognoses both specific ice mass and ice particle number density, allowing phase relaxation time to emerge naturally from microphysical relationships. This enables a more direct, microphysics-based simulation of the onset and persistence of ice supersaturation, which is particularly relevant for applications requiring detailed RHi forecasts, such as contrail avoidance. While this approach offers improved physical realism and consistency, it comes with increased computational cost and sensitivity to assumptions about nucleation and particle size distributions. Consequently, careful tuning and validation are necessary, especially in global applications.

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 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.

6.2 Model Resolution and Neighborhood Consideration

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.

6.3 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.

6.3 Model Resolution and Neighborhood 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 $RH_{ice} > 99.99\%$ 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 ~~neighbors~~ neighbors 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 neighbor values into a meta-model, the grid resolution we currently use (about 26 km horizontally and about ~~200-300~~ 200-300 m vertically in the height range of interest) may be too low to adequately account for horizontal neighbors. We expect that using a finer grid for ICON predictions may enable such an approach, and most likely improve the overall verification scores.

6.3 Prediction Improvement via Machine Learning

7 Outlook

~~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~~

875 Prediction Improvement via Machine Learning

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—, comparable to a binary 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 tree library CatBoost in classification mode. The results are shown in Appendix Fig. 10. The ROC curve of the CatBoost model shows a slight improvement in the upper left region of interest compared to the ~~k-out-of-10~~ 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

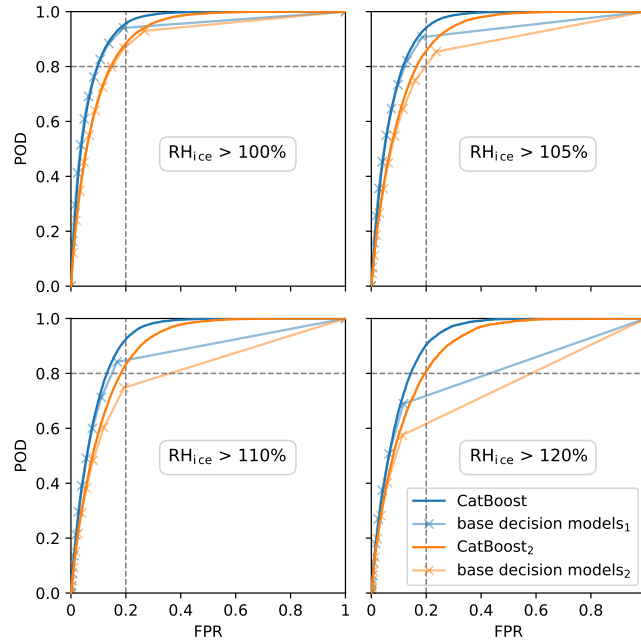


Figure 10. Comparison of ROC curves of EPS-based CatBoost and EPS-based decision models. CatBoost input features were the RH_{ice} values of all ten members. Solid blue ROC curves: Training and validation period from April to December 2024; test data from January to March 2025; ROC calculated for the 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.

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 [one-moment 40-member ICON 1-Mom EPS](#) 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.

8 Conclusions

This study demonstrates the [great potential of an EPS model for ISSR prediction](#) [strong potential of EPS-based classification models for ISSR](#), based on the ICON NWP model [with an adapted enhanced with a](#) two-moment ice microphysics scheme. The [two-moment scheme](#) [Compared to ICON 1-Mom, ICON 2-Mom](#) more accurately captures the physical conditions associated with (higher)-ice supersaturation, [which many one-moment schemes struggle to represent or fail to identify](#). Prior 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 where it significantly improves the POD while maintaining a moderate FPR. This improvement is also reflected by the MCC, indicating better overall classification skill.

895 The EPS model itself, ICON 2-Mom EPS, 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 Simple k-out-of-10 decision models spanned a wide range of POD-FPR ~~values~~ combinations, 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. outperforming the deterministic ICON 2-Mom model in terms of POD while maintaining comparable FPRs. For $RH_{ice} > 100\%$, ICON 2-Mom achieves a POD of ~ 0.6 and an FPR of ~ 0.1 , whereas ICON 2-Mom EPS allows for finer control: decision model 1 yields a POD > 0.9 at an FPR of 0.25, while decision
900 model 9 offers near-zero FPR with reduced POD. This flexibility enables users to select decision models based on operational cost trade-offs between false positives (e.g., unnecessary diversions) and false negatives (e.g., contrail formation).

