Towards Trustworthy Surrogate Models for Augmenting Certification: Fuel Tank Flammability Reduction System

Arash Heidari*, Lorin Werthen-Brabants [†], Sebastian Rojas Gonzalez [‡], and Ivo Couckuyt [§]
Ghent University - imec, Ghent 9000, Belgium

Can Onur[¶], Pieter van Gils [‖], and Ivana Jojic [**]
The Boeing Company, Seattle, WA, 98124, USA

I. Introduction

The Flammability Reduction System (FRS), or fuel tank inerting system, mitigates explosion risk by reducing oxygen concentration within aircraft fuel tanks using Nitrogen-Enriched Air (NEA) [1]. The system compensates for oxygen release during ascent and ambient air ingress during descent or fuel consumption through controlled NEA delivery. The FRS certification process, governed by the Code of Federal Regulations (CFR) 25.981(b), involves Monte Carlo (MC) analysis of over 10,000 flight missions to assess fleetwide flammability exposure [2, 3]. However, high-fidelity physics-based MC simulations are computationally expensive. Surrogate models are employed to alleviate such burden. These models must exhibit reliability, robustness, and trustworthiness, especially in safety-critical applications [4]. Advanced surrogate modeling methods, such as deep neural networks [5–15], enhance computational efficiency but often trade off interpretability and accuracy. Thus, rigorous uncertainty quantification and conservative predictions are essential for regulatory approval.

The FRS, with its non-stationary, transient time-series data, serves as an exemplary case for leveraging surrogate models to enhance design coverage and reduce simulation resource demands. This study employs deep learning to develop reliable surrogate models of the FRS MC database, comprising 12,000 high-fidelity simulated flight missions. The work supports emerging regulatory frameworks, including EASA's AI roadmap [16] and Concept Paper [17], which outline design assurance for safety-critical AI/ML systems. Effective integration into certification processes requires collaboration across industry, academia, regulators (e.g., EASA, FAA), and standardization bodies (e.g., SAE-G35 [18], SAE-G34/EUROCAE WG-114 [18]).

*PhD Researcher, Faculty of Engineering and Architecture, Ghent 9000 (Corresponding Author, Arash.Heidari@UGent.be)
[†]Postdoctoral Researcher, Faculty of Engineering and Architecture, Ghent 9000
[‡]Postdoctoral Researcher, Faculty of Engineering and Architecture, Ghent 9000
[§]Professor, Faculty of Engineering and Architecture, Ghent 9000
[¶]Technical Manager, Boeing Research & Technology, Dubai
[‖]Technical Lead, Boeing Research & Technology, Madrid 28042
[**]Associate Technical Fellow, BCA Propulsion, Seattle, WA 98124

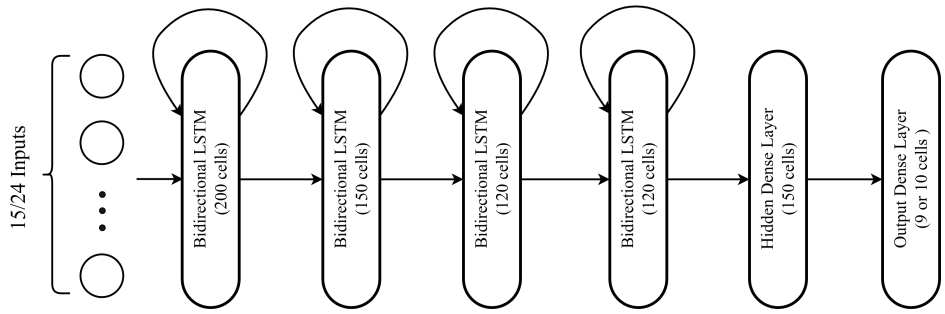

Fig. 1 Neural network architecture with layer sizes (in parentheses); 15 inputs for temperature models and 24 for oxygen models; output layer has 10 or 9 nodes for temperature or oxygen, respectively.

II. Methodology

While deep learning models have demonstrated strong performance across domains, their adoption in safety-critical applications hinges on trustworthiness. As outlined in [4], key pillars, uncertainty quantification, out-of-distribution (OOD) generalization, and explainability, are essential to ensure robustness, reliability, and interpretability.

