Environmental Data Science (2026), 5: e14, 1–11
doi:10.1017/eds.2026.10051

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Calibrated conformal prediction intervals for microphysical process rates

Miriam Simm[1], Corinna Hoose[1] and Tom Beucler[2,3]

[1]Institute of Meteorology and Climate Research Troposphere Research, Karlsruhe Institute of Technology, Germany
[2]Faculty of Geosciences and Environment, University of Lausanne, Switzerland
[3]Expertise Center for Climate Extremes, University of Lausanne, Switzerland
Corresponding author: Miriam Simm; Email: miriam.simm@kit.edu

Received: 03 June 2026 UTC; **Revised:** 03 June 2026 UTC; **Accepted:** 04 June 2026 UTC

Keywords: cloud microphysics; conformal prediction; machine learning; quantile regression; uncertainty quantification

Abstract

Conformal prediction (CP) can yield statistically valid prediction intervals for any regression model, with no model modifications and small computational costs. To assess its practical value, we apply conformal methods to quantify uncertainty in machine learning emulators of six microphysical process rates (MPRs). MPRs describe small-scale processes in atmospheric clouds such as precipitation formation and aerosol–cloud interactions and help understand weather and climate. The emulators are trained on simulation output from the ICOsahedral Nonhydrostatic (ICON) model in a limited-area numerical weather prediction configuration. We compare split CP for deterministic emulators with conformalized quantile regression (CQR) for quantile regression (QR) emulators. Both CP methods yield well-calibrated and sharp prediction intervals on average, but CQR provides more consistent intervals across several orders of magnitude, making it preferable for the uncertainty quantification of climate variables.

Impact Statement

Robust uncertainty quantification is important for the reliability of model predictions. Conformal prediction is a model-agnostic and distribution-free framework to obtain statistically valid prediction intervals, but remains underused in climate science. By applying and evaluating conformal prediction methods for microphysical process rates spanning several orders of magnitude, we demonstrate their potential for inexpensive uncertainty quantification in weather and climate modeling.

1. Introduction

In atmospheric prediction, purely deterministic methods increasingly reveal their limitations. For subgrid-scale parameterizations, grid-scale state variables may not uniquely determine the net effect of unresolved processes (Palmer, 2019; Christensen et al., 2024). This motivates probabilistic frameworks, stochastic parameterizations, and robust uncertainty quantification (Haynes et al., 2023; Christensen et al., 2024; Schreck et al., 2024; Mansfield and Christensen, 2025), often implemented through ensemble approaches (e.g., Gagne II et al., 2020; Behrens et al., 2025) that can be computationally costly.

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

As a computationally inexpensive alternative, conformal prediction (CP) (Gammerman et al., 1998; Angelopoulos and Bates, 2022; Vovk et al., 2022) provides distribution-free prediction intervals with finite-sample validity for any regression model, under the assumption that the data is exchangeable, which is weaker than the independent and identically distributed assumption commonly used in machine learning (ML). Yet, CP has seen limited use in weather and climate applications (e.g., Gopakumar et al. 2026; Mortier et al., 2025), in part because spatio-temporal dependence in geophysical data can violate exchangeability. Extending conformal methods to such settings remains an active research area (e.g., Gibbs and Candès, 2021; Xu and Xie, 2021; Sun, 2022).

In numerical models, the spatial and temporal discretization separates resolved from unresolved scales. Subgrid-scale physical processes must therefore be represented through parameterizations, which are a major source of model uncertainty (Boucher et al., 2014). Cloud microphysics describes phase transitions of condensed water in the atmosphere and interactions among liquid and frozen particles, water vapor, and aerosols. These processes control precipitation formation and intensity, impact storm evolution, and, through latent heating and cooling, influence cloud dynamics and large-scale cloud structure. Microphysical properties also affect radiative transfer and thus the Earth's radiation budget (Gettelman et al., 2019; Morrison et al., 2020; Lamb et al., 2026). Accurate representation of cloud microphysics remains particularly challenging due to its inherent complexity and nonlinearity, as well as incomplete process-level understanding (Khain et al., 2015; Zelinka et al., 2020). Improved understanding of cloud processes will help to reduce persistent uncertainties and advance the development of ML-based models (Lamb et al., 2026). Microphysical process rates (MPRs), computed internally to update the prognostic cloud variables, would provide detailed insights into cloud processes and support model development. Yet, storage constraints typically preclude saving MPR output from large-scale high-resolution simulations because MPRs comprise a two-digit number of three-dimensional variables. Moreover, offline recalculations of process rates are generally not accurate because of a temporal mismatch between the prognostic cloud variables in the model output and the MPRs computed during time stepping.

Here, we train ML models to emulate the computation of six MPRs from a two-moment bulk microphysics scheme (Seifert and Beheng, 2006) and quantify predictive uncertainty with CP methods. Using ML emulators forced by atmospheric state variables from limited-area ICOsahedral Nonhydrostatic (ICON) simulations (Zängl et al., 2015), we (i) apply split CP to calibrate prediction intervals for deterministic emulators, (ii) train quantile regression (QR) emulators and calibrate their intervals via conformalized quantile regression (CQR) (Romano et al., 2019), and (iii) benchmark interval calibration and sharpness across emulators and CP methods.

The structure is as follows. Section 2 reviews CP and CQR. Section 3 describes the MPRs, the ICON model output-based datasets, and the training and calibration procedure. Section 4 presents the results, and Section 5 concludes.

