



Kent Academic Repository

Kallarappayi, A., Nixon, Jonathan, Bégin-Drolet, A. and Godreau, C. (2026) *Machine-and deep-learning models for wind turbine icing prediction across multiple horizons: the influence of ice sensors and weather forecasts*. Cold Regions Science and Technology, 248 . ISSN 0165-232X.

Downloaded from

<https://kar.kent.ac.uk/115547/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1016/j.coldregions.2026.104940>

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

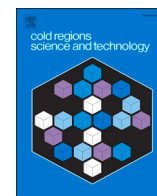
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).



Machine- and deep-learning models for wind turbine icing prediction across multiple horizons: the influence of ice sensors and weather forecasts

A. Kallarappayi^a, J.D. Nixon^{b,*}, A. Bégin-Drolet^c, C. Godreau^d

^a Coventry University, Priory St, Coventry CV1 5FB, UK

^b University of Kent, Canterbury, Kent CT2 7NZ, UK

^c Université Laval's 2255, rue de l'Université, Québec G1V 0A6, Canada

^d Nergica 70, rue Bolduc, Gaspé, QC G4X 1G2, Canada

ARTICLE INFO

Keywords:

Atmospheric icing
Icing protection system (IPS)
Numerical weather prediction
SCADA
Artificial Intelligence (AI)
Ice accretion

ABSTRACT

Around one in four onshore wind turbines operates in cold-climate regions – representing more than three times the global capacity of offshore wind – where annual energy production can be significantly reduced due to ice accretion. Despite growing interest in machine learning for turbine-icing prediction, previous studies have been mostly limited to SCADA data and short-term forecasts, offering little value for operational decision-making. We examine how newly available icing-sensor measurements and liquid water content (LWC) forecasts, together with AI model design, expand what is now possible in turbine-icing prediction. Using one winter (December 2023 to April 2024) of high-resolution data from a single cold-climate wind turbine, including SCADA, on-site meteorological data, icing measurements and weather forecast variables, five models were benchmarked: three ensemble learners (Random Forest, XGBoost, LightGBM) and two deep-learning architectures (CNN-LSTM, GRU) across three prediction horizons (1 h, 6 h, 24 h) and four input-richness configurations. For the studied site and winter season, the results reveal a clear pattern: physically meaningful inputs, rather than network depth, dominate predictive performance. At the 1 h horizon, adding icing indicators and LWC improved mean F1-score from 0.76 to 0.83, reflecting the benefit of direct microphysical information. At 6 h and 24 h, incorporating forecast-based inputs produced the largest gains, increasing mean F1 from 0.46 to 0.70 and from 0.41 to 0.64, respectively. Once supplied with rich physical inputs, lightweight ensemble models matched the accuracy of deep networks, supporting efficient edge-AI deployment directly at turbine level. These findings show that emerging ice-sensor and forecast data expand what is possible in AI-based icing prediction. They create new opportunities for short-horizon icing protection system (IPS) control through earlier intervention, and for more informed day-ahead market participation through better anticipation of icing-related production losses in cold climates.

1. Introduction

Approximately one in four onshore wind turbines worldwide operate under cold-climate conditions, representing around 300 GW of installed capacity – more than three times the global capacity of offshore wind (Karlsson, 2026). In cold climates, ice formation on wind turbine blades is more than just a seasonal nuisance, it's a serious operational challenge. When ice builds up, it reduces aerodynamic efficiency, increases mechanical wear, and can even force the turbines to shut down entirely. These effects can lead to annual energy production (AEP) losses of 10–20%, and in some cases more than 24% has been reported (Parent and Ilinca, 2011; Battisti, 2015; Dai et al., 2024b). Nordic test sites have

shown turbines losing over 200 h of production each winter due to icing (Kreutz et al., 2019; Parent and Ilinca, 2011). This uncertainty has direct economic consequences too: depending on the market structure, operators may be unable to deliver the electricity volumes they have committed and can face imbalance costs or the need to buy replacement power, often during periods of high prices (Norconsult, 2024). Icing also presents safety challenges: detached ice fragments can become hazards, prompting shutdowns under safety regulations (Afzal and Virk, 2018; Kreutz et al., 2019).

To address the icing problem, wind operators typically rely on anti-icing or de-icing strategies to remove ice and avoid shutdowns, and most wind turbine manufacturers now offer Icing Protection Systems (IPS)

* Corresponding author.

E-mail address: j.nixon-104@kent.ac.uk (J.D. Nixon).

<https://doi.org/10.1016/j.coldregions.2026.104940>

Received 24 November 2025; Received in revised form 23 March 2026; Accepted 6 April 2026

Available online 11 April 2026

0165-232X/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

(Stoyanov et al., 2021; Stoyanov and Nixon, 2019). Anti-icing systems try to stop ice before it forms, often by warming blade surfaces or applying special coatings that prevent adhesion. However, active heating systems, like electrothermal or hot-air solutions, typically activate when power output has dropped below an expected threshold (Godreau, 2020; Ischtschuk et al., 2019; Qu et al., 2020), meaning ice is only being removed after it has already formed (i.e. de-icing rather than anti-icing). These setups usually need the turbine to slow or stop while heat is applied to break up or melt the ice (Parent and Ilinca, 2011; Dierer et al., 2011). To avoid unnecessary shutdowns and power losses, it is critical for operators to be able to better forecast ice accretion, so IPS systems can activate and be controlled more efficiently (Godreau, 2020; IEA Wind TCP Task 19, 2017). Whilst short-term forecasts (1–6 h) could improve IPS control, directly improving grid reliability and revenue (Guo and Infield, 2021; Y. Wang and Zhao, 2022), day-ahead predictions (24 h) are equally critical because operators must commit to electricity markets in advance, where misjudgement can lead to penalties and balancing costs (Strauss et al., 2020; Kreutz et al., 2021).

Three ice prediction forecast horizons are considered in this study, corresponding to distinct operational decision contexts. The 1 h horizon supports near-term turbine control actions, such as triggering anti-icing or de-icing systems when ice accretion is imminent. The 6 h horizon aligns with short-term operational planning, where operators balance turbine availability, expected weather evolution, and intervention costs. The 24 h horizon reflects day-ahead market and maintenance planning, where decisions must be made under higher uncertainty and are driven primarily by anticipated atmospheric conditions rather than the current turbine state. These horizons therefore reflect realistic decision-making timescales in cold-climate wind farm operation.

Forecasting approaches generally fall into two main categories. Physics-based models simulate ice accumulation using thermodynamic and droplet-physics principles, but they require detailed meteorological inputs and are often sensitive to errors in weather forecasts (Molinder et al., 2018; Scher and Molinder, 2019). In contrast, data-driven models learn patterns from historical SCADA, meteorological, and icing data to predict ice accretion events (Ge et al., 2017; Xie et al., 2021; Chang et al., 2023). While physics-based models provide valuable physical insight, their performance depends strongly on data availability. This has led to a growing emphasis on data-driven forecasting, where model success depends on both the richness of input data and the learning capacity of the algorithm.

To build reliable data-driven forecasts, operators need more than just SCADA data. SCADA provides valuable indicators of turbine behaviour but offers limited foresight. For predictions extending beyond a few hours, numerical weather prediction data becomes indispensable (Strauss et al., 2020; Kreutz et al., 2023). Yet, some key variables, such as liquid water content (LWC), a major factor in icing formation (Makkonen, 2000; Cao and Hou, 2016), have only recently become available in some weather forecast models (Thompson and Eidhammer, 2014; Benjamin et al., 2016; Bernier et al., 2020). LWC essentially reflects how much ‘fuel’ there is for ice to grow, but continuous measurements are rare at sites. Consequently, many models still rely mainly on SCADA and basic meteorological inputs (Wang et al., 2021; Kreutz et al., 2023) and most machine-learning research to date has focused on short-term ice prediction, with limited emphasis on reliable day-ahead models (Roberge et al., 2023; Dai et al. 2024). Newly developed specialised sensors, such as IC-1 from Ictek (Roberge et al., 2022), can now provide valuable additional details like LWC and icing onset, offering richer data for predictive modelling.

1.1. Research objectives

This study examines how input informativeness and model architecture jointly influence wind turbine icing prediction across operationally relevant forecast horizons. Specifically, the study addresses the following objectives:

1.1.1. Horizon-dependent value of input sources

To quantify how the predictive contribution of different input families, namely SCADA and on-site meteorology, icing sensor measurements, and forecast-derived atmospheric variables, varies across short-term (1 h), medium-term (6 h), and day-ahead (24 h) forecasting horizons. This objective is evaluated through controlled comparisons of progressively enriched input configurations using F1-score, balanced accuracy, and PR-AUC.

1.1.2. Inputs versus model complexity

To assess whether performance gains obtained through input enrichment outweigh those achieved by increasing model complexity, by benchmarking lightweight ensemble models (RF, XGBoost, LightGBM) against recurrent deep learning architectures (CNN-LSTM, GRU) under identical leak-safe evaluation conditions across all horizons.

1.1.3. Horizon-dependent drivers of predictive skill

To identify how the dominant predictors of icing risk change with forecast lead time, distinguishing between short-horizon microphysical indicators, such as icing sensors and liquid water content, and longer-horizon forecast-derived atmospheric drivers, using model-specific feature attribution analyses.

Together, these objectives provide a structured framework for determining which input sources are most informative at each forecast horizon and the extent to which modelling complexity influences predictive performance, with direct relevance for operational deployment in cold-climate wind farms.

1.2. Contributions

This study makes two distinct contributions:

C1. Methodological contribution: controlled multi-horizon benchmarking.

A unified, leak-safe benchmarking framework is developed to compare machine-learning and deep-learning models across multiple forecast horizons. The framework integrates blocked time-series cross-validation, strict temporal alignment of forecast inputs, train-only pre-processing, and consistent threshold selection, enabling fair and reproducible comparison of model classes under realistic operational constraints.

C2. Empirical contribution: incremental value of physics-informed inputs.

Through systematic input enrichment, the study quantifies the marginal contribution of icing sensors and forecast-derived microphysical variables to icing prediction skill. The results show that microphysical measurements primarily improve short-horizon forecasts, whereas forecast-derived atmospheric inputs become increasingly influential at medium and day-ahead horizons. Once such inputs are available, lightweight ensemble models achieve predictive performance comparable to recurrent deep learning architectures, highlighting the central role of input informativeness relative to architectural complexity within the scope of the studied setting.

The next section reviews previous work on wind turbine icing forecasting and identifies the need for a systematic evaluation of input informativeness. Section 3 describes the dataset, input configurations, model architectures, and evaluation procedure, highlighting the controlled, leak-safe design used to ensure fair comparison across forecast horizons and model classes. Section 4 presents and analyses the results, examining how the inclusion of icing sensor and forecast-derived meteorological variables influences forecasting skill relative to model complexity, and how the most informative predictors vary with forecast lead time. Operational implications, key limitations, and directions for future research are also discussed. Section 5 concludes the paper.

2. Literature review

Research on wind-turbine icing has evolved along three major paths: physics-based modelling, classical machine learning (ML), and deep learning (DL). Each tackles the same fundamental challenge from a different angle; physics-based approaches model how ice forms, ML techniques detect patterns in operational data, and DL methods attempt to capture complex time-based relationships in those patterns. Yet, despite progress, relatively few studies demonstrate reliable, turbine-level forecast performance across seasons or at different sites under real-world constraints.

Early work leaned heavily on physical theories. Foundational models like Messinger's equilibrium approach (Messinger, 1953) and Makkonen's accretion theory (Makkonen, 2000) identified variables such as liquid water content (LWC), droplet size, and heat transfer as central to ice growth. Later studies refined this further, distinguishing between types of ice and incorporating energy balance equations (Finstad et al., 1988; Jones, 1990). While these models offer deep insight into icing mechanisms, they rely on microphysical inputs that are rarely available at operational wind farms and can be sensitive to weather variables (Molinder et al., 2018; Strauss et al., 2020). Even simplified versions – while reproducible in labs – can be hard to calibrate and computationally demanding. These models can explain why icing happens quite well, but struggle to predict when or how much icing will occur in practice.

As data availability from wind farms has improved, attention has shifted toward ML techniques that learn directly from SCADA and weather data. Early approaches used tools like logistic regression, support vector machines, and random forests to predict icing from turbine power, wind speed, and temperature (Skrimpas et al., 2019; Zhou et al., 2018). Ensemble models later helped improve robustness (Tao et al., 2021; Ding et al., 2021), although many still relied on proxy labels – such as deviations in the power curve – that could also reflect unrelated factors like derating or yaw errors (Lee et al., 2020; Ye and Ezzat, 2024). Forecast-oriented studies have shown that incorporating numerical weather prediction data can enhance early warning by capturing larger-scale weather patterns associated with icing onset (Molinder et al., 2018; Strauss et al., 2020). Separately, field investigations emphasise the physical importance of liquid water content for ice accretion and demonstrate that direct LWC measurements provide valuable microphysical context when such sensors are deployed (Roberge et al., 2021, 2023).