In [11], a deep neural network modeled 109 outputs of the FRS, achieving over 10,000× acceleration with minimal accuracy loss. However, not all outputs are needed for flammability exposure assessment. This study focuses on 19 key outputs—tank temperatures and oxygen levels—better reflecting the underlying physics. This subset-based model is termed the *focused model*, in contrast to the *general model* from [11].

Dataset The dataset comprises 12,000 high-fidelity simulated flights of varying durations, each with 15 inputs and 19 time-series outputs. Outputs include 10 temperature and 9 oxygen percentage measurements across fuel tanks. Initial oxygen levels are known and can be used as additional inputs. The data is partitioned into 75% for training and 25% for testing.

Preprocessing Preprocessing involves data normalization, imputation of missing values using the last known value, and one-hot encoding of flight phases. Previously, flights were padded to match the longest sequence by repeating the last valid input, causing models to partially learn from replicated values. Since flight profiles and phases are predefined, this study instead applies masking to ignore padded timesteps during training and inference.

Model Architecture The models used in this work (Fig. 1) are based on the architecture from [11]. The temperature model uses the original 15 inputs, while the oxygen model extends the input to 24 features by including the initial oxygen levels, replicated across the flight duration.

Uncertainty Quantification Three uncertainty quantification methods are employed: Deep Ensembles (DE) [19], Monte Carlo Dropout (MCD)[20], and Quantile Regression [21]. DE involves training multiple models with different

weight initializations and averaging their predictions. MCD similarly samples stochastic predictions, with 100 forward passes used in this study. Quantile Regression directly estimates the 2.5^{th} and 97.5^{th} percentiles of each output using the pinball loss [22], defined as follows, with τ representing the quantile level being estimated, $\boldsymbol{y}$ denoting the true observed values and $\hat{\boldsymbol{y}}$ the predicted values obtained from the model:

$$\mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}}; \tau) = \frac{1}{N} \sum_{i=1}^{N} \max \left[\tau(y_i - \hat{y}_i), (\tau - 1)(y_i - \hat{y}_i) \right] . \tag{1}$$

Prediction Intervals For quantile regression, the upper and lower prediction interval bounds are derived from distinct models estimating the 97.5^{th} and 2.5^{th} percentiles, or $L(x) = f(x|\tau = 0.025)$ and $U(x) = f(x|\tau = 0.975)$. Conversely, for DE and MCD, uncertainty is computed using the sample mean and standard deviation: $U(x) = \overline{f(x)} + 1.96\sigma(f(x))$, and $L(x) = \overline{f(x)} - 1.96\sigma(f(x))$. Assuming a normal distribution on the outputs of $f(x)$, this corresponds to the 95% prediction interval.

Flammability Calculation Flammability Calculation is done as a dedicated post-processing step. To classify a minute as flammable, two conditions must simultaneously be true: at least one temperature-related output must fall within the flammability envelope ($T_{\text{min}} < \hat{y}_{t_i} < T_{\text{max}}$), and simultaneously, one of the oxygen-related outputs must exceed the inertia limit ($\hat{y}_{o_i} > O_{\text{lim}}$). These quantities are defined for each flight minute for each flight. A simple scheme for tuning the amount of false positives and negatives is by widening and lowering the flammability envelope and inertia limit respectively.

III. Results

To ensure the confidentiality of the data while preserving its underlying structure and distribution, all results presented are scaled between 0 and 1.

RMSE compared to previous work The initial analysis compares the accuracy of the new *focused* models and the revised padding technique with the *general* model from [11]. Using normalized Root Mean Squared Error (RMSE), the results for the 19 outputs in two categories are shown in Fig. 2, demonstrating that the new methodology outperforms the general model in accuracy across all output variables.

Optimal number of members of Deep Ensemble A detailed analysis varies the number of models from 3 to 51 to identify the optimal number of models in the ensemble. Figure 3 shows that deep ensembles reduce RMSE, with 51 models providing the lowest error, though fewer models offer better computational efficiency. This effect is less pronounced for oxygen-related output.