2. Theory: Conformal prediction

CP provides distribution-free prediction intervals for any regression or classification model, which contain the true value of the target with a pre-defined probability, requiring only data exchangeability (Vovk et al., 2022). While originally formulated in a transductive setting (Gammerman et al., 1998), "full" CP is computationally expensive because it requires retraining the underlying model for each test point. We hence focus on inductive ("split") CP, which is computationally efficient but requires splitting the data into a proper training set and a calibration set (Lei et al., 2018). In split CP (Figure 1), a point-prediction model is trained on the proper training set, and nonconformity scores, which measure the prediction error, are computed on the calibration set. The $(1-\alpha)$-quantile of these scores sets the width of the prediction intervals. Here, α is a user-specified miscoverage rate, which quantifies the probability that the true value falls outside the prediction interval. This yields prediction intervals that, by construction, marginally contain the true target value with $100(1-\alpha)\%$ probability. Although this guarantee holds independently of model skill, data distribution, and nonconformity score function, interval informativeness mainly depends on the choice of score function (Angelopoulos and Bates, 2022).

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

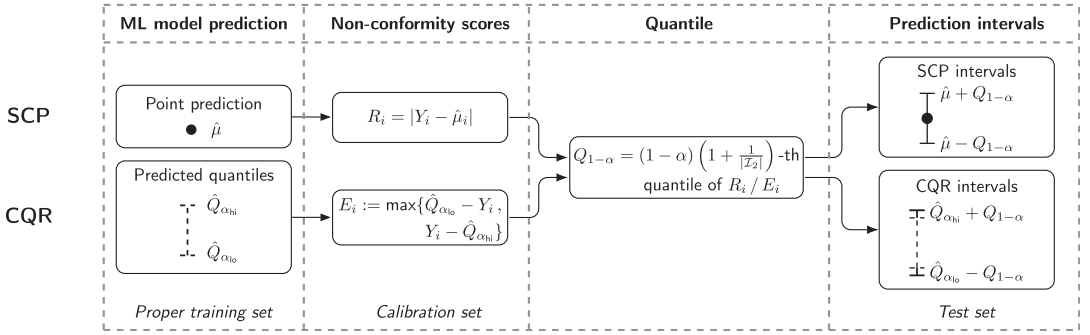


Figure 1. *Conformal prediction framework: split conformal prediction (top row) and conformalized quantile regression (bottom row).*

2.1. Split conformal prediction without retraining (SCP)

More formally, given n training samples $\{(X_i, Y_i)\}_{i=1}^{n}$, the main objective is to construct the prediction interval $C(X_{n+1}) \subseteq \mathbb{R}$ for a test point X_{n+1}. The interval C contains the unknown value Y_{n+1} with the miscoverage rate $\alpha \in (0,1)$, satisfying the marginal coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1})\} \geq 1 - \alpha, \tag{2.1}$$

for any joint distribution of the feature vectors $X \in \mathbb{R}^p$ of dimension $p \geq 1$ and labels $Y \in \mathbb{R}$ and any sample size n with the assumption that all samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are drawn exchangeably from the joint distribution.

The procedure for constructing the prediction intervals is straightforward. First, a regression model is trained on a *proper training* subset indexed by $\mathcal{I}_1 \subset \{1, \ldots, n\}$ to predict $\hat{\mu}(x)$. Then, a disjoint *calibration* subset indexed by $\mathcal{I}_2 \subset \{1, \ldots, n\}$ is used to compute the nonconformity scores

$$R_i = |Y_i - \hat{\mu}(X_i)| \, i \in \mathcal{I}_2, \tag{2.2}$$

i.e., the residuals. Computing

$$Q_{1-\alpha}(R, \mathcal{I}_2) := (1-\alpha)\left(1 + \frac{1}{|\mathcal{I}_2|}\right)\text{-th empirical quantile of } \{R_i\}_{i \in \mathcal{I}_2}, \tag{2.3}$$

with $|\mathcal{I}_2|$ the cardinality of the calibration set, yields the prediction interval

$$C(X_{n+1}) = [\hat{\mu}(X_{n+1}) - Q_{1-\alpha}(R, \mathcal{I}_2), \hat{\mu}(X_{n+1}) + Q_{1-\alpha}(R, \mathcal{I}_2)], \tag{2.4}$$

which satisfies (2.1). Setting $\alpha = 0.1$ corresponds to 90% prediction intervals.

The coverage property (2.1) is *marginal*, i.e., averaged over all samples $\{(X_i, Y_i)\}_{i=1}^{n+1}$ (Romano et al., 2019; Feldman et al., 2021). The coverage guarantee holds for the distribution of the test point X_{n+1}, but may not hold for the conditional distribution of Y_{n+1} given X_{n+1},

$$\mathbb{P}\{Y_{n+1} \in C(X_{n+1}) \mid X_{n+1}\} \geq 1 - \alpha, \tag{2.5}$$

which is referred to as conditional coverage and, in general, cannot be achieved (Lei and Wasserman, 2012). As a result, the coverage probability can fall below $100(1-\alpha)\%$ for subsets of the data, such as certain regimes or ranges of the target variable.

2.2. Conformalized quantile regression

While SCP allows for constructing marginally valid prediction intervals in a straightforward way at almost no additional computational cost, Eq. (2.4) reveals that the length of the prediction interval $C(X_{n+1})$, is fixed to $2Q_{1-\alpha}(R, \mathcal{I}_2)$, independent of the input X_{n+1}. SCP is based on the implicit assumption

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

that the spread of the residuals is constant for all inputs X_{n+1}, i.e., *homoscedasticity*, which is in practice often not the case. To address this issue, Romano et al. (2019) proposed CQR. In CQR, a QR model is used to derive prediction intervals that can handle heteroscedasticity and thus are potentially more informative.