Deep learning algorithms has shown a stronger ability to predict ice formation on wind turbine blades. LSTM networks and hybrid CNN-LSTM models have generally outperformed traditional ML methods for short-term predictions, especially by modelling the temporal behaviour of SCADA signals (Guo and Infield, 2021; Y. Wang and Zhao, 2022; Chang et al., 2023). Temporal convolutional networks (TCNs) have extended forecasting horizons further, showing credible multi-hour performance using SCADA alone (Zhang et al., 2024). Meanwhile, attention-based networks are being explored for icing diagnostics (J. Guo et al., 2024). However, DL models come with tradeoffs: they require large, well-labeled datasets, demand significant computing power, and often function as 'black boxes' that are hard to interpret. Moreover, relatively few studies test how these models perform when conditions change over time or across sites.

In recent years, hybrid and physics-informed models have started to emerge, integrating physical laws into learning architectures. The TIGER model by Ye and Ezzat (Ye and Ezzat, 2024) is one such example, using a physics-aware loss function that aligns predictions with actual power loss. Other efforts embed thermodynamic and droplet-flux equations into neural network structures (Tao et al., 2023; Jiang et al., 2023). These designs aim to balance adaptability with interpretability, though they have mostly been tested on small or single-season datasets, and comparisons with baseline models are still limited.

Field measurements continue to play a vital role in grounding predictive models with reality. Years of observations confirm that LWC and

droplet size are key factors in how fast and how much ice accumulates (Makkonen, 2000; Finstad et al., 1988; Jones, 1990; Parent and Ilinca, 2011). Field studies suggest that variability in LWC explains much of the difference between icing events, while sensor placement and calibration can significantly affect reliability (Roberge et al., 2021, 2023). When cross-checked with turbine performance, nacelle-mounted ice alarms often align closely with power-curve-based icing signals (Roberge et al., 2023). Combining ice sensor inputs with SCADA and weather forecasts is, therefore, seen as a promising option to improve ice formation predictions.

Several reviews have highlighted the challenges that remain for AI-based ice predictions. Cai et al. (2023) pointed out common issues like noisy labels, unbalanced datasets, and poor generalization. Dai et al. (2024a) echoed these concerns, noting that many ML and DL models perform well only under specific conditions, often failing when applied to new sites. Both reviews stress that while algorithm development has advanced rapidly, improvements in data quality and representation have lagged. In many cases, the choice of inputs is driven more by convenience than by strategy. In summary, physics-based models offer clarity but need more testing; ML models are efficient but often reactive, and DL models are powerful yet complex and resource-heavy. Across the board, differences in data coverage, input selection, and validation methods make direct comparisons between studies difficult. Reported F1-scores in the literature typically range from 0.6 to 0.85, depending on how rich the inputs are and how long the dataset spans (Guo and Infield, 2021; Y. Wang and Zhao, 2022; Chang et al., 2023; Dai et al., 2024b). However, many of these models don't transfer well beyond the environments they were trained in (Cai et al., 2023).

This study uses a single-site benchmark to isolate the role of inputs versus model complexity across 1, 6, and 24 h horizons. By focusing on deterministic forecasting in a controlled setting, the study provides a clean test of whether enriched inputs can enable lightweight ML to achieve accuracy comparable to DL, while clarifying the operational trade-offs that shape deployment.

3. Methodology

This section presents the data preparation, input configurations, forecasting models, model training, and evaluation protocol adopted in this study. To ensure a fair comparison across model classes and horizons, all pipelines follow identical leak-safe handling: no look-ahead, strictly chronological splits, train-only normalisation, and a single operating threshold per model-horizon tuned on validation data. Fig. 1 provides a conceptual overview of the modelling framework and input variants, while Fig. 2 details the leak-safe preprocessing, forecast alignment and evaluation pipeline used to implement this framework.

Fig. 2 summarises the complete leak-safe forecasting pipeline used in this study, including causal missing-data handling, valid-time forecast alignment, blocked time-series cross-validation with purge gaps, and the order of training, calibration and threshold selection. All preprocessing operations, including feature filtering and scaling, are fitted exclusively on the training data and applied forward in time, while probability calibration and operating-threshold selection are performed only on the validation segment prior to final testing. Inputs are aligned to a continuous 1-min timeline with causal missing-data handling and valid-time forecast matching. Blocked time-series cross-validation with purge gaps separates training, validation and test data. Feature filtering and scaling are fitted on training data only, calibration and threshold selection are performed on validation data, and final metrics are computed on the held-out test set.

3.1. Data preparation

Minute-resolution data from a cold-climate wind turbine in North America were analysed over one winter season (December 2023 – April 2024). The data were collected from a utility-scale wind turbine

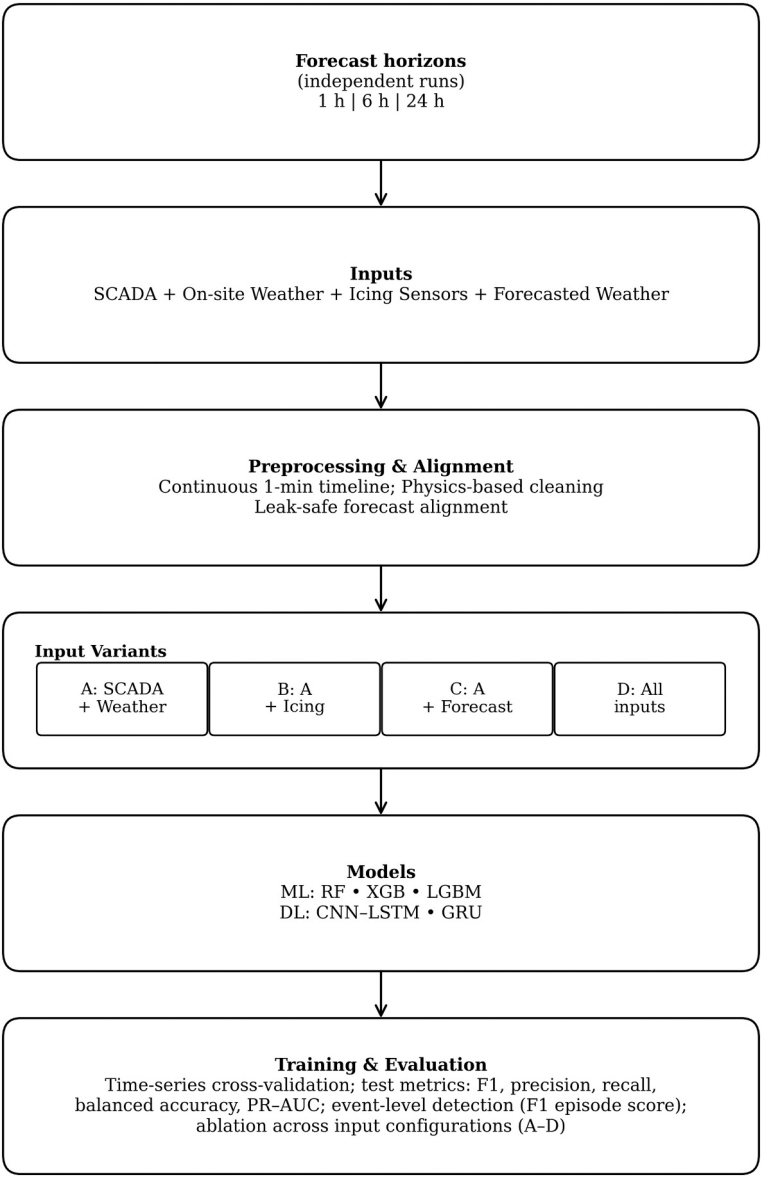


Fig. 1. Overall methodology: independent model runs for forecast horizons of 1 h, 6 h, and 24 h, covering input acquisition, leak-safe preprocessing, input variants (A-D), model classes (ML and DL), and evaluation.

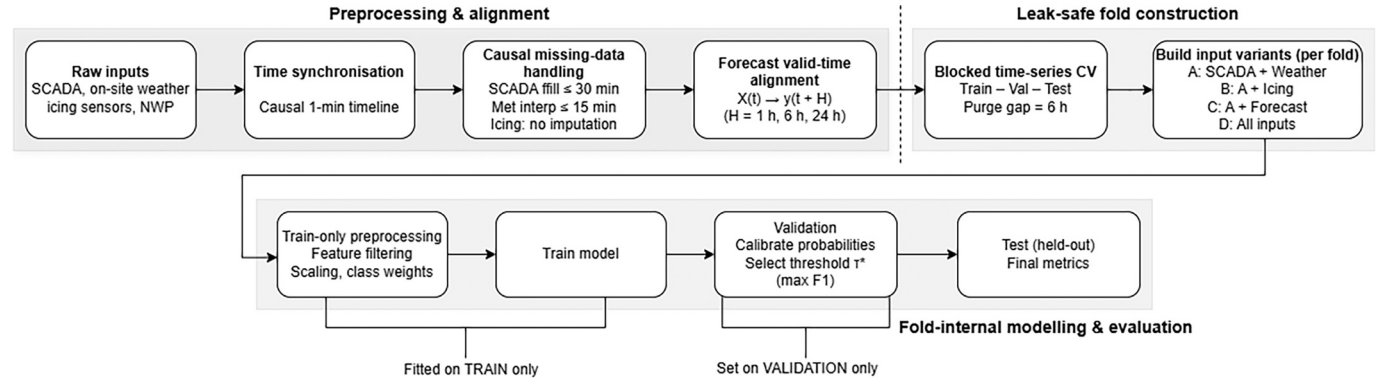


Fig. 2. Leak-safe forecasting and evaluation pipeline.

operating in a cold-climate region of North America. The site experiences sustained sub-zero winter temperatures and recurrent icing conditions during the December–April period analysed in this study. In accordance with commercial confidentiality agreements with the site operator, precise geographic coordinates, detailed terrain mapping, and forecast grid configuration cannot be disclosed. The multivariate dataset comprises: (i) SCADA variables (e.g. power, rotor speed, pitch); (ii) on-site meteorological measurements (e.g. temperature, humidity, pressure, wind, precipitation); (iii) icing-related inputs (binary ice label, LWC, severity, thickness – provided from an Ictek IC-1 sensor); and (iv) horizon-matched forecast variables from the ECCC NWP system (e.g. temperature, specific humidity, wind and LWC). Forecast data were merged only if valid at prediction time using strictly forward temporal alignment, preventing any access to future information.

Forecasted meteorological variables were provided by the site operator and are derived from Environment and Climate Change Canada (ECCC) numerical weather prediction framework configured for operational wind farm use. In addition to standard near-surface atmospheric variables (e.g. air temperature, humidity, wind speed and direction, and precipitation), the forecast dataset included liquid water content (LWC) as an icing-relevant diagnostic variable (<https://eccc-msc.github.io/open-data/>). The forecast inputs represent atmospheric conditions conducive to icing rather than explicit predictions of ice accretion or blade icing state. Details of the site-specific forecast configuration cannot be disclosed due to confidentiality constraints. These forecast variables are used solely as atmospheric input features for icing prediction and are not intended to represent or validate blade-scale icing or accretion physics. The task is framed as binary sequence classification. A 60-min look-back window is used to predict icing at 1 h, 6 h and 24 h ahead. All streams were reindexed to a continuous 1-min grid. Physically implausible values were clipped using limit values based on manufacturer guidance (e.g. temperature – 40 to 50 °C, RH 0–100%, power output less than 0), while icing variables were left unmodified to preserve high-frequency accretion signals.