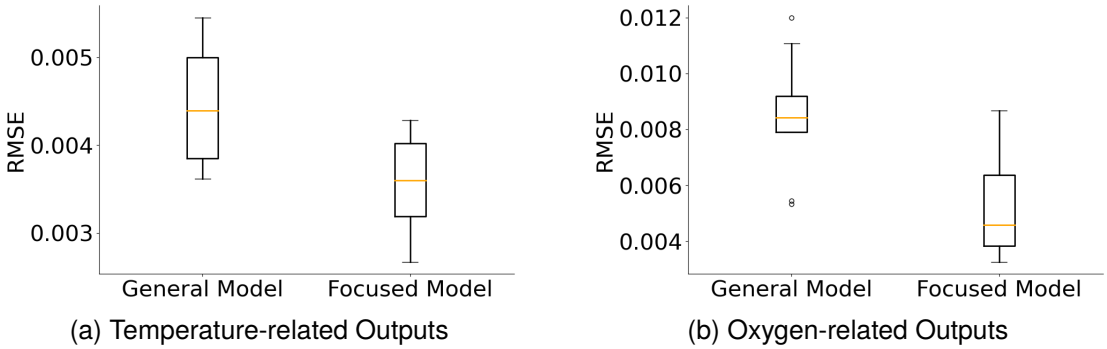

Fig. 2 Comparison of the accuracy of the different models for (a) the temperature-related outputs, and (b) the oxygen-related outputs.

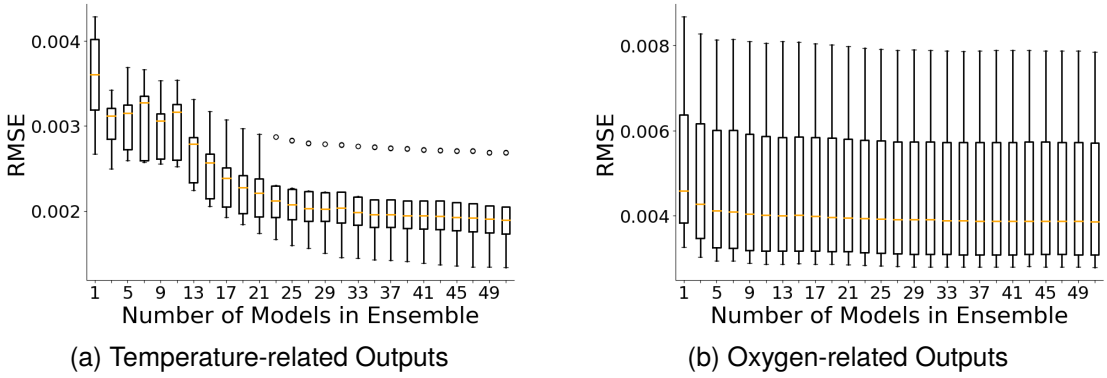

Fig. 3 Comparison of the accuracy of the different number of models for the deep ensemble for (a) the temperature-related outputs, and (b) the oxygen-related outputs.

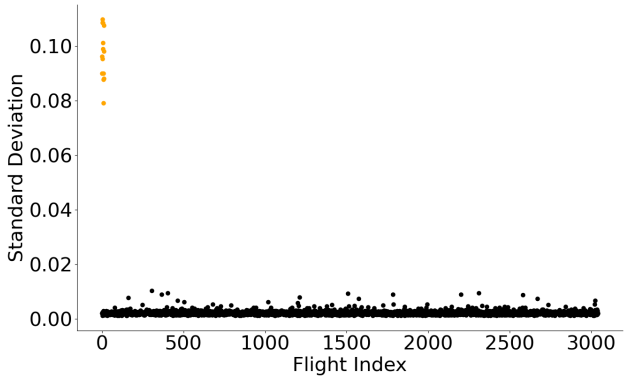

Fig. 4 The median of standard deviation for each flight. Orange points represent the maintenance flights.

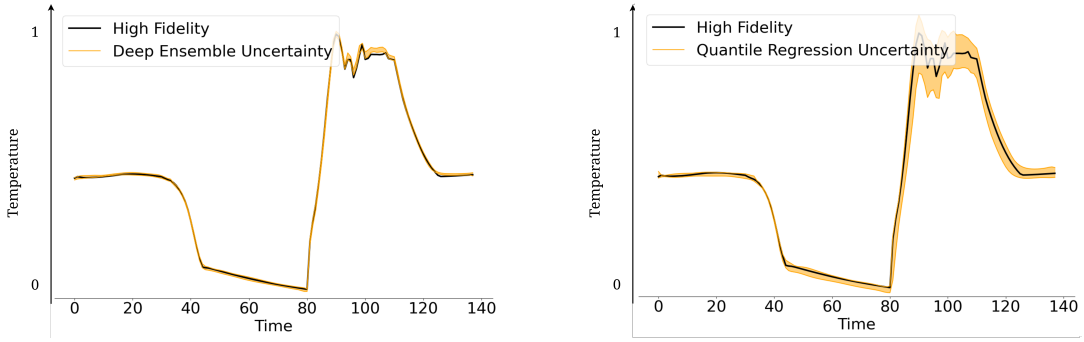

Fig. 5 Comparison of the provided uncertainty band of quantile regression and deep ensemble.