QR (Koenker and Bassett, 1978) is a statistical method to estimate conditional quantiles of Y_{n+1} given X_{n+1}, instead of the conditional mean. Here, a QR model is trained on the proper training dataset to predict the lower and upper quantile $\hat{Q}_{\alpha_{\text{lo}}}(x)$ and $\hat{Q}_{\alpha_{\text{hi}}}(x)$, which constitute initial estimates of the lower and upper bound of the prediction interval, $\widehat{C}(x) = \left[\hat{Q}_{\alpha_{\text{lo}}}(x), \hat{Q}_{\alpha_{\text{hi}}}(x)\right]$, with $\alpha_{\text{lo}} = \alpha/2$ and $\alpha_{\text{hi}} = 1 - \alpha/2$, e.g., $\widehat{C}(x) = \left[\hat{Q}_{0.05}, \hat{Q}_{0.95}\right]$. To quantify the error of this ad hoc prediction interval, nonconformity scores

$$E_i := \max\left\{\hat{Q}_{\alpha_{\text{lo}}}(X_i) - Y_i, Y_i - \hat{Q}_{\alpha_{\text{hi}}}(X_i)\right\}, \tag{2.6}$$

are computed on the calibration dataset for each $i \in \mathcal{I}_2$. For a new input X_{n+1}, the prediction interval is conformalized by computing $Q_{1-\alpha}(E, \mathcal{I}_2)$ analogously to SCP (2.3). This yields the conformalized prediction interval.

$$C(X_{n+1}) = \left[\hat{Q}_{\alpha_{\text{lo}}}(X_{n+1}) - Q_{1-\alpha}(E, \mathcal{I}_2), \hat{Q}_{\alpha_{\text{hi}}}(X_{n+1}) + Q_{1-\alpha}(E, \mathcal{I}_2)\right]. \tag{2.7}$$

3. Data and methods

3.1. Microphysical process rates

We aim to reconstruct MPRs from output of the ICON model, which were not included in the model output at the time of simulation, using a two-moment bulk microphysics scheme.

Bulk parameterization schemes describe cloud microphysical properties within each model grid volume with statistical bulk quantities, i.e., moments of the size distribution functions of distinct hydrometeor categories. The two-moment microphysics scheme (Seifert and Beheng, 2006) parameterizes cloud microphysics with prognostic number concentrations and mass mixing ratios of cloud droplets, raindrops, cloud ice, snow, graupel, and hail, corresponding to the first two moments of the respective size distribution functions. MPRs describe a change in the prognostic variables through phase transitions and interaction processes (Figure 2). (In this work, we only consider process rates related to a change in the mass mixing ratio.) Here, we consider two warm-rain processes, which both describe an increase in the rain mass mixing ratio through the formation of raindrops from coalescing cloud droplets (autoconversion) and cloud droplets and raindrops (accretion). Evaporation of rain causes a decrease in rainwater mass. Rain mass also decreases by freezing of rain to snow, graupel, and hail at temperatures $T < 0\,^\circ$C. At $T > 0\,^\circ$C, all frozen hydrometeors can melt to rain, leading to an increase in the rain mass mixing ratio. Furthermore, we consider riming, whereby liquid particles freeze upon contact with a frozen hydrometeor. The riming rate comprises eight individual processes between each category of hydrometeors in the liquid and solid phase. The melting, freezing, and riming rates represent the sum of all individual processes. As inputs to the ML models, we use mass mixing ratios and number concentrations together with temperature, pressure, and density. (These are listed in the Supplementary Material A.2. A short discussion of exchangeability can be found in Supplementary Material A.3.)

3.2. ICON model simulations and sampling procedure

For training, calibration, and evaluation, we use a dataset compiled from simulation output of the ICON model (version 2.6.6) (Zängl et al., 2015) in a limited-area configuration with $\overline{\Delta x} \approx 2$ km effective grid spacing. The simulations are performed with the two-moment microphysics scheme (Seifert and Beheng, 2006) with a fast-physics timestep $t_{\text{fast}} = 20$s. The modeling domain extends from $-0.4\,^\circ$W to $17.7\,^\circ$E and from $43.7\,^\circ$N to $57.26\,^\circ$N, corresponding to the operational ICON-D2 configuration (Reinert et al., 2025). Each vertical column contains 65 levels up to 22 km above ground. The dataset comprises 19 one-day simulations (24 hours each) from January 2022 to July 2023 with an output time step of 10 minutes. We perform simulations for 1 day for each month. For training purposes, the MPRs are added as additional output variables in our setup. Further details on the model configuration are given in Supplementary Material A.1. The set of input features for each target variable is aligned with the computation of the

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

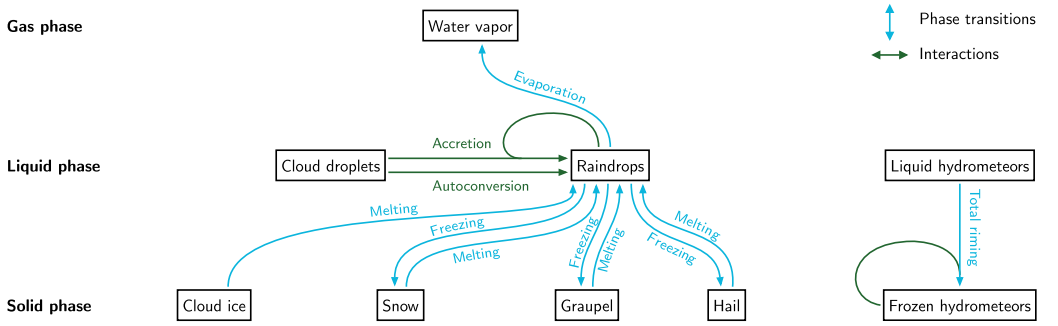


Figure 2. *Schematic of selected microphysical processes (arrows) between the six hydrometeor categories (boxes) and water vapor in the two-moment microphysics scheme of Seifert and Beheng (2006). Horizontal (green) arrows represent interaction processes, vertical (blue) arrows represent phase transitions. For simplicity, riming is shown separately. If a process occurs more than once, arrows represent contributions to the total process rate. From an ML perspective, boxes represent input features (mass mixing ratios and number concentrations) and arrows represent targets (process rates).*