Missing values were handled using strictly causal procedures. SCADA variables were forward filled (ffill) for up to 30 min to account for routine short-term communication delays and the high persistence of supervisory control signals, such as rotor speed, blade pitch, and turbine temperatures, which typically evolve slowly relative to icing onset. Meteorological variables were interpolated using forward linear interpolation across gaps not exceeding 15 min, ensuring that only past observations were used at each prediction time and that no future information contributed to feature construction. The 15-min limit was chosen as a conservative threshold to minimise the risk of introducing artificial persistence, and longer gaps were excluded rather than interpolated to avoid propagating spurious autocorrelation into the input features. Gaps longer than 15 min were not interpolated; instead, affected samples were excluded through a completeness mask. Icing-related variables were never imputed in order to preserve rapid ice accumulation and persistence dynamics. The completeness mask ensured that only windows with continuous SCADA and meteorological coverage over the preceding 60 min were used as model inputs, preventing artificial smoothing of critical transitions near the freezing point. Data preprocessing for model inputs differed according to model type and was paired with measures to address class imbalance. For deep learning architectures, all numeric channels were standardised using z-scores computed from the training partition only to avoid information leakage. Tree-based models were applied to unscaled inputs, as these algorithms are generally insensitive to feature magnitude and operate effectively on raw values. Over the study period, approximately 15% of minute-level observations corresponded to icing conditions, indicating a moderate imbalance between classes. This was mitigated using class-weight scaling in the loss function and operating-threshold optimisation during validation. To reflect operational conditions more realistically and improve metric stability, event-level grouping was applied during evaluation so that models were assessed on their ability to detect

entire icing episodes rather than isolated timestamps.

3.2. Input configurations

Four input configurations were defined to assess how data richness influences forecasting performance. Table 1 summarises the four input variants, designed to isolate the respective contributions of operational monitoring, icing detection, and forecast-enhanced information. Variant A uses only SCADA and on-site weather data and acts as the operational baseline. Variant B adds icing sensor variables (IC-1 icing detection flag, liquid water content, severity and thickness), enabling investigation of whether direct observations of icing improve short-term detection beyond what can be inferred from operational signals.

Variant C combines SCADA and weather data with horizon-specific forecast inputs aligned in a leak-safe manner. This reflects conditions where operators have access to meteorological forecasts and allows assessment of how far-ahead prediction benefits from anticipating future atmospheric states. Variant D merges sensor-based and forecast-based enrichment, representing a fully informed setting that combines immediate icing indicators with projected weather evolution.

Organising the input variants progressively allows the study to isolate whether performance gains arise from enhanced information content or from the forecasting architecture itself. Each variant defines a deterministic and horizon-specific set of physically meaningful variables, independent of any model-specific feature engineering. The configuration design reflects realistic implementation pathways, starting from standard SCADA-based monitoring and extending to forecast-assisted operational decision-making. Forecast predictors were incorporated using a valid-time alignment rule, such that for each observation timestamp t and forecast horizon H , forecast variables correspond to atmospheric conditions valid at time $t + H$. Redundancy among predictors was controlled during subsequent model-specific preprocessing using correlation-based filtering applied within each cross-validation fold. Final variable sets were stored separately for each configuration and forecast horizon to ensure reproducibility. To limit redundancy among highly collinear predictors, correlation-based feature filtering was applied as a secondary preprocessing step within each cross-validation fold using training data only. Pairwise Pearson correlation coefficients were computed, and when the absolute correlation between two features exceeded 0.9, the feature with the lower univariate association with the target variable, measured on the training set, was removed. Across all model runs, the filter removed between two and six features per configuration and horizon, consistently from the wind-speed and atmospheric pressure cluster; no IC-1 sensor channel was removed in any configuration. The resulting feature set was then applied unchanged to the corresponding validation and test data, ensuring that feature selection did not introduce temporal leakage.

Table 1
Overview of the four input configurations.

Variant	Included variables
A	Turbine operational data (power output, rotor speed, blade pitch, air density) and on-site meteorology (air temperature, relative humidity, barometric pressure, wind speed and direction, precipitation, solar radiation)
B	Variant A plus icing-sensor measurements, including binary icing state, ice thickness, icing severity and measured liquid water content (LWC)
C	Variant A plus horizon-matched meteorological forecasts from the ECCC numerical weather prediction system, including forecast air temperature, specific humidity, wind speed and direction, pressure, precipitation rates, radiation, cloud cover and forecast liquid water content (LWC)
D	Variant B plus horizon-matched ECCC meteorological forecasts, including forecast air temperature, specific humidity, wind speed and direction, pressure, precipitation rates, radiation, cloud cover and forecast liquid water content (LWC)

3.3. Forecasting models

This study evaluates five forecasting models: two deep learning architectures, Convolutional Neural Network–Long Short-Term Memory (CNN-LSTM) and Gated Recurrent Unit (GRU), and three tree-based ensemble methods, Random Forest (RF), Extreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (LGBM). Icing formation is governed by both rapid short-term fluctuations, for example sudden increases in liquid water content, and slower accumulation under sustained sub-zero conditions. To represent these coupled dynamics, the CNN-LSTM model applies temporal convolutions to identify localised patterns and uses an LSTM layer to integrate contextual information across the input sequence. The GRU model is included to probe whether a lighter recurrent architecture can deliver comparable performance with lower computational cost, supporting operational deployment. Random Forest, XGBoost and LightGBM serve as ensemble baselines well suited to SCADA and meteorological data and also provide interpretable feature importance. Together, these models allow a structured comparison between input informativeness and modelling complexity.

3.3.1. CNN-LSTM model

The CNN-LSTM model is formulated, where $\mathbf{X}_{t-W+1:t} \in \mathbb{R}^{W \times F}$ denotes a window of $W = 60$ minutes across F features. A one-dimensional convolutional layer generates feature maps as

$$\mathbf{z}_{t,c} = f_{\text{ReLU}} \left(\sum_{i=0}^{k-1} \sum_{f=1}^F K_{c,f,i} \mathbf{x}_{t+i,f} + b_c \right)$$

where $K_{c,f,i}$ are kernel weights, b_c is a bias term, and $f_{\text{ReLU}}(x) = \max(0, x)$ denotes the rectified linear activation. Here, k represents the temporal kernel width (3–5 min) corresponding to the number of past time steps aggregated by each filter, and all variables are indexed on a one-minute grid. For clarity, $\sigma_{\text{sig}}(\cdot)$ denotes the sigmoid activation and $\odot$ indicates elementwise multiplication.

Two temporal convolutions (64 filters, kernel sizes 3 and 5) with batch normalisation and dropout (0.2) extract local temporal patterns. The resulting sequence of CNN features $\mathbf{Z} \in \mathbb{R}^{W \times 64}$ is passed to a single-layer unidirectional LSTM (hidden size 64), which updates its hidden and cell states as

$$\mathbf{f}_t = \sigma_{\text{sig}}(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_f), \quad \mathbf{i}_t = \sigma_{\text{sig}}(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_i),$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_c), \quad \mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t,$$

$$\mathbf{o}_t = \sigma_{\text{sig}}(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{z}_t] + \mathbf{b}_o), \quad \mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$

The final hidden state $\mathbf{h}_W$ summarises the 60-min context. A dense head transforms $\mathbf{h}_W$ into a scalar logit through

$$\text{Linear}(64 \rightarrow 64) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.2) \rightarrow \text{Linear}(64 \rightarrow 1)$$

followed by a sigmoid activation that outputs the predicted icing probability $\hat{p}_{t+H}$. Model architecture is shown in the online Annex A.

3.3.2. GRU model

In the GRU model, $\mathbf{x}_t$ denotes the feature vector at time t and $\mathbf{h}_{t-1}$ the previous hidden state. In this model, $\sigma_{\text{sig}}(\cdot)$ represents the sigmoid activation, $f_{\text{ReLU}}(\cdot)$ the rectified linear unit, and $\odot$ denotes elementwise multiplication. The GRU cell updates its gates and hidden state as

$$\mathbf{z}_t = \sigma_{\text{sig}}(\mathbf{W}_z[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_z), \quad \mathbf{r}_t = \sigma_{\text{sig}}(\mathbf{W}_r[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_r),$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h[\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_h),$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t$$

A single unidirectional GRU layer with 64 hidden units and dropout of 0.2 encodes the temporal sequence. Its final hidden state $\mathbf{h}_W$ sum-

marises information over the look-back window W . The output head consists of two fully connected layers (64→32→1) with f_{ReLU} activation and dropout, followed by a sigmoid that produces the predicted icing probability $\hat{p}_{t+H}$.

In this study, the GRU operates with a longer look-back window (60 min) and a stride of 5 min, enabling an efficient yet expressive configuration suitable for operational deployment. Model architecture is shown in the online Annex A.

3.3.3. Machine-learning models (RF, XGB and LGBM)

Tree-based ensemble models were included as baseline predictors due to their suitability for structured SCADA and meteorological inputs. These algorithms capture non-linear interactions without requiring explicit feature scaling and are generally robust to measurement noise and moderate multicollinearity, making them appropriate for operational sensor data. Random Forest constructs multiple decorrelated decision trees using bootstrap aggregation, thereby reducing variance and improving generalisability. Both XGBoost and LightGBM employ gradient-boosted tree frameworks that optimise predictions through stage-wise residual learning. XGBoost enhances robustness using second-order gradient information and regularisation, while LightGBM improves computational efficiency through histogram-based binning and leaf-wise growth strategies, enabling faster training on high-dimensional feature sets.

The use of these models provides two principal advantages. First, they act as robust non-temporal baselines against which the benefits of sequential learning in the deep models can be objectively assessed, helping to determine whether forecast skill is driven predominantly by the informativeness of the inputs or by temporal pattern extraction. Second, they provide interpretable feature importance measures, enabling direct examination of which variables contribute most strongly to predictive performance across input variants and horizons, thereby supporting physical interpretation of model behaviour.

Class imbalance was addressed using proportional class weighting based on the ratio of negative to positive samples in the training data. As these models operate on static feature vectors rather than full temporal sequences, they do not capture progression dynamics, which allows their results to isolate the influence of input composition independently of sequence modelling. Hyperparameter optimisation, probability calibration, and threshold selection were conducted following the procedures detailed in Section 3.5 to ensure consistent and leak-free evaluation across all model types.

For tree-based ensemble models, temporal information from the 60-min look-back window was encoded using a fixed-length feature representation. Each numeric variable was summarised using window-level statistics, namely the mean, standard deviation, minimum, and maximum, capturing the persistence and variability of atmospheric and operational conditions over the look-back period. In addition, temporal context was incorporated through lagged features, whereby each variable was represented by its observed values at fixed lag offsets of 10, 30, and 60 min prior to the prediction time. These lagged values capture short-term persistence effects relevant to icing formation while maintaining a controlled feature dimensionality. All lagged and summary features were strictly causal and derived exclusively from observations available at prediction time. No temporal flattening, trend extraction, or explicit sequence modelling was applied to the tree-based models.

3.4. Model training, calibration, and evaluation protocol

This subsection consolidates all leak-safe design elements, including time-series splitting, purge intervals, forecast valid-time alignment, train-only preprocessing and calibration, and fold-safe feature filtering. All models were trained and evaluated using blocked time-series cross-validation to reflect realistic operational deployment. Three non-overlapping temporal folds were defined, each consisting of contiguous training, validation, and test segments ordered chronologically,

with the validation period immediately preceding the test window. A purge interval of 6 h was enforced between adjacent segments to reduce the influence of short-term autocorrelation and prevent information leakage across fold boundaries.

For each forecast horizon H and each cross-validation fold, the modelling workflow followed a fixed and deterministic sequence. First, horizon-specific input variants were constructed using the variable groupings defined in Section 3.2, with forecast predictors aligned to the prediction target using valid-time matching to the corresponding target time (t plus H). When multiple forecast runs provided values for the same target time ($t + H$), only forecasts issued at or before the prediction timestamp t were considered eligible. Among these, the forecast with the most recent issue time prior to t was selected. Forecasts issued after t were excluded. This deterministic rule ensures strict operational causality and reproducible forecast alignment. All preprocessing operations were then fitted using the training segment only and applied unchanged to the validation and test segments. These operations included causal missing-value handling, correlation-based feature filtering, and normalisation for deep learning models. No information from the validation or test periods was used during preprocessing. This leak-safe, time-series-aware evaluation protocol ensures that reported performance reflects true operational forecasting skill rather than artefacts of temporal leakage.

Model training was performed on the processed training data using fixed hyperparameter settings. Hyperparameters were tuned once using the earliest fold only, based exclusively on its training and validation segments, to emulate realistic forecasting practice in which models are configured prior to operational deployment. Following this initial tuning, hyperparameter values were held constant and reused across all subsequent folds without further adjustment.

Deep learning models were trained for up to 30 epochs using mini-batches of size 256 with binary cross-entropy loss and class weighting proportional to the imbalance ratio. The Adam optimiser was used with a learning rate of 0.001 and a weight decay of 0.0001, together with early stopping based on validation F1-score (patience of 5 epochs). Gradient clipping with a global norm of 1.0 was applied to the GRU to improve training stability. Machine-learning models followed the same fold structure and were trained using inverse class-frequency weighting, without feature scaling.