~~These approaches were statistical in nature, meaning that we used classical statistical methods and verification results to define a~~ Further refinement was achieved by incorporating ensemble spread into the decision making. Grouping ROC curves by the standard deviation of RH_{ice} revealed that low-spread conditions correspond to high categorical skill, whereas high-spread
910 scenarios tend toward random performance. This insight was used to define an adaptive meta-model that ~~functions as a newly developed forecast model.~~ selects k based on ensemble spread, keeping FPR below a target level. This approach relies solely on model data and can be seamlessly integrated into more advanced models.

Building on this ~~methodology, we trained statistical framework,~~ a gradient boosting tree classifier ~~representing a more advanced~~ was trained as a more sophisticated meta-model. Despite ~~being trained in under a minute using minimal training time~~
915 and default hyperparameters, ~~the model outperformed the k-out-of-N~~ it outperformed the k-out-of-10 models in the POD-FPR region of interest. Additional advantages ~~of this model include an almost~~ include a nearly continuous ROC curve and ~~its ability to integrate additional features in a straightforward manner~~ the ability to incorporate additional features with ease.

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~~ ICON 2-Mom EPS rerouted over 100 flights, demonstrating the
920 operational relevance of this forecasting approach. The results presented ~~in this study demonstrate here show~~ that EPS-based meta-models bring us even closer to reliably identifying ~~the potential for~~ conditions conducive to persistent contrail formation.

~~The results of this study can also be informative for~~ Finally, these findings may inform the European Union's ~~recently established's~~ Monitoring, Reporting and Verification (MRV) system, ~~where which uses~~ climate response models are used to quantify the trade-off to quantify trade-offs between contrails, CO₂ emissions, and other greenhouse gases. ~~Climate response~~
925 ~~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 a critical input for contrail modeling, yet remains~~ poorly predicted by state-of-the-art operational NWP models many operational NWP systems. This study ~~is a step~~

~~towards improved prediction of ISSR and~~ represents a step toward more accurate RH_{ice} forecasting and improved support for climate-conscious aviation strategies.

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

Appendix A: ~~The Two-Moment History and Details of the One-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
935 as the Lokalmmodell (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
940 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~~
945 ~~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.~~

950 ~~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 $D_{i,\text{conv}} = 200 \mu\text{m}$ is assumed, corresponding to a mean~~
955 ~~mass of $m_{i,\text{conv}} = 10^{-9}$ 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.~~

A1 ~~Homogeneous Ice Nucleation~~

~~For homogeneous ice nucleation, the parameterization by Köhler et al. (2006) is used. It accounts for the presence of pre-existing~~
960 ~~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 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.

A1 Heterogeneous Ice Nucleation

- 965 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 $N_{\text{dust}} = 1 \times 10^3 \text{ m}^{-3}$ is assumed in the upper troposphere above $p_0 = 200 \text{ hPa}$. Below that pressure height the profile increases following

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

- 970 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 $1 \mu\text{m}$ and a standard deviation of 2.5. The number of nucleated ice particles is then diagnosed as:

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

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).

- 975 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.

Appendix B: Ensemble Verification of ICON 1-Mom EPS

Fig. B1 compares the ROC curves of the ICON 1-Mom EPS and the ICON 2-Mom EPS for $RH_{ice} > 100\%$ events. The We also evaluated the ensemble data of the operational ICON 1-Mom EPS is noteworthy as it is the operational ICON model. The 1-Mom 10-member data was obtained by resampling. Overall, 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 2-Mom 10-member EPS performs as well as the 1-Mom operational 40-member EPS, while the 1-Mom EPS as well as for 10-member EPS performs slightly but significantly worse. For larger RH_{ice} thresholds, the 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 breaks down, as shown in the inset. 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 aligns with studies that emphasize the importance of model quality as a key factor in the success of ensemble prediction systems (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.