process rate in the ICON subroutine and listed in Supplementary Material B.1. The size of the initial dataset is reduced by random sampling of model grid points. Furthermore, the output from high altitudes is discarded under the assumption that they are cloud-free. We use the simulation output for 2022 for training (even-numbered months) and validation (odd-numbered months); the simulated days for 2023 are used for the test dataset. The training dataset is randomly split into the proper training and the calibration set. The datasets for each target MPR are retrieved from these initial datasets by filtering out all grid points where the value of the MPR fulfills $|\mathrm{MPR}| \geq 10^{-12}\,\mathrm{kg\,kg^{-1}\,t_{fast}}^{-1}$ and $\sum_k q_k \geq 10^{-12}\,\mathrm{kg\,kg^{-1}}$ over all mass mixing ratios q_k in the set of input features, which corresponds to the threshold $q_{\mathrm{crit}} = 10^{-12}\,\mathrm{kg\,kg^{-1}}$ in the two-moment microphysics scheme. (In the ICON model, q_{crit} is a critical mass mixing ratio threshold used in the parameterization of certain processes, such as autoconversion.) This allows us to train the regression models to predict logarithmically transformed targets, as the range of the target variables spans multiple orders of magnitude. Thus, the proper training dataset contains 1.12×10^7 samples, and each validation, calibration, and evaluation dataset contains 1.6×10^6 samples, corresponding to a 70%/10%/10%/10% split. The input features are scaled to the range $[0, 1]$ with *min-max* scaling.

3.3. Training and calibration

A separate model is trained for each process rate. In order to obtain prediction intervals with SCP, we train random forest (RF), gradient boosting (XGB), and neural network (NN) models to obtain point predictions for the process rates. For this, we use the `scikit-learn` library for the RF models (Pedregosa et al., 2011), the XGBoost library (Chen and Guestrin, 2016) for the XGB models, and PyTorch (Paszke et al., 2019) for the NN models. Calibration is performed with the absolute error residual (2.2) and $\alpha = 0.1$. For CQR, we employ three QR architectures: quantile RF (QRF) models using `quantile_forest`'s *RandomForestQuantileRegressor* (Meinshausen, 2006; Johnson, 2024), quantile gradient boosting (QXGB), and quantile feed-forward neural networks (QNN). We use the Adam optimizer (Kingma and Ba, 2017) for training the QNN with a learning rate scheduler with a minimum learning rate of 10^{-6}. The maximum number of epochs is 150. The QNN is trained with the *quantile loss* (Koenker and Bassett, 1978), frequently also referred to as the *pinball loss*,

$$L_\alpha(y, \hat{y}) = \alpha(y - \hat{y})\mathbb{1}\{y > \hat{y}\} + (1 - \alpha)(\hat{y} - y)\mathbb{1}\{y \leq \hat{y}\}, \tag{3.1}$$

where $\hat{y}$ is the model's prediction of the "true" y. All three QR models are trained to predict the 5%- and 95%-quantiles. More details on the training and the model hyperparameters are given in

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

Supplementary Material B.2. The predicted upper and lower quantiles are calibrated with the nonconformity score function (2.6) and $\alpha = 0.1$.

4. Results

4.1. Deterministic performance of microphysical process rate emulation

The performance of the deterministic RF, XGB, and NN models is evaluated with the coefficient of determination R^2 (Table 1). In order to obtain an estimate of the performance of the QR models, we compute the R^2 score with the median of the predicted (uncalibrated) upper and lower quantiles. The deterministic models show moderate to high R^2 scores ranging from 0.47 to 0.99, with the NN models achieving the highest R^2 scores for all MPRs. The lowest score is obtained for rain freezing with the XGB model. The QR models yield R^2 scores that are largely comparable to those of the deterministic models. An exception is the QRF model for autoconversion, where $R^2 \approxeq 0.02$. For all other QR models and process rates, R^2 ranges from 0.43 to 0.99.

4.2. Conformal prediction uncertainty estimates

We evaluate the calibration of the CP intervals with the prediction interval coverage probability (PICP)

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i \in C(X_i)\}, \tag{4.1}$$

which is the proportion of the true values that are covered by the prediction intervals, i.e., the PICP evaluates the marginal coverage property (2.1). Here, n is the size of the test set, Y_i the true value, and $C(X_i)$ the prediction interval. With $\alpha = 0.1$, we aim for PICP $\approx 90\%$. The results are given in Table 2; for comparison, we include the PICP for the uncalibrated prediction intervals obtained with the QR models. We find that the PICP $\approx 90\%$ for both the deterministic models calibrated with SCP and the CQR models. Yet, in some cases, the PICP falls slightly below the target 90%. This may be due to sampling variability but it could also indicate mild violations of the exchangeability assumption. Furthermore, we observe that calibration does not always result in a PICP that is closer to 90%, compared to the uncalibrated intervals. This can be seen, for instance, for autoconversion and the QRF model or accretion and the QXGB model. In these cases, calibration decreases the PICP slightly. This can occur because the CQR nonconformity score function (2.6) can be negative and thus shrink the predicted interval, as the score function is designed to account for both undercoverage and overcoverage (Romano et al., 2019). Nevertheless, in other cases we observe a significant improvement, for example with the QRF models for the melting to rain and total riming rate.

Table 1. *Deterministic performance in terms of the R^2 score*

	Autoconversion	Accretion	Rain evap.	Rain fr.	Rain melt.	Tot. riming
RF	0.67	0.97	0.93	0.52	0.90	0.78
XGB	0.66	0.70	0.81	0.47	0.97	0.90
NN	**0.77**	**0.98**	**0.98**	**0.78**	**0.99**	**0.91**
QRF	0.02	**0.99**	**0.98**	0.45	0.98	**0.91**
QXGB	**0.60**	0.74	0.83	0.46	0.88	0.82
QNN	0.43	0.88	0.66	**0.68**	**0.99**	0.82

Note: The scores for the QR models are computed with the median of the predicted uncalibrated upper and lower quantiles. For each process rate, the highest score for the deterministic and QR models is highlighted in bold.