After model fitting, probabilistic outputs from the tree-based ensemble models were calibrated on the validation segment using isotonic regression. Deep learning architectures were not separately calibrated, as they were trained directly under binary cross-entropy loss and evaluated using validation-based threshold optimisation. An operating classification threshold was then selected by maximising F1-score on the validation data and fixed prior to testing. Final performance metrics were computed exclusively on the held-out test segment. At no stage were test observations used during preprocessing, model fitting, calibration, or threshold selection.

All forecast variables were aligned to each prediction timestamp using valid-time matching, as defined in Section 3.2, ensuring that only information corresponding to atmospheric conditions valid at the target time was used. No post-event observations were included during preprocessing. To enhance reproducibility, random seeds were fixed across all software libraries, and deterministic CuDNN (Convolutional Neural Network with CUDA Deep Neural Network library) operations were enabled where supported. Detailed hyperparameter ranges, final tuned values, calibration settings, and reproducibility controls are provided in Annexes B and C to ensure full auditability.

3.5. Evaluation metrics and reproducibility

Model performance was evaluated using time-series cross-validation. For each cross-validation fold, metrics were computed on the held-out test segment, and the final performance values reported in this paper represent the mean across folds. F1-score was selected as the primary

metric due to its relevance in applications where both false alarms and missed detections carry operational consequences. Precision, recall, balanced accuracy and PR-AUC were reported alongside F1 to provide a broader view of classification reliability and the trade-off between conservative and reactive decision making. Metrics assess model performance at individual timestamps and are reported to two decimal places for consistency.

To assess practical detectability, an event-level F1-score was also computed. Icing episodes are defined as continuous time intervals during which the IC-1 icing detection flag indicates sustained icing conditions, grouped to reflect operationally relevant accretion events rather than isolated timestamp-level detections. Consecutive positive predictions were grouped into full icing episodes and matched against observed events based on temporal overlap criteria. This evaluates whether models capture icing onset and duration consistently rather than only isolated moments of detection, which is more relevant for operational decision making.

Consecutive positive readings in both the observed and predicted sequences separated by a gap of ten minutes or fewer were merged into a single episode, reflecting typical sensor response times and avoiding artificial splitting on brief noise. Episodes shorter than five minutes after merging were discarded as spurious detections. This pre-processing was applied symmetrically to both sequences so that identical rules govern the construction of observed and predicted episodes. A predicted episode was counted as a true positive if either: (a) the predicted episode's temporal span overlapped the ground-truth episode by at least 50% of the ground-truth episode's duration, or (b) the predicted onset fell within ± 30 min of the observed onset. Sensitivity to the overlap threshold is demonstrated in Annex Table D5, where model rankings remain consistent at 30% and 70% overlap, confirming the robustness of the reported event-level conclusions.

Uncertainty in all metrics was quantified using block bootstrapping with 1000 resamples, preserving temporal dependence within blocks. Confidence intervals at the 95% level were generated for both timestamp and event-level results. Threshold robustness was assessed by perturbing the selected operating threshold by ± 0.05 and re-evaluating performance to test stability under realistic operational variability. Comparisons between input configurations and model architectures were based on fold-averaged metrics, enabling separation of performance improvements attributable to input quality from those due to modelling depth.

Block bootstrapping was applied to the held-out test segment of each cross-validation fold using contiguous, fixed-length temporal blocks sampled with replacement. A block length of 60 min was selected to match the input window duration and to reflect the typical persistence of icing events and near-surface meteorological conditions. This design preserves short-range temporal dependence within icing episodes while avoiding the independence assumptions of standard bootstrap resampling.

Bootstrapping was performed independently within each fold. For comparisons between input configurations and model architectures, a paired bootstrap was used by applying identical bootstrap resample indices across configurations and models within a fold. This pairing ensures that reported performance differences reflect methodological contrasts rather than sampling variability. Final confidence intervals were obtained by aggregating bootstrap distributions across folds.

Feature importance was assessed using model-appropriate methods. For tree-based ensembles, split-gain-based attribution was computed and averaged across folds. For deep learning architectures, mean absolute SHAP (SHapley Additive exPlanations) values were calculated over all test timestamps and aggregated across the full 60-min input window. This approach reflects the influence of each feature on the prediction irrespective of its exact timing within the sequence, in line with common practice in time-series interpretability. Feature importance was used qualitatively to support physical interpretation of performance trends rather than as a primary comparison criterion.

Reproducibility was ensured by fixing random seeds across all software libraries and enabling deterministic CuDNN operations where available. All preprocessing steps, hyperparameters, data splits, calibrated thresholds and prediction outputs were logged for traceability and experimental consistency.

4. Results and discussion

Model performance is evaluated across three forecast horizons (1, 6, and 24 h) and four input configurations (Variants A–D; Table 1). Results are assessed primarily using the F1-score, which reflects the operational trade-off between missed icing events and false alarms under class imbalance. Threshold-dependent metrics, including precision, recall, and balanced accuracy, are reported alongside threshold-free discrimination performance measured using the precision–recall area under the curve (PR-AUC). Across all horizons, PR-AUC trends were consistent with F1-based rankings, indicating that observed performance differences reflect genuine changes in class separability rather than artefacts of threshold selection. To support physical interpretation beyond aggregate metrics, performance is further analysed using event-based evaluation and model-specific feature importance. Summary results are presented in the main text, with complete metric tables provided in Appendix Tables D1–D3. Illustrative performance comparisons are shown in Figs. 3, 4, and 6, while Figs. 3a–b, 5a–b, and 7a–b examine feature contributions across horizons. While F1-score is used as the primary operating metric in this study, threshold-free performance was evaluated using the precision–recall area under the curve (PR-AUC) to ensure that reported performance gains are not artefacts of a specific decision threshold. Across all models and forecast horizons, configurations incorporating forecast information (Variants C and D) consistently achieved higher PR-AUC than SCADA-only inputs, with the largest improvements observed at the 6 h and 24 h horizons, confirming improved class separability under severe class imbalance. Operating thresholds were optimised using F1-score to balance missed icing events and false alarms, reflecting the operational requirements of early-warning icing systems. Event-level evaluation further demonstrates that improvements in timestamp-based metrics translate into more reliable detection of complete icing episodes, which constitutes the operationally relevant unit for turbine control and mitigation.

Results are presented by forecast horizon to reflect operational relevance. Section 4.1 focuses on the 1-h horizon, which is dominated by short-term microphysical drivers. Section 4.2 examines 6-h forecasting, where near-term atmospheric evolution becomes increasingly influential. Section 4.3 discusses the 24-h horizon, representing day-ahead decision-making under higher uncertainty. Within each section, tree-

based models are discussed alongside deep learning architectures, and findings are interpreted with reference to icing physics and operational use.

4.1. One-hour horizon: high accuracy and incremental gains from enriched inputs

At the 1-h horizon, all models demonstrate strong predictive capability across input configurations, with F1-scores typically ranging from 0.74 to 0.86 (Fig. 3). The highest performance is obtained using Variant D (SCADA, on-site weather, icing sensors and forecast variables), where the GRU achieves an F1 of 0.86, followed closely by Random Forest and CNN-LSTM (both ≈ 0.83). This indicates that while recurrent architectures can capture short-term transitional signals, much of the predictive skill at this horizon is already contained within the input data. The small gap between GRU and well-configured RF (≈ 0.03) suggests that under physically informative inputs, lightweight models can perform competitively.

Performance improves progressively from Variant A to Variant D. Integrating icing sensors (Variant B) yields initial gains (typically +0.05 to +0.07), while the addition of forecast variables (Variant C) provides further refinement. Improvements from Variant C to Variant D are modest ($\approx +0.02$ to $+0.03$ F1) but consistent and associated with reduced classification variability. Complementary metrics presented in online Annex D (PR-AUC and balanced accuracy) support these findings, confirming that higher F1 is accompanied by improved precision–recall trade-off and more balanced class discrimination.

These results suggest that, at short horizons, icing onset is strongly associated with immediate microphysical and thermodynamic conditions, which models are able to detect effectively when such information is explicitly provided. Deep learning offers incremental benefit through temporal pattern extraction but is not essential when features are sufficiently informative. These findings favour operational use of tree-based algorithms for short-term forecasting, particularly in scenarios where computational efficiency and implementation simplicity are prioritised.

To assess whether the 1-h performance gains reflect genuine predictive skill rather than persistence of the current icing state, model results were compared against a naïve persistence baseline (F1 = 0.80). Variant A models perform at or below this baseline (F1 = 0.74–0.80), as expected given the absence of any current icing state information – the model is required to infer future icing conditions from atmospheric precursors alone. In contrast, Variants B and D consistently exceed the persistence baseline (F1 = 0.78–0.84 and 0.81–0.86 respectively), demonstrating that the gains from sensor integration reflect genuine improvement in transition detection rather than exploitation of icing state continuity.

4.1.1. Feature importance analysis

Feature importance from the Random Forest model supports the performance trends observed in the F1 analysis. Under Variant A (SCADA and meteorological variables only), relative humidity emerges as the dominant predictor, accounting for approximately a quarter of the total importance, followed by precipitation, air temperature and atmospheric pressure (Fig. 4a–b). This suggests that in the absence of direct icing-related inputs, the model infers icing risk from atmospheric conditions that are strongly associated with physical icing formation processes.

When icing sensor data are added (Variant B), importance shifts toward microphysical indicators, with ice thickness and icing sensor data jointly contributing around 24–25% of total importance, and icing severity adding a further 7–8% (Fig. 4b). This results in reduced reliance on meteorological proxies (e.g. relative humidity reduces to approximately 14%), indicating that the model no longer infers icing onset but is able to detect it directly. The alignment between performance gains in Variant B and changes in feature

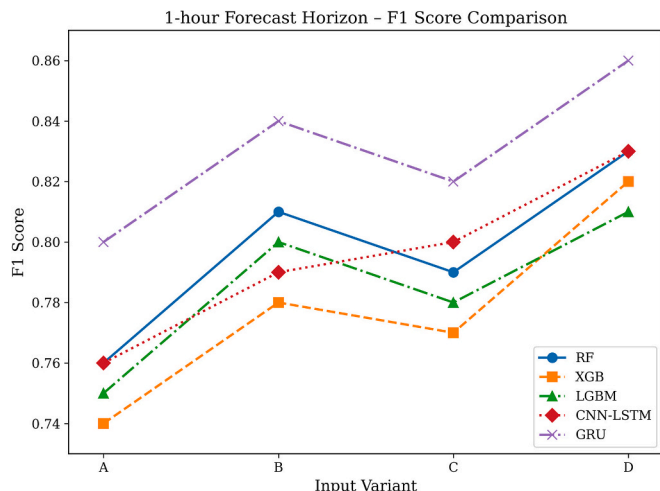


Fig. 3. Comparison of F1 score for all models at 1 h across all input variants.

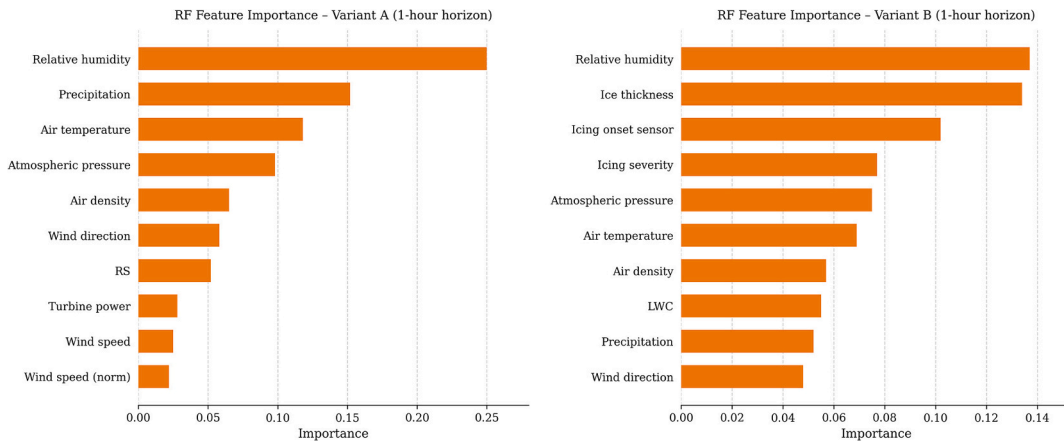


Fig. 4. a-b: Feature importance for RF at 1 h for variant A and Feature importance for RF at 1 h for variant B.