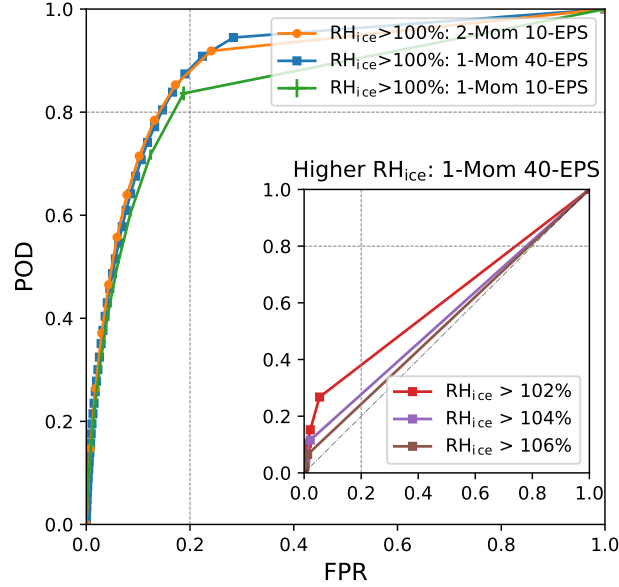


Figure B1. ROC curves for the ICON 2-Mom EPS (orange), the operational ICON 1-Mom EPS with 40 members (blue) and for the corresponding ICON 1-Mom EPS subsets with 10 members (green). For the latter, we randomly selected 1000 10-member EPS subsets, calculated the ROC curve for each and plotted the mean 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 RH_{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 12 h; Northern Hemisphere.

1000 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~~-vapor partial pressure (e) using the formula from Hardy (1998), ensuring consistency with the processing applied by radiosonde manufacturers, such as Vaisala. We further calculate
 1005 the saturation ~~vapour~~-vapor pressure over ice (e_i) consistently with the formula used in ICON which is given by

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

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
1010 receive

$$\text{RH}_{\text{ice}} = \frac{e}{e_i} 100 \%. \quad (\text{C2})$$

C2 Computation of RH_{ice} for ICON Data

First we calculate the water ~~vapour~~vapor partial pressure e by

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

1015 where the temperature T (in K), the density of moist air ρ (in kg/m^3), and the specific water ~~vapour~~vapor content qv (in kg/kg) are output variables of ICON, and $r_v = 461.51$ is the gas constant for water ~~vapour~~vapor. 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~~vapor 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
1020 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.

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

1025 In Fig. ??-D1 a zoom of the top right of Fig. ??-6 is provided, where the details of the precision-recall-PR curve for the non-ISSR event and for the events $\{RH_{ice} \leq \text{threshold}\}$ 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 $\{RH_{ice} \leq \text{threshold}\}$ ensemble members with the event $\{RH_{ice} \leq \text{threshold}\}$.

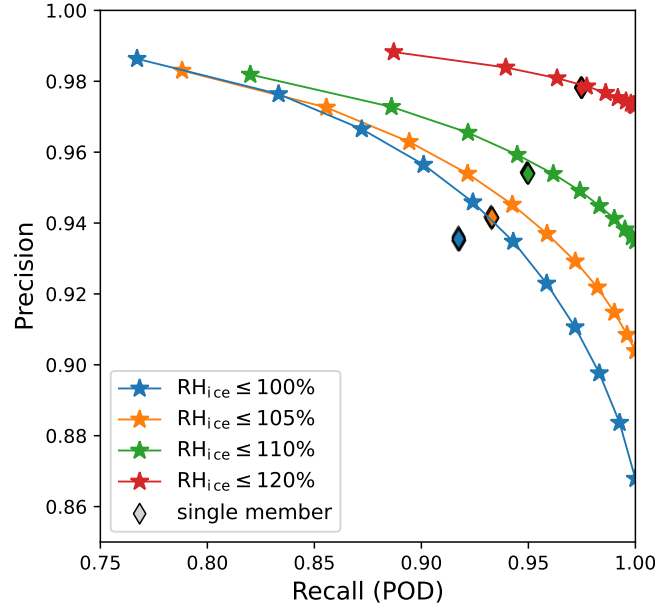


Figure D1. Precision-recall curve of events $\{RH_{ice} \leq \text{threshold}\}$ with threshold in $\{100\%, 105\%, 110\%, 120\%\}$ (zoom of top right of Figure ??-6). Stars on the lines indicate the scores corresponding to the decision models based on the EPS. The pink cross highlights decision model 4 for which the maximum F_1 score is obtained. Thin diamonds indicate the scores of the single ensemble members.