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

Table 2. *Prediction interval coverage probability (PICP)*

	Autoconversion	Accretion	Rain evap.	Rain fr.	Rain melt.	Tot. riming
SCP, RF	90.18%	91.96%	**89.21%**	88.01%	90.96%	91.21%
SCP, XGB	89.82%	**90.63%**	87.61%	89.52%	90.17%	91.06%
SCP, NN	**89.98%**	91.43%	87.61%	**90.47%**	**89.98%**	**90.74%**
QRF, uncalibrated	89.47%	88.46%	97.91%	87.61%	94.53%	97.71%
CQR, QRF	86.44%	88.51%	**90.05%**	87.61%	89.18%	88.63%
QXGB, uncalibrated	89.24%	89.50%	89.58%	**90.66%**	89.00%	**90.28%**
CQR, QXGB	89.27%	88.83%	88.14%	89.18%	88.09%	89.14%
QNN, uncalibrated	89.68%	87.19%	95.71%	83.74%	90.87%	92.60%
CQR, QNN	**89.77%**	**90.41%**	90.11%	89.04%	**90.60%**	89.43%

Note: For comparison, we show the results for the uncalibrated prediction intervals obtained with QR in gray. For each process rate, the best PICP obtained with SCP and CQR is highlighted in bold, which is defined as $\min(|\text{PICP} - 90\%|)$.

The sharpness of the prediction intervals is evaluated with the normalized mean prediction interval width (NMPIW):

$$\text{NMPIW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{U(X_i) - L(X_i)}{\max(Y) - \min(Y)}\right), \tag{4.2}$$

where U and L refer to the calibrated upper and lower prediction interval bounds. We present the results in Table 3. (Note that the length of the calibrated and uncalibrated prediction intervals is similar in each case, therefore, we do not include both results.) We find that for each process rate, the smallest NMPIW is obtained with SCP. For accretion, rain evaporation, rain melting, and total riming, the NMPIW values obtained with SCP with the deterministic models and CQR with the QR models differ by not more than one order of magnitude for each MPR. Yet, for autoconversion and rain freezing, the NMPIW obtained with SCP is smaller by up to two and three orders of magnitude, respectively. Moreover, a higher PICP does not necessarily imply a larger or smaller NMPIW, as can be seen, e.g., for the rain freezing and the total riming rate, where the smallest NMPIW values correspond to neither the highest nor lowest PICP.

We visualize the prediction intervals derived with SCP and CQR exemplarily for autoconversion in Figure 3. (Additional results for all process rates are included in the Supplementary Material C.) It is apparent that the SCP results mainly depend on $Q_{1-\alpha}$. For true values smaller than $Q_{1-\alpha}$, all prediction intervals have the same width and mostly cover the true value. Above $Q_{1-\alpha}$, this is not the case anymore, with prediction intervals with decreasing PICP typically missing the true value of the MPR. In contrast, with CQR, valid prediction intervals are present across the full range of true values. This indicates that in our case, single-valued evaluation metrics might not be reflective of the true behavior. To further investigate the dependence of the PICP and NMPIW on the size range of the target variables, in

Table 3. *Normalized mean prediction interval width (NMPIW)*

	Autoconversion	Accretion	Rain evap.	Rain fr.	Rain melt.	Tot. riming
SCP, RF	4.89×10^{-6}	1.23×10^{-4}	8.64×10^{-4}	$\mathbf{1.66 \times 10^{-8}}$	2.79×10^{-3}	4.22×10^{-4}
SCP, XGB	4.29×10^{-6}	$\mathbf{4.88 \times 10^{-5}}$	3.21×10^{-4}	2.97×10^{-8}	$\mathbf{9.04 \times 10^{-4}}$	$\mathbf{1.38 \times 10^{-4}}$
SCP, NN	$\mathbf{4.06 \times 10^{-6}}$	7.31×10^{-5}	$\mathbf{1.80 \times 10^{-4}}$	2.96×10^{-8}	1.08×10^{-3}	1.94×10^{-4}
CQR, QRF	$\mathbf{8.86 \times 10^{-5}}$	$\mathbf{1.08 \times 10^{-4}}$	7.49×10^{-4}	5.62×10^{-5}	$\mathbf{1.48 \times 10^{-3}}$	$\mathbf{4.62 \times 10^{-4}}$
CQR, QXGB	1.12×10^{-4}	3.17×10^{-4}	$\mathbf{7.36 \times 10^{-4}}$	7.46×10^{-5}	2.30×10^{-3}	6.28×10^{-4}
CQR, QNN	1.10×10^{-4}	1.21×10^{-4}	9.13×10^{-4}	$\mathbf{3.60 \times 10^{-5}}$	1.50×10^{-3}	7.16×10^{-4}

Note: For each process rate, the smallest NMPIW obtained with SCP and CQR, is highlighted in bold.