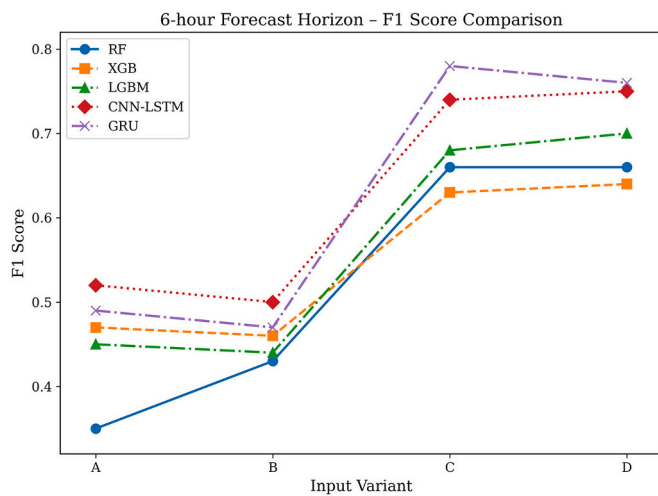


Fig. 5. Comparison of F1 score for all models at 6 h across all input variants.

contribution highlights the operational value of sensor integration, particularly for early warning at short horizons.

At the 1-h horizon, the prominence of temperature, relative humidity, liquid water content (LWC), and icing-sensor variables is consistent with the physics of rapid ice accretion near 0 °C, where small changes in thermodynamic conditions determine whether supercooled liquid water remains available in the atmosphere. LWC directly represents the amount of liquid water that can be captured by the blade, and for a given sub-zero temperature, accretion intensity is therefore strongly controlled by LWC rather than by temperature alone. As a result, periods with similar thermal conditions can lead to very different icing outcomes depending on the available liquid water. The substantial performance increase from Variant A to Variant B reflects the models' ability to exploit this microphysical information through direct icing and LWC measurements, rather than relying only on indirect operational or meteorological proxies. For the studied turbine and winter season, short-horizon icing prediction is most strongly influenced by immediate microphysical and thermodynamic conditions. The improvement observed from Variant A to Variant B, together with the comparatively smaller contribution from forecast inputs, indicates that locally measured icing indicators and liquid water content are the most informative inputs at 1 h. Under these enriched input conditions, ensemble and deep-learning models exhibit similar performance, suggesting that input informativeness outweighs architectural complexity for near-term operational detection.

The feature importance pattern further supports the process-based

interpretation of predictive skill at this horizon. In Variant A, the absence of any icing measurement forces the model to rely exclusively on atmospheric precursors reflecting thermodynamic conditions governing supercooled liquid water availability. This represents a lower bound on process-based skill in the data, independently of icing state continuity. The shift in importance toward microphysical variables when sensors are added in Variants B and D confirms that performance gains arise from direct physical signal rather than state-copying behaviour.

4.2. Six-hour horizon: transition from sensor-driven refinement to forecast-driven skill

Model performance decreases compared with the 1-h horizon, indicating that current turbine and atmospheric conditions do not fully capture the processes leading to icing six hours ahead. Fig. 5 shows that under Variant A, deep models maintain a slight advantage (CNN-LSTM F1 = 0.52; GRU F1 = 0.49) over tree-based methods, but all models exhibit moderate precision (PR-AUC $\leq$ 0.58) and balanced accuracy below 0.80. This suggests that without forward-looking context, models are sensitive to short-term patterns but struggle to anticipate delayed onset.

Introducing icing sensors in Variant B results in heterogeneous response. Random Forest improves from F1 = 0.35 to 0.43, reflecting its ability to exploit empirical indicators of recent accretion. In contrast, gradient-boosted models (XGB, LGBM) show minimal change, indicating limited sensitivity to isolated sensor signals. CNN-LSTM and GRU experience slight reductions in performance (e.g. CNN-LSTM 0.52 to 0.50), consistent with the fact that icing sensors inform present-state detection but provide weak guidance on future event development. Improvements in precision-recall observed in online Annex D confirm this variability stems from changes in detection confidence rather than threshold instability.

A clear shift occurs once forecast variables are introduced (Variant C). All models exhibit marked improvement, with LGBM, XGB and RF reaching F1 $\approx$ 0.63–0.68 and deep models rising to 0.74 (CNN-LSTM) and 0.78 (GRU). Balanced accuracy increases to $\geq$ 0.82 for both model families. These results suggest that six-hour icing risk is increasingly influenced by evolving thermodynamic conditions rather than turbine behaviour alone. Deep architectures benefit most due to their ability to integrate forecast-driven temporal patterns, whereas tree-based models exploit the structured meteorological signals at face value. Adding icing sensors to forecasts (Variant D) yields only incremental gains ($\leq$ +0.02 F1), suggesting that once future weather trajectory is known, sensor inputs mainly help refine borderline decisions rather than drive major improvement.

F1 is retained as the primary metric given its relevance for early-

warning scenarios. PR-AUC and balanced accuracy support the same trends (Appendix A), confirming that performance gains arise from improved identification of positive cases rather than threshold tuning effects.

4.2.1. Feature importance analysis (6-h horizon)

Feature importance patterns were derived using SHAP values from the GRU model, which achieved the highest performance among deep architectures at this horizon and best captures temporal learning behaviour (Fig. 6a). Under Variant A (SCADA and on-site meteorology only), the model assigns greatest importance to air temperature and barometric pressure, indicating reliance on present-state atmospheric conditions. When forecast variables are introduced (Variant C), importance shifts toward horizon-matched predictors such as forecasted temperature, humidity and wind speed, showing that predictive skill increasingly derives from anticipated rather than immediate environmental states (Fig. 6b). This transition is consistent with the observed performance gains between Variants A and C.

At the 6-h horizon, predictive skill depends primarily on the evolution of forecasted large-scale atmospheric conditions rather than on instantaneous local icing indicators. The GRU feature-importance results for the forecast-enriched configuration (Variant C, Fig. 6a) show that air temperature, forecast cloud cover, pressure and wind speed dominate the prediction, indicating sensitivity to synoptic-scale conditions associated with icing-favourable conditions. In contrast, relative humidity appears only in the on-site configuration (Variant A, Fig. 6b), because forecast relative humidity was not used; instead, forecast moisture is represented through the specific humidity, cloud and liquid water variables provided by the ECCC numerical weather prediction system. These forecasted variables describe whether the atmospheric environment is transitioning toward conditions supportive of future icing-favourable moisture and temperature conditions, even before local accretion is detected by sensors. This explains the substantial performance gains obtained when forecast inputs are included at this horizon. Consistent trends are observed for the tree-based models (Annex D), confirming that the improvement is driven by access to forecasted atmospheric structure rather than by model architecture.

At the 6-h horizon, the relative importance of forward-looking atmospheric information increases. The consistent performance gains from Variant C compared with Variants A and B indicate that horizon-matched forecast variables provide the dominant contribution to medium-range predictability for this dataset, while icing sensors mainly refine present-state detection. Deep learning models show slightly higher sensitivity to these evolving forecast patterns, although ensemble methods remain competitive once forecast inputs are included.

4.3. 24-h horizon: long-range prediction governed by forecasted atmospheric evolution

Model accuracy declines at the 24-h horizon, reflecting the limited ability of current turbine or meteorological conditions to anticipate icing risk one day ahead. As shown in Fig. 7, performance under Variant A remains low across all architectures, with CNN-LSTM achieving the

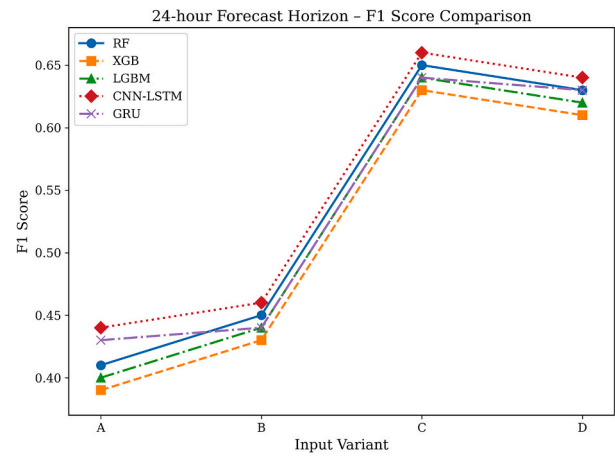


Fig. 7. Comparison of F1 score for all models at 24 h across all input variants.

highest discrimination ($F1 = 0.44$) and GRU only slightly lower. These scores suggest that models attempt to infer future icing indirectly from current operational response and ambient conditions, but the absence of forward-looking information restricts their practical utility. Adding icing sensors (Variant B) results in only marginal improvement (e.g. RF 0.41 to 0.45) and negligible changes in balanced accuracy, indicating that localised accumulation signals hold limited predictive value beyond immediate horizons.

The introduction of forecasted weather variables (Variant C) produces a substantial increase in performance across models, with F1 scores typically reaching 0.63–0.66 and balanced accuracy exceeding 0.72. Under this configuration, differences between deep and ensemble methods narrow significantly (e.g. CNN-LSTM 0.66 vs RF 0.65), indicating that once relevant future atmospheric information is available, input informativeness is more influential than model depth. Only marginal gains, or slight reductions, are observed under Variant D, suggesting that short-horizon icing sensor variables do not materially improve day-ahead forecasts when forecast-based drivers already dominate the prediction. Additional performance indicators reported in Appendix A confirm that these F1-based trends are reflected consistently across PR-AUC and balanced accuracy. Overall, these results show that 24-h icing prediction is strongly associated with expected atmospheric evolution, with reduced influence from current turbine behaviour or sensor-based accumulation indicators. Within the studied site and season, deep learning remains valuable when future weather information is unavailable, but when reliable forecasts are present, lighter models such as LGBM achieve comparable performance, supporting their suitability for operational deployment at this horizon.

Overall, these results indicate that long-range icing prediction is strongly associated with expected atmospheric evolution rather than current operational state or sensor-based accumulation cues. While CNN-LSTM performs best under sparse inputs, tree-based models such as LGBM provide comparable accuracy once forecast data are incorporated, supporting their suitability for day-ahead operational

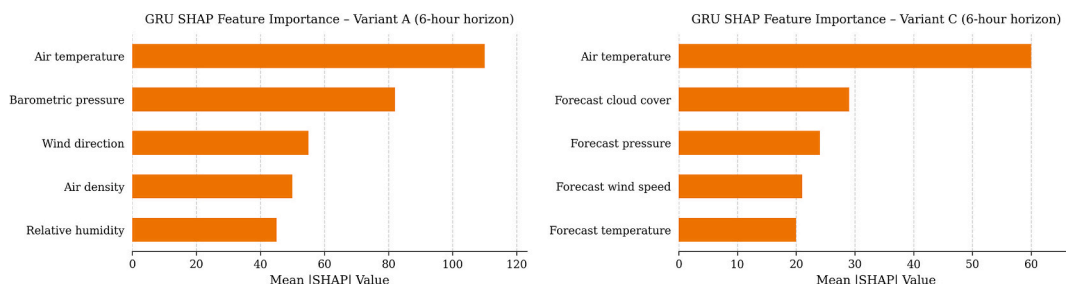


Fig. 6. a-b: Feature importance for GRU at 6 h for input A and C.

applications given their efficiency and ease of deployment.

4.3.1. Feature importance analysis (CNN-LSTM, 24-h horizon)

Feature importance results align with the performance behaviour observed at longer horizons. Under Variant A, the CNN-LSTM primarily relies on air temperature, air density and relative humidity, alongside operational variables such as rotor speed and turbine power (Fig. 8a).

These parameters reflect the turbine's immediate thermodynamic environment but act mainly as reactive indicators, limiting their relevance over a 24-h lead time. This is consistent with the relatively modest discrimination achieved ($F1 \approx 0.44$).

When forecast variables are introduced (Variant C), importance shifts toward horizon-matched predictors such as forecasted wind speed, temperature and air density (Fig. 8b). This transition indicates that the model moves from inferring risk based on current turbine response to anticipating atmospheric evolution. The subsequent increase in $F1$ (≈ 0.44 to ≈ 0.65) therefore stems from access to physically meaningful future indicators rather than enhanced temporal reasoning alone.

These findings reinforce that, at day-ahead scales, input informativeness particularly through meteorological forecasts has greater influence on prediction skill than network depth. While enriched inputs allow long-range risk anticipation, uncertainty remains inherently higher at this horizon. The next section evaluates uncertainty propagation and event-level performance to assess robustness under realistic decision-making criteria. For day-ahead forecasts in this single-site winter dataset, prediction skill is primarily governed by access to horizon-matched meteorological forecasts rather than by local turbine or icing-sensor measurements. The large improvement observed when moving from Variants A and B to Variant C indicates that usable 24 h icing prediction depends on anticipated atmospheric evolution. When such forecast information is available, both ensemble and deep-learning models achieve comparable performance, implying that long-range skill is largely input-limited rather than architecture-limited in this setting.