Appendix E: Binary Classification Models: CatBoost

1030 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:

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

where N is the total number of samples (spatio-temporal matching points of model and observation), y_i is 1 if an event was
1035 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~~ minimize the effect of potential correlations in the data. Figure 10 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.

~~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~~
1045 ~~RH_{ice} threshold.~~

Author contributions. LS and AS supervised and conceived the study, with contributions from MH and CGK. AS implemented the adapted two-moment ice microphysics scheme in ICON, based on a previous COSMO/GME implementation by CGK, and set up the dedicated ICON system at DWD. MH developed the verification methods and conducted the analysis. All authors discussed the results. MH drafted the manuscript with input from all authors, and all contributed to the final version.

1050 *Competing interests.* The authors declare that they have no competing interests.

Acknowledgements. MH and CGK 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
1055 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,
1060 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~~ [Marsigli](#) 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
1065 implementation and insightful discussions on the results.

[The two anonymous referees are acknowledged for their constructive comments which improved the final manuscript.](#)

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, <https://doi.org/10.1175/JAS-D-23-0210.1>, 2024.
- Appleman, H.: The formation of exhaust condensation trails by jet aircraft, *Bulletin of the American Meteorological Society*, 34, 14–20, <https://doi.org/10.1175/1520-0477-34.1.14>, 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, <https://doi.org/10.1002/qj.289>, 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, <https://doi.org/10.1175/2010MWR3594.1>, 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, <https://doi.org/10.1175/JCLI-D-19-0467.1>, 2020.
- Bickel, M., Ponater, M., Burkhardt, U., Righi, M., Hendricks, J., and Jöckel, P.: Contrail Cirrus Climate Impact: From Radiative Forcing to Surface Temperature Change, *Journal of Climate*, 38, 1895–1912, <https://doi.org/10.1175/JCLI-D-24-0245.1>, 2025.
- 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, e2024JD040981, <https://doi.org/10.1029/2024JD040981>, 2024.
- Borella, A., Vignon, É., Boucher, O., Meurdesoif, Y., and Fairhead, L.: A new prognostic parameterization of subgrid ice supersaturation and cirrus clouds in the ICOLMDZ AGCM, *Journal of Advances in Modeling Earth Systems*, 17, e2024MS004918, <https://doi.org/2024MS004918>, 2025.
- 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, <https://doi.org/10.1016/B978-0-12-819457-6.00014-8>, 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.
- Dekoutsidis, G., Groß, S., Wirth, M., Krämer, M., and Rolf, C.: Characteristics of supersaturation in midlatitude cirrus clouds and their adjacent cloud-free air, *Atmospheric Chemistry and Physics*, 23, 3103–3117, <https://doi.org/10.5194/acp-23-3103-2023>, 2023.
- Dietmüller, S., Matthes, S., Dahmann, 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-CO₂ algorithmic climate change functions: CLIMaCCF V1.0, *Geoscientific Model Development Discussions*, 2022, 1–33, <https://doi.org/10.5194/gmd-2022-121>, 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, <https://library.wmo.int/records/item/68808-report-of-wmo-s-2022-upper-air-instrument-intercomparison-campaign>, 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, <https://doi.org/10.7289/V5/ON-NCEP-493>, 2018.
- Dyroff, C., Zahn, A., Christner, E., Forbes, R., Tompkins, A. M., and van Velthoven, P. F.: Comparison of ECMWF analysis and forecast humidity data with CARIBIC upper troposphere and lower stratosphere observations, *Quarterly Journal of the Royal Meteorological Society*, 141, 833–844, <https://doi.org/10.1002/qj.2400>, 2015.
- ECMWF: IFS Documentation CY49R1 - Part IV: Physical Processes, chap. 4, ECMWF, <https://doi.org/10.21957/c731ee1102>, 2024.
- Epstein, E. S.: Stochastic dynamic prediction, *Tellus*, 21, 739–759, <https://doi.org/10.3402/tellusa.v21i6.10189>, 1969.
- Field, P. R., Hogan, R. J., Brown, P. R., Illingworth, A. J., Choullarton, T. W., and Cotton, R. J.: Parametrization of ice-particle size distributions for mid-latitude stratiform cloud, 131, 1997–2017, <https://doi.org/10.1256/qj.04.134>, 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, <https://doi.org/10.3390/aerospace4030034>, 2017a.