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

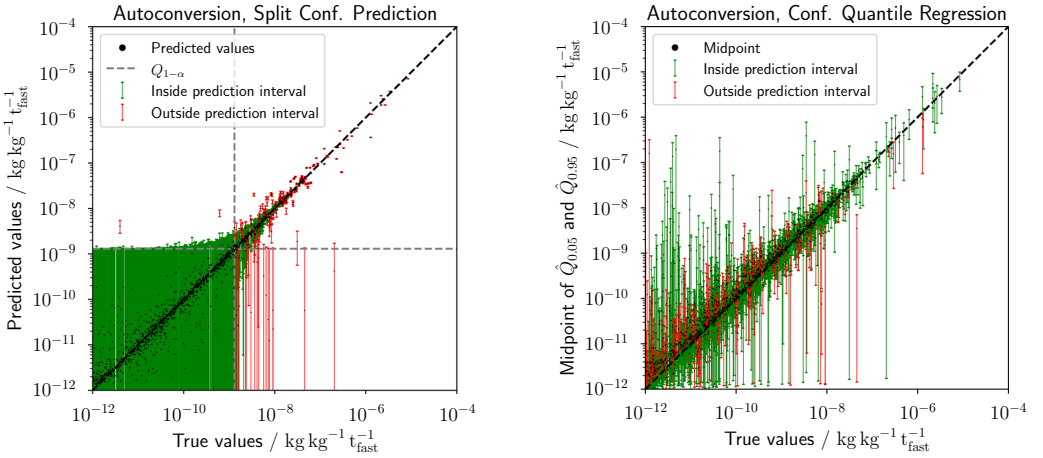


Figure 3. *Calibrated prediction intervals with SCP and the NN (left) and CQR and the QNN (right) for autoconversion. For better visualization, we only show 1500 randomly selected samples.*

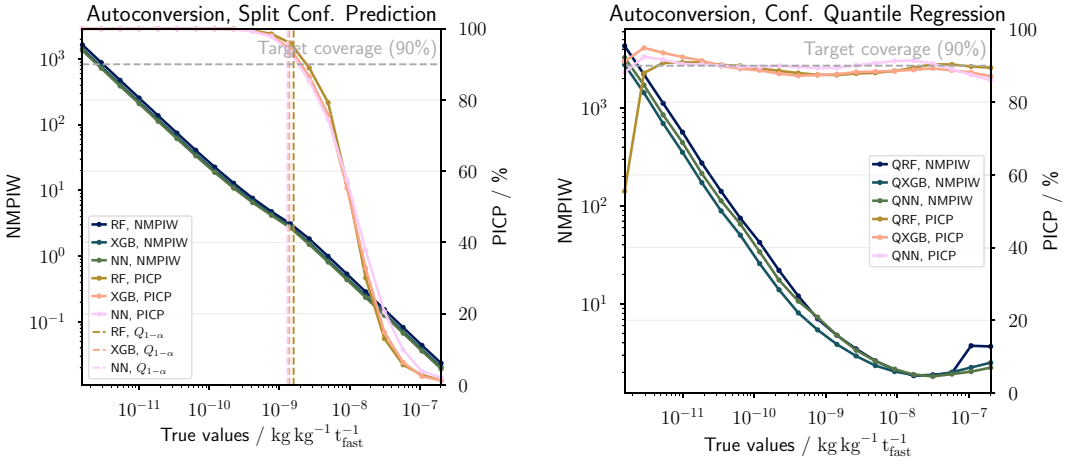


Figure 4. *Normalized mean prediction interval width (NMPIW) and prediction interval coverage probability (PICP) binned by value of the autoconversion rate for SCP (left) and CQR (right).*

Figure 4, we show both quantities binned by the size of the true value for SCP and CQR, respectively. Consistent with Figure 3, we observe that with SCP, the PICP is almost at 100% for size bins of the true values below $Q_{1-\alpha}$ and drops to PICP $\ll 90\%$ above $Q_{1-\alpha}$. With CQR, the PICP is approximately 90% across all size bins, in agreement with Figure 3. With both methods, the NMPIW decreases with increasing size of the target value. Thus, unlike SCP, CQR can predict state-dependent intervals across the entire range of MPRs, yielding more reliable prediction intervals in general. (Supplementary Material C.3 includes an analysis of the absolute residuals, which is a diagnostic for heteroscedasticity.)

5. Conclusions

We applied and evaluated conformal methods to derive calibrated prediction intervals for microphysical process rates (MPRs), enabling the retrieval of detailed process information from high-resolution atmospheric simulations. We compared split conformal prediction, which is simple and inexpensive,

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

with conformalized quantile regression, which requires training a quantile regression model. Applied to well-performing machine learning emulators, both approaches yield satisfactory results for calibration (measured with PICP) and sharpness (measured with NMPIW) on average. In certain cases, the PICP falls slightly below the nominal level, which might indicate mild violations of the exchangeability assumption. Moreover, split conformal prediction exhibits a strongly bin-dependent PICP and unreliable intervals for process rate values exceeding $Q_{1-\alpha}$. In contrast, conformalized quantile regression adapts to heteroscedasticity and thus better reflects the distribution of MPRs, making it the preferable approach when large values are most important. Future work could explore alternative nonconformity scores (Papadopoulos et al., 2008; Lei et al., 2018; Gopakumar et al., 2026) and loss functions for quantile regression training (e.g., Feldman et al., 2021) to further improve efficient, reliable uncertainty quantification across the diverse distributions encountered in weather and climate science.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2026.10051.

Acknowledgments. The authors gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT (NHR@KIT). This center is jointly supported by the Federal Ministry of Education and Research and the Ministry of Science, Research, and the Arts of Baden-Württemberg, as part of the National High-Performance Computing (NHR) joint funding program (https://www.nhr-verein.de/en/our-partners). HoreKa is partly funded by the German Research Foundation (DFG). The GitHub Copilot extension for Visual Studio Code and ChatGPT-5.2 was used to support the development of the model code. Generative AI tools have only been used for proofreading; not for text preparation. Supplementary Figures 4, 7, 8, 9, and 10 make use of the Scientific color maps 7.0 (Crameri et al., 2020). Figures 1 and 2 were created using the LaTeX package Ti*k*Z (Tantau, 2025).

Author contribution. Conceptualization: M.S., T.B., C.H.; Data curation: M.S.; Data visualization: M.S.; Methodology: M.S.; Writing original draft: M.S.; Writing review & editing: M.S., T.B., C.H. All authors approved the final submitted draft.