4.4. Horizon-dependent impact of input enrichment

To synthesise how different sources of information influence icing prediction across lead times, and to provide a coherent framework for identifying which inputs matter at which horizons, performance contrasts between input configurations were examined under controlled information conditions. Rather than relying on visual overlap of confidence intervals alone, paired comparisons were used to assess how forecast-derived predictors and icing-sensor variables contribute to predictive skill as forecast horizon increases. The analysis focuses on two conditional transitions that remain meaningful across all horizons: the transition from Variant A to Variant C, which isolates the effect of

Table 2

Median paired changes in $F1$ -score associated with forecast and icing-sensor input enrichment. Ranges reflect variability across model architectures; direction and magnitude were consistent across all models.

Horizon	Transition	Median $\Delta F1$	Model-wise range
1 h	A $\rightarrow$ C	+0.02	+0.02 to +0.03
1 h	C $\rightarrow$ D	+0.07	+0.06 to +0.07
6 h	A $\rightarrow$ C	+0.13	+0.12 to +0.14
6 h	C $\rightarrow$ D	+0.01	+0.01 to +0.02
24 h	A $\rightarrow$ C	+0.30	+0.29 to +0.31
24 h	C $\rightarrow$ D	+0.02	+0.01 to +0.03

incorporating forecast information, and the transition from Variant C to Variant D, which quantifies the additional value of icing-sensor inputs when forecasts are already available.

Table 2 summarises paired within-model changes in $F1$ -score associated with successive input enrichment (A $\rightarrow$ C and C $\rightarrow$ D) across forecast horizons, together with model-wise ranges that reflect uncertainty across architectures.

At the 1 h horizon, the transition from Variant A to Variant C yields only modest improvements across all models, indicating that short-term icing predictability is weakly influenced by forecasted atmospheric evolution. In contrast, the transition from Variant C to Variant D produces a clear and consistent performance increase, reflecting the dominant role of locally observed microphysical conditions, including liquid water content and icing sensor measurements, for near-term icing detection. This pattern holds across both ensemble and deep learning architectures, demonstrating that immediate icing risk is increasingly influenced by on-site physical sensing rather than by anticipated weather changes.

At the 6 h horizon, the relative contribution of forecast information increases markedly. Across all models, Variant C consistently outperforms Variant A by a substantial margin, indicating that medium-range icing predictability depends strongly on access to horizon-matched atmospheric forecasts. Once forecast inputs are incorporated, the additional gain from icing-sensor variables is limited, with only marginal improvements observed in the transition from Variant C to Variant D. This shift reflects a transition regime in which atmospheric evolution begins to dominate predictability, while the influence of instantaneous local measurements diminishes.

At the 24 h horizon, forecast-derived inputs become the principal driver of predictive skill. Models trained without forecast information (Variants A and B) perform poorly regardless of architecture, whereas the inclusion of forecast predictors (Variant C) leads to a step change in performance across all model classes. The subsequent addition of icing-sensor variables yields only minor incremental gains, indicating that

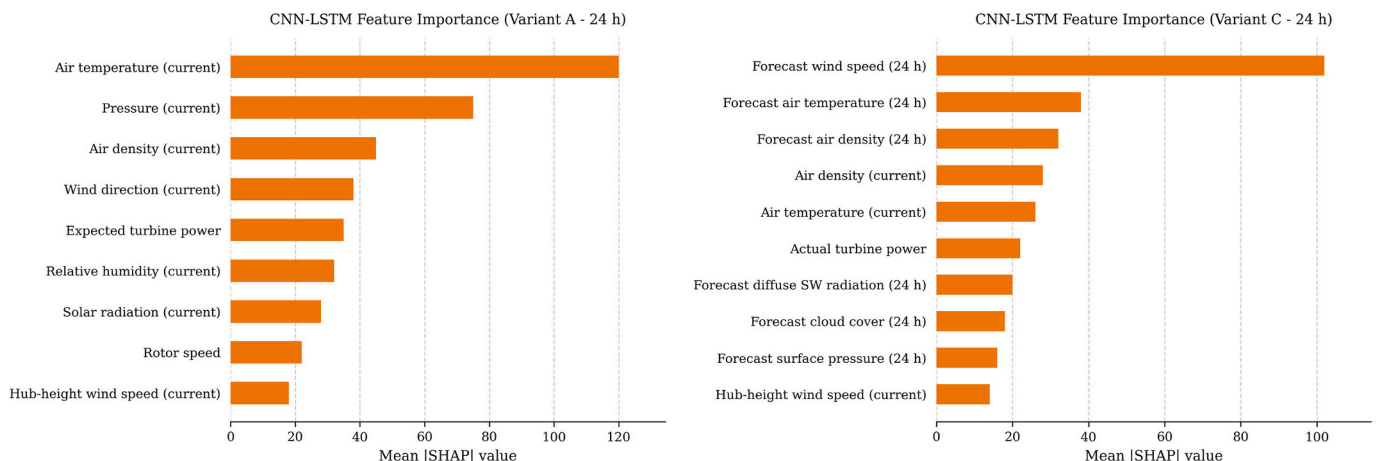


Fig. 8. a-b Feature importance for CNN-LSTM at 24 h for variant A and C.

local sensing provides limited additional information once predictions extend into the day-ahead regime. These results suggest that reliable long-horizon icing forecasts are strongly forecast-driven for the studied setting and cannot be achieved through local measurements alone.

Overall, the paired contrasts demonstrate a systematic and horizon-dependent reweighting of input importance. Icing-sensor variables dominate short-horizon predictions, forecast-derived atmospheric predictors dominate medium and day-ahead forecasts, and model architecture plays a secondary role once physically informative inputs are available. This finding is consistent with the central premise of the study: that icing forecast skill is governed primarily by the relevance and temporal alignment of the available information rather than by the complexity of the learning architecture.

The horizon-dependent value of continuous IC-1 measurements is further confirmed by two complementary analyses. First, a binary-flag ablation in which the RF model for Variant B was restricted to the binary icing flag alone (removing LWC, thickness, and severity) shows that continuous channels provide a modest but consistent improvement at the 1-h horizon ($\Delta F1 = +0.025$, full channels 0.807 vs binary-flag 0.782), consistent with their physical relevance at short lead times. At 6 and 24 h this advantage does not hold, consistent with current icing intensity carrying no direct information about icing state beyond its immediate physical timescale. Second, RF feature importance confirms the same pattern at the channel level: the four IC-1 channels collectively account for 61% of total importance in Variant B at 1 h, declining sharply to 32% at 6 h and 14% at 24 h, where NWP variables dominate entirely. Thickness shows the most persistence across horizons, consistent with ice accumulation being a slower-evolving process than instantaneous LWC or severity. Together these analyses confirm that continuous microphysical measurements provide modest short-horizon value beyond the binary icing flag, while their contribution diminishes beyond the physical timescale of the icing state they measure.

4.5. Uncertainty and event-level forecast skill

Model robustness was assessed using block bootstrapping of timestamp-level predictions (1000 resamples) to account for temporal dependence. Across configurations, 95% confidence intervals for F1-score typically ranged between ± 0.02 and ± 0.05 . Event-level performance values are shown in Fig. 9 (full summary in online Appendix D), which shows that all models benefit from input enrichment, with CNN-LSTM and GRU achieving the highest skill under forecast-supported configurations (C/D) and machine learning models closely following once future atmospheric evolution is incorporated. Model ranking remained unchanged within these intervals, suggesting that the observed performance differences reflect meaningful predictive ability rather than sampling variability. Although confidence intervals partially

overlapped in some cases, paired block-bootstrap comparison indicated that the enriched configuration achieved higher F1 in over 85% of resamples, confirming that improvement trends are systematic rather than due to random variation. While the analysis covers a single winter season, blocked time-series cross-validation combined with event-level bootstrapping ensures that these trends are evaluated across multiple independent temporal segments within that season.

Probability calibration using isotonic regression reduced over-confidence in tree-based models, particularly at short horizons, while deep architectures retained slight optimism in upper probability ranges. Sensitivity testing of the operating threshold τ^* , perturbed by ± 0.05 around the validation-optimised value, changed timestamp-level F1 by less than 0.01 on average, indicating that comparative outcomes are not overly dependent on threshold selection.

Since timestamp-level metrics do not fully reflect operational usefulness, an event-level F1-score was computed by grouping continuous detections into icing episodes. An event was considered correctly identified if the predicted interval overlapped with the observed episode for at least 50% of its duration or occurred within a ± 30 -min window. Event-level scores were consistently higher than timestamp-level values across all horizons, improving by approximately 0.03–0.11 depending on model and input configuration. The greatest gains occurred for configurations with fragmented timestamp detections, such as Random Forest and CNN-LSTM under sparse inputs, confirming that although timing precision may be limited, core icing evolution is frequently captured.

Event-level performance metrics are sensitive to the chosen temporal matching criteria, including the 50% overlap threshold and ± 30 -min window (Appendix Table D5). Sensitivity checks using alternative thresholds (e.g., stricter 70% overlap or ± 15 min, and more lenient 30% overlap or ± 60 min) showed that while absolute event F1 scores vary, the relative ranking of model configurations and the directional improvements from input enrichment remain robust. The selected criteria were chosen to reflect realistic operational tolerance for timing uncertainty in icing onset detection. The key finding—that forecast-informed configurations substantially improve event detection capability at medium and long horizons—therefore holds across reasonable threshold variations. Event-level metrics should be interpreted as indicative of operational detection capability within this evaluation framework rather than as threshold-independent measures.

Event-level analysis also reveals systematic differences in predictability across icing-event characteristics. Short-duration and moderate-to-severe icing episodes, particularly those associated with clear thermodynamic transitions near the freezing point and elevated liquid water content, tend to be detected more reliably across the evaluated horizons. In contrast, long-duration or weak icing events with gradual onset, intermittent accretion signals, or marginal atmospheric conditions are

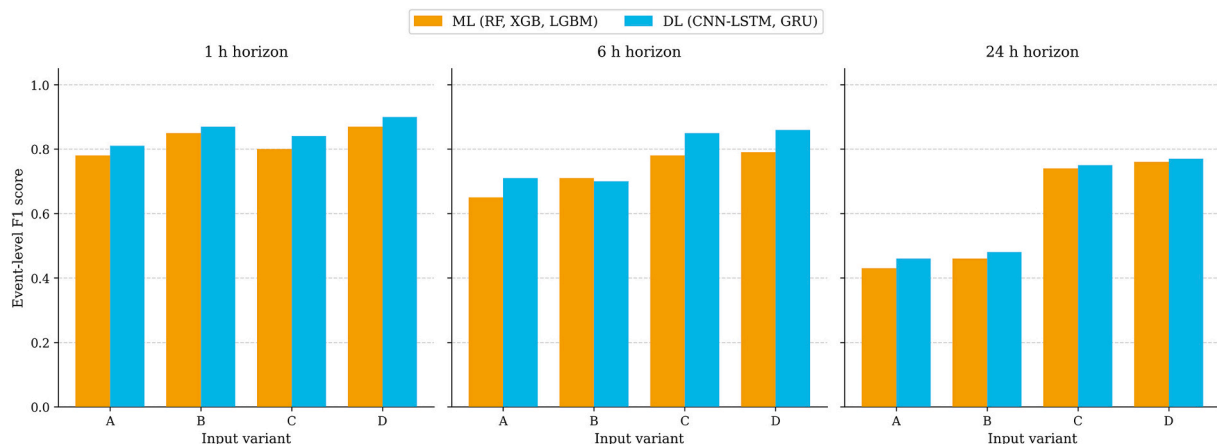


Fig. 9. Event-level F1-scores with 95% confidence intervals across all models, input configurations (A-D) and forecast horizons.

more challenging to predict consistently, especially at longer lead times. These limitations reflect both intrinsic uncertainty in atmospheric evolution and reduced signal-to-noise ratios in icing indicators under marginal conditions.

At the 1 h horizon, ML and DL models achieved event-level $F1 \geq 0.80$ for enriched input configurations, supporting their suitability for short-term operational decision-making. However, IC-1 ice sensor (Variant B) enabled similar or better accuracy to achieved than using only forecasted weather data (Variant C), which is often not available.

Performance remained moderate at 6 h, with deep architectures providing a slight advantage under forecast-driven inputs, though well-configured ensemble learners remained competitive. The use of forecasted data (C and D) became significantly more important than the IC-1 ice sensor data. At 24 h, this became even more evident, with ≈ 0.66 to ≈ 0.75 achieved for Variant C/D.