- Grewe, V., Matthes, S., Frömming, C., Brinkop, S., Jöckel, P., Gierens, K., Champougny, T., Fuglestedt, J., Haslerud, A., Irvine, E., et al.: Feasibility of climate-optimized air traffic routing for trans-Atlantic flights, *Environmental Research Letters*, 12, 034003, <https://doi.org/10.1088/1748-9326/aa5ba0>, 2017b.
- 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, <https://doi.org/10.1038/s41467-021-24091-y>, 2021.
- Grundner, A., Beucler, T., Gentine, P., and Eyring, V.: Data-driven equation discovery of a cloud cover parameterization, *Journal of Advances in Modeling Earth Systems*, 16, e2023MS003763, <https://doi.org/10.1029/2023MS003763>, 2024.
- 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, <https://www.decatur.de/javascript/dew/resources/its90formulas.pdf>, 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, <https://doi.org/10.1016/j.physd.2006.11.008>, 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, <https://doi.org/10.1029/2008JD010341>, 2009.
- Kärcher, B.: Formation and radiative forcing of contrail cirrus, *Nature communications*, 9, 1824, <https://doi.org/10.1038/s41467-018-04068-0>, 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, <https://doi.org/10.1029/2005JD006219>, 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, e2021JD035805, <https://doi.org/10.1029/2021JD035805>, 2022.
- Kärcher, B., Jensen, E., Pokrifka, G., and Harrington, J. Y.: Ice supersaturation variability in cirrus clouds: Role of vertical wind speeds and deposition coefficients, *Journal of Geophysical Research: Atmospheres*, 128, e2023JD039324, <https://doi.org/10.1029/2023JD039324>, 2023.
- Klöwer, M., Allen, M., Lee, D., Proud, S., Gallagher, L., and Skowron, A.: Quantifying aviation’s contribution to global warming, *Environmental Research Letters*, 16, 104027, <https://doi.org/10.1088/1748-9326/ac286e>, 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, <https://doi.org/10.3402/tellusb.v67.24494>, 2015.
- 1145 Krüger, K., Schäfler, A., Wirth, M., Weissmann, M., and Craig, G. C.: Vertical structure of the lower-stratospheric moist bias in the ERA5 reanalysis and its connection to mixing processes, *Atmospheric Chemistry and Physics*, 22, 15 559–15 577, <https://doi.org/10.5194/acp-22-15559-2022>, 2022.
- Kunz, A., Spelten, N., Konopka, P., Müller, R., Forbes, R. M., and Wernli, H.: Comparison of Fast In situ Stratospheric Hygrometer (FISH) measurements of water vapor in the upper troposphere and lower stratosphere (UTLS) with ECMWF (re) analysis data, *Atmospheric chemistry and physics*, 14, 10 803–10 822, <https://doi.org/10.5194/acp-14-10803-2014>, 2014.
- 1150 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, <https://doi.org/10.3390/aerospace4030042>, 2018.
- Lee, D. S., Allen, M. R., Cumpsty, N., Owen, B., Shine, K. P., and Skowron, A.: Uncertainties in mitigating aviation non-CO₂ emissions for climate and air quality using hydrocarbon fuels, *Environmental Science: Atmospheres*, 3, 1693–1740, <https://doi.org/10.1039/D3EA00091E>, 2023.
- 1155 Lewis, J. M.: Roots of ensemble forecasting, *Monthly weather review*, 133, 1865–1885, <https://doi.org/10.1175/MWR2949.1>, 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, [https://doi.org/10.1175/1520-0450\(1983\)022<1065:BPOTSF>2.0.CO;2](https://doi.org/10.1175/1520-0450(1983)022<1065:BPOTSF>2.0.CO;2), 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, <https://doi.org/10.3390/aerospace8050120>, 2021.
- 1160 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, <https://doi.org/10.5194/egusphere-2024-1123>, 2024.
- Magnus, G.: Versuche über die Spannkkräfte des Wasserdampfs, *Annalen der Physik*, 137, 225–247, https://zs.thulb.uni-jena.de/receive/jportal_jparticle_00114539, 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, [https://doi.org/10.1175/1520-0493\(2002\)130<0319:TOGIHG>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<0319:TOGIHG>2.0.CO;2), 2002.
- 1165 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, <https://doi.org/10.3390/aerospace4030042>, 2017.
- 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, <https://doi.org/10.5194/gmd-2023-92>, 2023.
- 1170 Murray, F. W.: On the computation of saturation vapor pressure, *Journal of Applied Meteorology and Climatology*, 6, 203–204, [https://doi.org/10.1175/1520-0450\(1967\)006<0203:OTCOSV>2.0.CO;2](https://doi.org/10.1175/1520-0450(1967)006<0203:OTCOSV>2.0.CO;2), 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, <https://doi.org/10.1111/j.1600-0870.2004.00076.x>, 2004.
- Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C., Gebhardt, C., Marsigli, C., Förstner, J., Zängl, G., Schlemmer, L., Blahak, U., and Welzbacher, C.: DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting Sys-