Data availability statement. The released version of the data and code necessary to reproduce the manuscript's figures are publicly available on Zenodo at https://zenodo.org/records/19114006 (DOI: https://doi.org/10.5281/zenodo.19114006) and in the GitHub repository https://github.com/miriamsimm/ConformalMPR. The trained models as well as the ICON model simulation output are available upon request. The newest version of the source code that was used to perform the simulations with the ICON model can be found at icon-model.org.

Funding statement. M. S. and C. H. acknowledge funding from the NHR Call for Collaboration Project MICRO. C. H. acknowledges funding from the European Union's Horizon Europe Programme under Grant Agreement No. 101137639 (CleanCloud). T. B. acknowledges funding from the Swiss State Secretariat for Education, Research and Innovation (SERI) for the Horizon Europe project AI4PEX (Grant agreement ID: 101137682 and SERI no 23.00546).

Competing interests. The authors declare none.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Provenance statement. This article is part of the Climate Informatics 2026 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer review process.

References

Angelopoulos AN and Bates S (2022) *A gentle introduction to conformal prediction and distribution-free uncertainty quantification.* https://doi.org/10.48550/arXiv.2107.07511.

Behrens G, **Beucler T**, **Iglesias-Suarez F**, **Yu S**, **Gentine P**, **Pritchard M**, **Schwabe M and Eyring V** (2025) Simulating atmospheric processes in earth system models and quantifying uncertainties with deep learning multi-member and stochastic parameterizations. *Journal of Advances in Modeling Earth Systems 17*(4). https://doi.org/10.1029/2024MS004272; Available at https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2024MS004272.

Boucher O, **Randall D**, **Artaxo P**, **Bretherton C**, **Feingold G**, **Forster P**, **Kerminen VM**, **Kondo Y**, **Liao H**, **Lohmann U**, **Rasch P**, **Satheesh SK**, **Sherwood S**, **Stevens B and Zhang XY** (2014) Clouds and aerosols. In *Climate Change 2013 – The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, pp. 571–658. https://doi.org/10.1017/CBO9781107415324.016.

Chen T and Guestrin C (2016) XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: ACM, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Christensen HM, **Kouhen S**, **Miller G and Parthipan R** (2024) Machine learning for stochastic parametrization. *Environmental Data Science 3*, e38. https://doi.org/10.1017/eds.2024.45.

Crameri F, **Shephard GE and Heron PJ** (2020) The misuse of colour in science communication. *Nature Communications 11*(1), 5444. https://doi.org/10.1038/s41467-020-19160-7.

Feldman S, **Bates S and Romano Y** (2021) *Improving Conditional Coverage via Orthogonal Quantile Regression.* arXiv: 2106.00394 [cs.LG] . Available at https://arxiv.org/abs/2106.00394.

Gagne II DJ, **Christensen HM**, **Subramanian AC and Monahan AH** (2020) Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz '96 model. *Journal of Advances in Modeling Earth Systems 12*(3). https://doi.org/10.1029/2019MS001896; :Available at https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001896.

Gammerman A, **Vovk V and Vapnik V** (1998) Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. Madison, WI: Morgan Kaufmann Publishers Inc, pp. 148–155.

Gettelman A, **Morrison H and Thompson G** (2019) Cloud microphysics across scales for weather and climate. In Randall DA, Srinivasan J, Nanjundiah RS and Mukhopadhyay P (eds), *Current Trends in the Representation of Physical Processes in Weather and Climate Models*. Singapore: Springer Singapore, pp. 71–94. https://doi.org/10.1007/978-981-13-3396-5_4.

Gibbs I and Candès E (2021) *Adaptive Conformal Inference Under Distribution Shift.* arXiv: 2106.00170 [stat.ME]. Available at https://arxiv.org/abs/2106.00170.

Gopakumar V, **Gray A**, **Oskarsson J**, **Zanisi L**, **Giles D**, **Kusner MJ**, **Pamela S and Deisenroth MP** (2026) Uncertainty quantification of surrogate models using conformal prediction. *Machine Learning: Science and Technology 7*(1). https://doi.org/10.1088/2632-2153/ae2e7b.

Haynes K, **Lagerquist R**, **McGraw M**, **Musgrave K and Ebert-Uphoff I** (2023) Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artificial Intelligence for the Earth Systems 2*(2), 2769–7525. https://doi.org/10.1175/AIES-D-22-0061.1.

Johnson RA (2024) Quantile-forest: A python package for quantile regression forests. *Journal of Open Source Software 9*(93), 5976. https://doi.org/10.21105/joss.05976.

Khain AP, **Beheng KD**, **Heymsfield A**, **Korolev A**, **Krichak SO**, **Levin Z**, **Pinsky M**, **Phillips V**, **Prabhakaran T**, **Teller A**, **van den Heever SC and Yano JI** (2015) Representation of microphysical processes in cloud-resolving models: Spectral (bin) microphysics versus bulk parameterization. *Reviews of Geophysics 53*(2), 247–322. https://doi.org/10.1002/2014RG000468. Available at https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014RG000468.

Kingma DP and Ba J (2017) *Adam: A Method for Stochastic Optimization.* arXiv: 1412.6980 [cs.LG] . Available at https://arxiv.org/abs/1412.6980.

Koenker R and Bassett G (1978) Regression quantiles. *Econometrica 46*(1), 33–50. http://www.jstor.org/stable/1913643.

Lamb KD, **Singer CE**, **Loftus K**, **Morrison H**, **Powell M**, **Ko J**, **Buch J**, **Hu AZ**, **van Lier Walqui M and Gentine P** (2026) Perspectives on systematic cloud microphysics scheme development with machine learning. *Journal of Advances in Modeling Earth Systems 18*(1), e2025MS005341. https://doi.org/10.1029/2025MS005341.