4.6. Reliability stratification across environmental regimes

Understanding operational reliability under varying meteorological conditions is critical for deployment of icing prediction systems in cold-climate wind energy operations. To address this requirement, model performance was stratified by environmental regime across forecast horizons and input configurations (Table 3). Regime-specific F1 values were derived by aggregating out-of-fold predictions from the cross-validation procedure reported in Annex D into regime-specific confusion matrices, evaluated at the same decision thresholds. This post-hoc stratification provides reliability interpretation patterns to guide operational decision-making. Regime columns represent independent (non-exclusive) stratifications rather than a partition of the feature space, and act as statistical proxies for underlying microphysical processes rather than direct classification of icing type.

Regime membership was determined exclusively from on-site observed values—not forecast fields—evaluated at the prediction timestamp, t . Specifically, the threshold criteria listed in Table 3 (e.g. $T < 0^\circ\text{C}$, $\text{LWC} \geq 0.1\text{ g/m}^3$, $\text{RH} \geq 90\%$, wind speed $\geq 8\text{ m/s}$) were applied to IC-1 sensor and on-site meteorological measurements recorded at the time the prediction was issued. This design ensures regime assignment is operationally reproducible without requiring knowledge of future conditions. It also means regimes characterise the atmospheric state from which the model is predicting, rather than the state at the target horizon—for example, a prediction labeled “near-freezing” reflects conditions at t , regardless of whether icing occurred at $t + H$.

Performance stratification reveals systematic patterns across conditions. Regimes associated with stronger detection signals—high relative humidity ($\geq 90\%$), high-signal conditions ($T < 0^\circ\text{C}$ AND $\text{LWC} \geq 0.1\text{ g/m}^3$ AND $\text{RH} \geq 90\%$), and high wind speeds ($\geq 8\text{ m/s}$)—consistently exhibited elevated performance across all forecast horizons. At the 1-h horizon under enriched inputs (Variant D), these conditions yielded F1 scores of 0.88, 0.92, and 0.86 respectively, reflecting strong predictive capability when atmospheric indicators provide unambiguous signal. These conditions are associated with more persistent and detectable accretion patterns, consistent with established relationships between

environmental state and ice accumulation rate.

Regimes with weaker detection signals—near-freezing temperatures (-2°C to $+2^\circ\text{C}$), low liquid water content ($< 0.1\text{ g/m}^3$), and medium relative humidity (70–90%)—exhibited degraded performance across all horizons. At 1 h, these marginal conditions yielded F1 scores of 0.76, 0.68, and 0.74 under enriched inputs, compared to 0.68, 0.62, and 0.64 under baseline configurations. The performance gap between high-signal and low-signal regimes (approximately 15–20 percentage points) reflects the difficulty of detecting icing onset when atmospheric indicators are weak or oscillate near threshold values. Near-freezing conditions introduce uncertainty in precipitation phase transitions, while low liquid water content provides minimal signal strength for detection algorithms. Medium-humidity regimes represent transitional atmospheric states where icing likelihood is inherently ambiguous due to competing thermodynamic processes.

Performance degradation was more pronounced at longer forecast horizons, particularly in low-signal regimes. At 24 h, low-LWC conditions remained below $F1 = 0.54$ even with full enrichment, compared to $F1 = 0.66$ – 0.68 under high-signal regimes—a persistent 12–14 point gap. The relative performance advantage of high-signal over low-signal regimes persisted across horizons, demonstrating that regime-dependent limitations are systematic across lead times. These patterns reflect the propagation of atmospheric uncertainty with lead time, which disproportionately affects weak-signal regimes where forecast skill depends critically on accurate representation of subtle microphysical processes.

Input enrichment (Variant A to D) improved performance across all regimes but did not eliminate the signal-strength gap. At 1-h, enrichment improved low-LWC performance from $F1 = 0.62$ to 0.68 (+6 points) while improving high-RH performance from 0.80 to 0.88 (+8 points), indicating that sensor and forecast inputs benefit both regime types. However, the absolute performance ceiling remains regime-dependent: even with full enrichment, low-signal regimes at 24 h remain below $F1 = 0.58$, whereas high-signal regimes maintain $F1 \geq 0.62$. The persistence of a regime gap under enrichment, and its amplification with lead time, indicates a practical predictability limit dominated by weak microphysical signal and forecast uncertainty rather than model architecture or training methodology.

These regime-specific patterns provide actionable operational guidance for turbine operators and infrastructure managers. High-signal regimes support direct operational action—such as pre-emptive anti-icing system activation or turbine shutdown protocols—with lower false-alarm propensity relative to marginal regimes, even at day-ahead horizons. Low-signal regimes warrant ensemble-based probabilistic interpretation, conservative threshold adjustment, or human oversight to account for elevated forecast uncertainty. The systematic nature of these patterns across model architectures and input configurations suggests they reflect systematic differences in signal strength and forecast uncertainty across conditions, providing potentially transferable insights beyond wind energy applications, subject to validation in other sites and seasons. Overall, regime-dependent reliability emerges as a structural property of day-ahead icing prediction, indicating that

Table 3

Reliability stratification across environmental regimes and forecast horizons ($\dagger - T < 0^\circ\text{C}$ AND $\text{LWC} \geq 0.1\text{ g/m}^3$ AND $\text{RH} \geq 90\%$).

Horizon	Input	Overall F1	High RH ($\geq 90\%$)	High-signal conditions $\dagger$	High wind ($\geq 8\text{ m/s}$)	Near-freezing (-2 to 2°C)	Low LWC ($< 0.1\text{ g/m}^3$)	Medium RH (70–90%)
1-h	Variant A	0.74	0.80	0.83	0.78	0.68	0.62	0.64
	Variant D	0.82	0.88	0.92	0.86	0.76	0.68	0.74
6-h	Variant A	0.47	0.53	0.52	0.48	0.42	0.39	0.41
	Variant D	0.64	0.71	0.72	0.67	0.57	0.54	0.55
24-h	Variant A	0.39	0.43	0.44	0.40	0.36	0.33	0.33
	Variant D	0.61	0.68	0.66	0.62	0.58	0.54	0.55

forecast skill is constrained more by environmental signal strength than by model architecture alone.

4.7. Operational implications

The horizon-dependent results define a clear operational decision framework. For short-term icing control (1 h), on-turbine icing sensors and liquid water content measurements provide the largest performance gains and are therefore most relevant for triggering blade heating and de-icing actions. For medium-range planning (6 h), forecast meteorological variables dominate predictive skill and are most valuable for scheduling maintenance, crew deployment and turbine availability. For day-ahead planning (24 h), weather forecasts are the primary driver of usable prediction skill and are particularly relevant for energy trading, production scheduling and grid commitments, while icing-sensor inputs add little additional benefit. This ranking of input families by forecast horizon indicates where investment in sensing versus forecasting infrastructure yields the greatest operational value.

Reliable ice forecasting has tangible operational value, as it enables wind farm operators to anticipate production losses and optimise the timing of anti-icing or de-icing actions. When icing can be identified earlier and with greater certainty, interventions can be targeted more efficiently, reducing unnecessary curtailment and minimising energy losses. The strongest benefits are expected at medium and long horizons, where improved recall supports proactive planning, while short-horizon accuracy allows precise triggering of blade heating when accretion is already imminent. To illustrate the potential scale of this impact, a simple energy recovery estimate is provided based on the frequency and duration of observed icing events over the study period.

Across the five-month dataset, 211 icing episodes were recorded, totalling approximately 528 h of icing conditions, with a mean duration of around 2.5 h per event. For operational estimation, a representative 2 MW cold-climate turbine was assumed, which typically delivers approximately 1.0 MW under icing-prone conditions due to reduced aerodynamic efficiency and moderate winter wind speeds. If improved forecasts prevent just one hour of unnecessary curtailment per episode, the potential seasonal energy recovery is approximately 211 MWh per turbine ($211 \text{ events} \times 1 \text{ h} \times 1.0 \text{ MW}$). To reflect typical variability during icing events, turbine output was conservatively varied between 0.8 MW and 1.2 MW. The lower value represents reduced power under limited wind conditions, whereas the upper value assumes favourable wind conditions despite icing risk. This results in an estimated recovery between approximately 170 MWh (211×0.8) and 250 MWh (211×1.2) per turbine per winter, aligning with reported winter loss estimates of 10 to 20% of annual production. These estimates are intended to be illustrative and are derived from aggregated event statistics rather than individual case studies, given data confidentiality.

Electro-thermal de-icing systems typically consume around 5 to 10% of rated turbine power while active, corresponding to approximately 100 to 200 kW for a 2 MW turbine (Stoyanov and Nixon 2021; Getz et al. 2020). Even after accounting for this auxiliary energy use, the net recovery remains positive because the avoided loss from unnecessary curtailment generally exceeds the power demand for heating. Moreover, IPS heating can be modulated to maintain anti-icing conditions, rather than operating at constant power throughout an icing event.

These estimates are intended as order-of-magnitude values. Actual recoverable energy will depend on turbine class, site-specific climatology, curtailment strategy and icing frequency. However, the results indicate that improved forecasting, particularly when using both sensor-based inputs and forecast-derived atmospheric predictors, can support significantly more targeted and proactive mitigation of icing impacts.

4.8. Limitations and future directions

The study involved a single turbine and weather season, which constrains spatial transferability and does not capture inter-turbine

variability or local environmental differences in icing conditions. Forecast inputs were sourced from a single numerical weather prediction provider and, although incorporated in a leak-safe manner, may embed systematic biases that were not fully characterised through calibration. The binary icing target is derived from a single IC-1 ice sensor, and sensor noise, drift, or installation-specific effects may propagate into the learned relationships and reported performance. While the results demonstrate clear gains when horizon-matched forecasts are used, future work should compare multiple weather forecasting systems and test whether adaptive bias-correction or ensemble forecasting improves long-horizon robustness. Deep learning models were trained using fixed architectures with tuned hyperparameters. Future work should investigate adaptive model designs, uncertainty-aware formulations and online learning strategies to improve resilience under changing atmospheric conditions. Although the conclusions are derived from a single turbine, site and winter season, the prominence of forecast information at medium and long horizons aligns with established synoptic-scale icing processes, suggesting that the qualitative hierarchy of input informativeness may generalise to other cold-climate locations. Extending the analysis to multi-site, multi-winter datasets and establishing a publicly available benchmark with standardised temporal splits and reproducible code would provide stronger empirical validation and facilitate deployment in operational forecasting systems.

5. Conclusion

This study evaluated five forecasting models across three horizons and four input configurations to determine the relative importance of input quality and model architecture. The results show that enriched inputs are the dominant driver of predictive performance. Forecasted weather increased F1 scores most substantially at 6 h and 24 h horizons, while icing sensor and LWC features primarily improved 1 h predictions. Model complexity had a secondary influence, with lightweight ensemble models achieving performance comparable to deep learning once physically informative inputs were included.

Deep architectures provided marginal benefits, particularly when future atmospheric trends were not fully reflected in the inputs. Without forecast data, day-ahead predictions remained unreliable, whereas combining horizon-matched forecasts with physical sensing yielded operationally viable performance. Minor degradation in some configurations highlights the need for careful temporal alignment of sensor inputs.

Overall, for the studied turbine, site, and winter season, high-quality, physics-informed data appear more influential than architectural depth in determining icing forecast skill. Operational strategies may therefore benefit from prioritising improved sensing and access to reliable forecasts over increasing model complexity. Future work should assess generalisability across sites, seasons and forecast providers, and explore hybrid models that integrate physical understanding with data-driven methods. These conclusions are drawn from a single-site, single-season analysis and should be interpreted within this operational context.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the author(s) used CHATGPT in order to refine the writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

CRediT authorship contribution statement

A. Kallarappayi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **J.D. Nixon:** Writing – review & editing, Writing – original draft,

Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **A. Bégin-Drolet:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **C. Godreau:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Code availability

The preprocessing and model-training code (RF/XGB/LGBM, GRU, and CNN-LSTM) is institutionally restricted due to IP agreements.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: André Bégin-Drolet has patent #System and method for determining an icing condition status of an environment, July 14, 2020. US Patent 10,712,301 (CA 3,012,392) issued to André Bégin-Drolet, Jean Ruel and Jean Lemay. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the funding this research received from the Royal Society International Exchange Programme, which supported the project Intelligent Wind Turbines for Cold Climates (IES/R2/222063).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.coldregions.2026.104940>.

Data availability

The SCADA, icing-sensor, and site datasets used in this study are proprietary and not publicly available due to commercial confidentiality agreements.