- tem, Tech. rep., Deutscher Wetterdienst, https://isabel.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/icon_d2/icon_d2_dbbschr_aktuell.pdf?view=nasPublication&nn=16102, 2025.
- 1180 Reisner, J., Rasmussen, R. M., and Brientjes, R. T.: Explicit forecasting of supercooled liquid water in winter storms using the MM5 mesoscale model, 124, 1071–1107, <https://doi.org/10.1002/qj.49712454806>, 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, <https://doi.org/10.1029/JD091iD13p14385>, 1986.
- 1185 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, <https://doi.org/10.1127/metz/5/1996/4>, 1996.
- 1190 Schumann, U.: A contrail cirrus prediction model, *Geosci. Model Dev.*, 5, 543–480, <https://doi.org/10.5194/gmd-5-543-2012>, 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, <https://doi.org/10.2514/6.2011-3376>, 2011.
- Seifert, A. and Beheng, K. D.: A double-moment parameterization for simulating autoconversion, accretion and selfcollection, 59-60, 265–281, [https://doi.org/10.1016/S0169-8095\(01\)00126-0](https://doi.org/10.1016/S0169-8095(01)00126-0), 2001.
- 1195 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, <https://doi.org/10.5194/acp-23-6409-2023>, 2022.
- Shapiro, M., Engberg, Z., Teoh, R., Stettler, M., and Dean, T.: pycontrails: Python library for modeling aviation climate impacts, *Zenodo* [code], 10, <https://doi.org/10.5281/zenodo.10015108>, 2023.
- 1200 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, <https://doi.org/10.3390/aerospace9030146>, 2022.
- Snyder, C. and Zhang, F.: Assimilation of simulated Doppler radar observations with an ensemble Kalman filter, *Monthly Weather Review*, 131, 1663–1677, [https://doi.org/10.1175/1520-0493\(2003\)131<1663:AOSDRO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<1663:AOSDRO>2.0.CO;2), 2003.
- 1205 Sonntag, D.: Advancements in the field of hygrometry, *Meteorologische Zeitschrift* (Berlin);(Germany), 3, <https://doi.org/10.1127/metz/3/1994/51>, 1994.
- Spichtinger, P. and Leschner, M.: Horizontal scales of ice-supersaturated regions, *Tellus B: Chemical and Physical Meteorology*, 68, 29 020, <https://doi.org/10.3402/tellusb.v68.29020>, 2016.
- 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, <https://doi.org/10.5194/acp-23-2035-2023>, 2023.
- 1210 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, <https://doi.org/10.1007/s00703-002-0568-2>, 2003.
- Teoh, R., Schumann, U., Gryspeerd, E., Shapiro, M., Molloy, J., Koudis, G., Voigt, C., and Stettler, M. E.: Aviation contrail climate effects in the North Atlantic from 2016 to 2021, *Atmospheric Chemistry and Physics*, 22, 10 919–10 935, <https://doi.org/10.5194/acp-22-10919-2022>, 2022.
- 1215