Lei J and Wasserman L (2012) *Distribution Free Prediction Bands.* arXiv: 1203.5422 [stat.ME] . Available at https://arxiv.org/abs/1203.5422.

Lei J, **G'Sell M**, **Rinaldo A**, **Tibshirani RJ and Wasserman L** (2018) Distribution-free predictive inference for regression. *Journal of the American Statistical Association 113*(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116.

Mansfield LA and Christensen HM (2025) *Epistemic and Aleatoric Uncertainty Quantification inWeather and Climate Models.* arXiv: 2511.23448 [physics.ao-ph] . Available at https://arxiv.org/abs/2511.23448.

Meinshausen N (2006) Quantile regression forests. *Journal of Machine Learning Research 7*(35), 983–999. Available at http://jmlr.org/papers/v7/meinshausen06a.html.

Morrison H, **Van Lier-Walqui M**, **Fridlind AM**, **Grabowski WW**, **Harrington JY**, **Hoose C**, **Korolev A**, **Kumjian MR**, **Milbrandt JA**, **Pawlowska H**, **Posselt DJ**, **Prat OP**, **Reimel KJ**, **Shima SI**, **Van Diedenhoven B and Xue L** (2020) Confronting the challenge of Modeling cloud and precipitation microphysics. *Journal of Advances in Modeling Earth Systems 12*(8), e2019MS001689. https://doi.org/10.1029/2019MS001689.

Mortier T, **Decancq C**, **Sale Y**, **Javanmardi A**, **Waegeman W**, **Hüllermeier E and Miralles DG** (2025) Valid prediction intervals for weather forecasting with conformal prediction. *EGU General Assembly 2025*, Vienna, Austria, 27 Apr–2 May 2025, EGU25–17783. https://doi.org/10.5194/egusphere-egu25-17783.

Palmer TN (2019) Stochastic weather and climate models. *Nature Reviews Physics 1*(7), 463–471. https://doi.org/10.1038/s42254-019-0062-2.

Papadopoulos H, **Gammerman A and Vovk V** (2008) Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the 26th IASTED International Conference on Artificial Intelligence and Applications*. Innsbruck, Austria: ACTA Press, pp. 64–69.

Paszke A, **Gross S**, **Massa F**, **Lerer A**, **Bradbury J**, **Chanan G**, **Killeen T**, **Lin Z**, **Gimelshein N**, **Antiga L**, **Desmaison A**, **Köpf A**, **Yang E**, **DeVito Z**, **Raison M**, **Tejani A**, **Chilamkurthy S**, **Steiner B**, **Fang L**, **Bai J and Chintala S** (2019) *PyTorch: An Imperative Style, High-Performance Deep Learning Library.* arXiv: 1912.01703 [cs.LG] . Available at https://arxiv.org/abs/1912.01703.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V**, **Vanderplas J**, **Passos A**, **Cournapeau D**, **Brucher M**, **Perrot M and Duchesnay E** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press

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** (2025) *DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System.* Ver 2.5.2. German Weather Service (DWD). Available at https://www.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/icon/icon_dbbeschr_aktuell.pdf?view=nasPublication&nn=346850.

Romano Y, **Patterson E and Candes E** (2019) Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, Vol. *32*. Curran Associates, Inc.

Schreck JS, **Gagne DJ**, **Becker C**, **Chapman WE**, **Elmore K**, **Fan D**, **Gantos G**, **Kim E**, **Kimpara D**, **Martin T**, **Molina MJ**, **Przybylo VM**, **Radford J**, **Saavedra B**, **Willson J and Wirz C** (2024) Evidential deep learning: Enhancing predictive uncertainty estimation for earth system science applications. *Artificial Intelligence for the Earth Systems 3*(4), 230093. https://doi.org/10.1175/AIES-D-23-0093.1; Available at https://journals.ametsoc.org/view/journals/aies/3/4/AIES-D-23-0093.1.xml.

Seifert A and Beheng KD (2006) A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description. *Meteorology and Atmospheric Physics 92*(1-2), 45–66. https://doi.org/10.1007/s00703-005-0112-4.

Sun S (2022) *Conformal Methods for Quantifying Uncertainty in Spatiotemporal Data: A Survey.* arXiv: 2209.03580[cs.AI] . Available at https://arxiv.org/abs/2209.03580.

Tantau T (2025) The TikZ and PGF Packages. Manual for version 3.1.11a. Available at https://github.com/pgf-tikz/pgf.

Vovk V, **Gammerman A and Shafer G** (2022) *Algorithmic Learning in a RandomWorld, 2nd*. Cham: Springer. https://doi.org/10.1007/978-3-031-06649-8.

Xu C and Xie Y (2021) Conformal prediction interval for dynamic time-series. In Meila M and Zhang T (eds), *Proceedings of the 38th International Conference on Machine Learning.139*. Proceedings of Machine Learning Research. PMLR, 11559–11569. Available at https://proceedings.mlr.press/v139/xu21h.html.

Zängl G, **Reinert D**, **Rípodas P and Baldauf M** (2015) TheICON(ICOsahedral non-hydrostatic) modelling framework ofDWDandMPI-M: Description of the non-hydrostatic dynamical Core. *Quarterly Journal of the Royal Meteorological Society 141*(687), 563–579. https://doi.org/10.1002/qj.2378.

Zelinka MD, **Myers TA**, **McCoy DT**, **Po-Chedley S**, **Caldwell PM**, **Ceppi P**, **Klein SA and Taylor KE** (2020) Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters 47*(1), e2019GL085782. https://doi.org/10.1029/2019GL085782; Available at https://doi.org/10.1029/2019GL085782.

Cite this article: Simm M, Hoose C and Beucler T (2026). Calibrated conformal prediction intervals for microphysical process rates. *Environmental Data Science*, 5: e14. doi:10.1017/eds.2026.10051

https://doi.org/10.1017/eds.2026.10051 Published online by Cambridge University Press