References

- Afzal, F., Virk, M.S., 2018. Review of icing effects on wind turbine in cold regions. *E3S Web Conf.* 72, 01007. <https://doi.org/10.1051/e3sconf/20187201007>.
- Battisti, L., 2015. Wind turbine icing and de-icing: a review of ice-prevention systems and related technologies. *Renew. Sust. Energ. Rev.* 49, 1126–1137. <https://doi.org/10.1016/j.rser.2015.04.137>.
- Benjamin, S.G., Weygandt, S.S., Brown, J.M., Hu, M., Smirnova, T.G., Olson, J.B., Alexander, C.R., et al., 2016. A north american hourly assimilation and model forecast cycle: the rapid refresh and High-Resolution rapid refresh. *Mon. Weather Rev.* 144 (4), 1669–1694. <https://doi.org/10.1175/MWR-D-15-0242.1>.
- Bernier, N.B., Milbrandt, J.A., Tong, H., Mailhot, J., Abrahamowicz, M., 2020. Towards improved icing forecasts with environment and climate change Canada's high-resolution deterministic prediction system (HRDPS). *Cold Reg. Sci. Technol.* 175, 103063. <https://doi.org/10.1016/j.coldregions.2020.103063>.
- Cai, W., Sun, Z., Liu, S., Wang, L., Li, J., 2023. A review of data-driven approaches for wind turbine blade icing detection and prediction. *Sustainability* 15 (2), 1219. <https://doi.org/10.3390/su15021219>.
- Cao, H., Hou, Y., 2016. Sensitivity of wind turbine icing to liquid water content and droplet size in cold climates. *Wind Energy* 19 (9), 1549–1564. <https://doi.org/10.1002/we.1934>.
- Chang, X., Zhang, H., Liu, Y., 2023. Deep-learning forecasting of icing severity and duration on wind turbines. *Energy* 264, 126217. <https://doi.org/10.1016/j.energy.2023.126217>.
- Dai, H., Li, Y., Zhang, H., Zhao, X., Yang, Y., 2024a. A review of machine learning approaches for wind turbine icing detection and prediction. *IEEE Access* 12, 14726–14747. <https://doi.org/10.1109/ACCESS.2024.3356627>.
- Dai, H., Li, Y., Zhang, H., Zhao, X., Yang, Y., 2024b. Quantifying power loss and icing severity on wind turbines using multi-sensor data and deep learning. *Renew. Energy* 222, 1205–1217. <https://doi.org/10.1016/j.renene.2023.118765>.
- Dierer, S., Oechslin, R., Cattin, R., 2011. Wind turbines in icing conditions: Performance and prediction. *Adv. Sci. Res.* 6, 245–250. <https://doi.org/10.5194/asr-6-245-2011>.
- Ding, S., Liu, Y., Zhang, X., Wang, J., 2021. A PCC-ensemble-TCN model for wind turbine icing detection using class-imbalanced and label-missing SCADA data. *Int. J. Distrib. Sens. Netw.* 17 (12), 1–13. <https://doi.org/10.1177/15501477211057737>.
- Finstad, K.J., Lozowski, E.P., Makkonen, L., 1988. On the median volume diameter and its significance in icing. *Atmos.-Ocean* 26 (1), 115–119. <https://doi.org/10.1080/07055900.1988.9649284>.
- Ge, Y., Li, X., Zhou, Y., 2017. SCADA-based icing detection for wind turbines using supervised machine learning. *Renew. Energy* 101, 269–278. <https://doi.org/10.1016/j.renene.2016.08.047>.
- Godreau, C., 2020. Ice protection systems and retrofits: Performance and experiences. https://windren.se/WW2020/13_4_39_Godreau_Ice_protection_systems_and_retrof_its_Performance_and_experiences_Pub.pdf.
- Guo, Y., Infield, D., 2021. Wind turbine icing detection and prediction using machine learning. *IET Renew. Power Gener.* 15 (16), 3732–3744. <https://doi.org/10.1049/rpg2.12232>.
- Guo, J., Song, X., Tang, S., 2024. Fault diagnosis of wind turbine blade icing based on feature engineering and the PSO-ConvLSTM-Transformer. *Ocean Eng.* 302, 117726. <https://doi.org/10.1016/j.oceaneng.2024.117726>.
- IEA Wind TCP Task 19, 2017. Wind Energy Projects in Cold Climates. <https://iea-wind.org/wp-content/uploads/2021/09/2017-IEA-Wind-TCP-Recommended-Practice-13-2nd-Edition-Wind-Energy-in-Cold-Climates.pdf>.
- Ischtschuk, J., Thúrarinsdóttir, T.L., Stefánsson, G., Valdimarsson, H.P., 2019. Evaluation of surface coatings for anti-icing performance on wind turbine blades. *Cold Reg. Sci. Technol.* 158, 62–72. <https://doi.org/10.1016/j.coldregions.2018.11.010>.
- Jiang, G., Zhang, X., Wei, S., Zhou, Z., 2023. Imbalanced learning for wind turbine blade icing detection using spatio-temporal attention. *Expert Syst. Appl.* 224, 120028. <https://doi.org/10.1016/j.eswa.2023.120028>.
- Jones, K.F., 1990. A simple model for freezing rain ice loads. *Cold Reg. Sci. Technol.* 17 (3), 207–220. [https://doi.org/10.1016/0165-232X\(90\)90048-N](https://doi.org/10.1016/0165-232X(90)90048-N).
- Karlsson, T., 2026. IEA wind TCP task 54 global cold climate wind market size 2025. Winterwind 2026. https://windren.se/WW2026/10_3_24_Karlsson_IEA_Wind_Task_54_Global_cold_climate_wind_market_size_Pub_v2.pdf.
- Kreutz, M., Ait Alla, A., Varasteh, K., Ohlendorf, J.-H., Lütjen, M., Freitag, M., Thoben, K.-D., 2019. Machine learning-based icing prediction on wind turbines. *Procedia CIRP* 81, 651–656. <https://doi.org/10.1016/j.procir.2019.03.165>.
- Kreutz, M., Ait Alla, A., Varasteh, K., Ohlendorf, J.-H., Lütjen, M., Freitag, M., Thoben, K.-D., 2021. Convolutional neural network with dual inputs for time series ice prediction on rotor blades of wind turbines. *Procedia CIRP* 104, 446–451. <https://doi.org/10.1016/j.procir.2021.11.075>.
- Kreutz, M., Ait Alla, A., Lütjen, M., Ohlendorf, J.-H., Freitag, M., Thoben, K.-D., Zimmol, F., Greulich, A., 2023. Ice prediction for wind turbine rotor blades with time series data and a deep learning approach. *Cold Reg. Sci. Technol.* 206, 103741. <https://doi.org/10.1016/j.coldregions.2022.103741>.
- Lee, J.C.Y., Pao, L.Y., Clifton, A., Simley, E., King, R., et al., 2020. The power curve working groups assessment of wind turbine power performance modelling. *Wind Energy Sci.* 5, 199–227. <https://doi.org/10.5194/wes-5-199-2020>.
- Makkonen, L., 2000. Models for the growth of rime, glaze, icicles and wet snow on structures. *Philos. Trans. R. Soc. A* 358 (1776), 2913–2939. <https://doi.org/10.1098/rsta.2000.0683>.
- Messinger, B.L., 1953. Equilibrium temperature of an unheated icing surface as a function of air speed. *J. Aeronaut. Sci.* 20 (1), 29–42. <https://doi.org/10.2514/8.2495>.
- Molinder, J., Scher, S., Huang, Y., Olsson, P.-O., Widén, J., 2018. Probabilistic forecasting of wind power production losses in cold climates due to icing. *Wind Energy Sci.* 3, 667–682. <https://doi.org/10.5194/wes-3-667-2018>.
- Norconsult, 2024. The economic costs of icing and the potential of icing forecasts: A case study in Finland. https://windren.se/WW2024/10_1_03_Kurppa_The_economic_costs_of_icing_and_the_potential_of_icing_forecasts_Pub_v2.pdf.
- Parent, O., Ilincă, A., 2011. Anti-icing and de-icing techniques for wind turbines: critical review. *Cold Reg. Sci. Technol.* 65 (1), 88–96. <https://doi.org/10.1016/j.coldregions.2010.01.005>.
- Qu, M., Lu, X., Zhang, J., Wang, J., Zhang, X., 2020. Development and evaluation of a superhydrophobic anti-icing coating for wind turbine blades. *Surf. Coat. Technol.* 389, 125664. <https://doi.org/10.1016/j.surfcoat.2020.125664>.
- Roberge, P., Lemay, J., Ruel, J., Bégin-Drolet, A., 2021. In situ estimation of effective liquid water content on a wind turbine using a thermal-based sensor. *Cold Reg. Sci. Technol.* 184, 103235. <https://doi.org/10.1016/j.coldregions.2021.103235>.
- Roberge, P., Lemay, J., Ruel, J., Bégin-Drolet, A., 2022. Unsettled topics in wind energy production in cold climate. *SSRN Electron. J.* <https://doi.org/10.2139/ssrn.4110453>.
- Roberge, P., Lemay, J., Ruel, J., Bégin-Drolet, A., 2023. Understanding ice accretion on wind turbines with field data. *Cold Reg. Sci. Technol.* 210, 103853. <https://doi.org/10.1016/j.coldregions.2023.103853>.
- Scher, S., Molinder, J., 2019. Machine learning-based prediction of icing-related wind power production losses. *J. Appl. Meteorol. Climatol.* 58 (9), 1971–1981. <https://doi.org/10.1175/JAMC-D-19-0030.1>.
- Skrimpas, G.A., Mijatovic, N., Holboell, J., 2019. Wind turbine icing detection using SCADA data and machine learning. *Renew. Energy* 132, 1037–1048. <https://doi.org/10.1016/j.renene.2018.08.101>.
- Stoyanov, D., Nixon, J.D., 2019. Alternative operational strategies for wind turbines in cold climates. *Renew. Energy* 145, 2694–2706. <https://doi.org/10.1016/j.renene.2019.08.023>. IF 8/Q1.
- Stoyanov, D., Nixon, J.D., Sarlak, H., 2021. Analysis of derating and anti-icing strategies for wind turbines in cold climates. *Appl. Energy* 288, 116610. <https://doi.org/10.1016/j.apenergy.2021.116610>. IF 9.7/Q1.

- Strauss, L., Thúrarinsdóttir, T.L., Sigurðsson, S., Ólafsdóttir, T., Magnússon, K., 2020. Skill and potential economic value of forecasts of ice accretion on wind turbines. *J. Appl. Meteorol. Climatol.* 59 (11), 1823–1839. <https://doi.org/10.1175/JAMC-D-20-0025.1>.
- Tao, T., Zhang, M., Li, Y., Liu, J., Xu, Y., 2021. Wind turbine blade icing diagnosis using hybrid features and stacking ensembles. *Renew. Energy* 170, 1260–1273. <https://doi.org/10.1016/j.renene.2021.02.112>.
- Tao, C., Zhou, C., Wang, C., Li, X., Sun, J., 2023. Wind turbine blade icing prediction using focal loss and a CNN-attention-GRU algorithm. *Energies* 16 (15), 5621. <https://doi.org/10.3390/en16155621>.
- Thompson, G., Eidhammer, T., 2014. A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. *J. Atmos. Sci.* 71 (10), 3636–3658. <https://doi.org/10.1175/JAS-D-13-0305.1>.
- Wang, Y., Zhao, H., 2022. A wavelet-enhanced transformer network for wind turbine icing forecasting. *Energies* 15 (21), 7865. <https://doi.org/10.3390/en15217865>.
- Wang, Z., Yu, Y., Chen, Q., 2021. Field measurement of liquid water content and icing losses on a utility-scale wind turbine. *Appl. Energy* 288, 116638. <https://doi.org/10.1016/j.apenergy.2021.116638>.
- Xie, N., Cao, Q., Zeng, Z., 2021. Real-time prediction of blade icing on wind turbines using MLP neural networks. *Processes* 9 (6), 1024. <https://doi.org/10.3390/pr9061024>.
- Ye, F., Ezzat, A., 2024. Icing detection and prediction for wind turbines using physics-informed deep learning. *Energies* 17 (1), 158. <https://doi.org/10.3390/en17010158>.
- Zhang, X., Li, J., Liu, Y., Zhao, Z., Wang, C., 2024. Prediction of icing on wind turbines based on SCADA data via temporal convolutional network. *Energies* 17 (3), 765. <https://doi.org/10.3390/en17030765>.
- Zhou, J., Li, K., Chen, X., Liu, Y., 2018. Detection of blade icing on wind turbines using SCADA data and SVM. *Energies* 11 (10), 2603. <https://doi.org/10.3390/en11102603>.