- Teoh, R., Engberg, Z., Schumann, U., Voigt, C., Shapiro, M., Rohs, S., and Stettler, M. E.: Global aviation contrail climate effects from 2019 to 2021, *Atmospheric Chemistry and Physics*, 24, 6071–6093, <https://doi.org/10.5194/acp-24-6071-2024>, 2024.
- Tetens, O.: Über einige meteorologische Begriffe, *Z. geophys.*, 6, 297–309, 1930.
- Thompson, G., Rasmussen, R. M., and Manning, K.: Explicit forecasts of winter precipitation using an improved bulk micro-
1220 physics scheme. Part I: Description and sensitivity analysis, *Monthly Weather Review*, 132, 519–542, [https://doi.org/10.1175/1520-0493\(2004\)132<0519:EFOWPU>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2), 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, 107663, <https://doi.org/10.1016/j.atmosres.2024.107663>, 2024.
- 1225 Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, *Mon. Weather Rev.*, 117, 1779–1800, [https://doi.org/10.1175/1520-0493\(1989\)117<1779:ACMFSF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2), 1989.
- Tiedtke, M.: Representation of clouds in large-scale models, *Monthly Weather Review*, 121, 3040–3061, [https://doi.org/10.1175/1520-0493\(1993\)121<3040:ROCILM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1993)121<3040:ROCILM>2.0.CO;2), 1993.
- Tompkins, A. M., Gierens, K., and Rädcl, G.: Ice supersaturation in the ECMWF integrated forecast system, *Quarterly Journal of the
1230 Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 133, 53–63, <https://doi.org/10.1002/qj.14>, 2007.
- 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, <https://doi.org/10.1175/JAS-D-16-0074.1>, 2017.
- 1235 Unterstrasser, S. and Gierens, K.: Numerical simulations of contrail-to-cirrus transition–Part I: An extensive parametric study, *Atmospheric Chemistry and Physics*, 10, 2017–2036, <https://doi.org/10.5194/acp-10-2017-2010>, 2010.
- Vaisala: Vaisala Radiosonde RS41 Measurement Performance, in: White Paper, pp. 1–28, VAISALA, <https://www.vaisala.com/sites/default/files/documents/White%20paper%20RS41%20Performance%20B211356EN-A.pdf>, 2013.
- Wang, J., Chen, J., Du, J., Zhang, Y., Xia, Y., and Deng, G.: Sensitivity of ensemble forecast verification to model bias, *Monthly Weather
1240 Review*, 146, 781–796, <https://doi.org/10.1175/MWR-D-17-0223.1>, 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, <https://doi.org/10.5194/acp-25-2845-2025>, 2025.
- Wilhelm, J., Akytas, T., Bölöni, G., Wei, J., Ribstein, B., Klein, R., and Achatz, U.: Interactions between mesoscale and submesoscale
1245 gravity waves and their efficient representation in mesoscale-resolving models, *Journal of the Atmospheric Sciences*, 75, 2257–2280, <https://doi.org/10.1175/JAS-D-17-0289.1>, 2018.
- Wilson, D. R. and Ballard, S. P.: A microphysically based precipitation scheme for the UK Meteorological Office Unified Model, 125, 1607–1636, <https://doi.org/10.1002/qj.49712555707>, 1999.
- WMO: Report of WMO's 2022 Upper-Air Instrument Intercomparison Campaign, Instruments and Observing Methods Report 143, WMO, <https://library.wmo.int/idurl/4/68808>, 2024.
- 1250 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, <https://doi.org/10.5194/gmd-9-3363-2016>, 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, 1255 <https://doi.org/10.1002/qj.2378>, 2014.