Out-of-Distribution detection of flights Some flight profiles, such as maintenance flights with unique initial oxygen levels, are considered out-of-distribution as they differ from training data. To assess OOD detection, the model is trained without maintenance flights, and the uncertainty quantification boundaries are examined to identify outliers. Figure 4 shows that the median standard deviation per flight minute predicted by the Deep Ensemble model strongly correlates with these anomalous flight profiles.

Quality of Uncertainty Quantification Uncertainty estimates from DE, QR, and MCD are compared, with an example shown in Fig. 5. The focus is on the coverage metric, which measures the proportion of true outputs within the predicted lower and upper bounds. For a dataset with N samples, and $L(x)$ and $U(x)$ representing the lower and upper bounds for input x_i, coverage is defined as:

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{L(x_i) \leq y_i \leq U(x_i)\} \tag{2}$$

Figure 6 shows the coverage of two output categories using Quantile Regression (Q), Monte Carlo Dropout (MC), and different numbers of models in the deep ensemble, presented as boxplots. While RMSE plateaus after a few

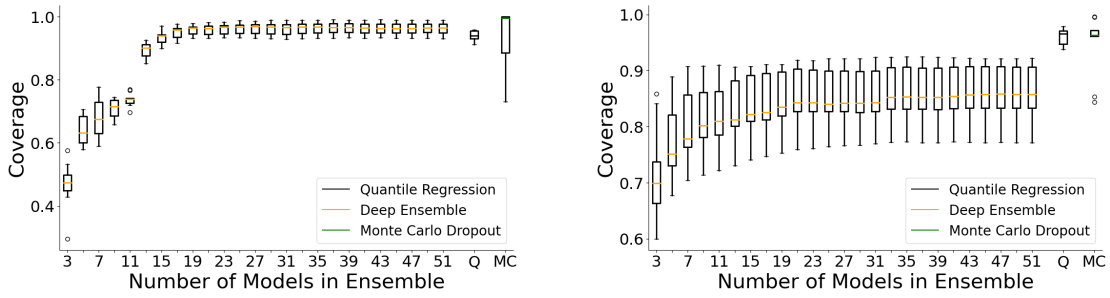

Fig. 6 Comparison of the coverage of quantile regression and deep ensemble with varying number of models for (a) the temperature-related outputs, and (b) the oxygen-related outputs.

Table 1 Comparison of performance of different models in flammability calculation.

	General	Focused	Ensemble
Precision	0.9397	0.9588	**0.9753**
Recall	0.9370	0.9844	**0.9856**
F1-score	0.9383	0.9714	**0.9804**
Accuracy	99.918%	99.961%	**99.973**%

ensemble members (Fig. 3), adequate coverage requires around 23 models to plateau.

Flammability Calculation Although the proposed method outperforms prior work in accuracy and provides uncertainty estimates, the primary goal is to determine whether a minute is flammable. Given the dataset's imbalance, where flammable minutes are fewer, accuracy is insufficient. Thus, the F1-score [23], which balances precision and recall, is used for model evaluation (Table 1). The Focused model achieves an F1-score of 0.97, outperforming the General model (0.94). The ensembled version further improves to 0.98, with precision increasing from 0.96 (single Focused model) to 0.98 (deep ensemble).

IV. Conclusion

This work marks an initial step toward developing a trustworthy surrogate model for flammability exposure estimation, capable of accurately predicting key outputs using deep learning ensembles while providing uncertainty estimates and handling edge cases like maintenance flights through augmented inputs. Among the three implemented uncertainty quantification methods, deep ensembles offered the most reliable estimates but were computationally intensive and occasionally overconfident; quantile regression was more efficient and statistically grounded but produced broader intervals, while Monte Carlo Dropout showed inconsistent coverage and limited suitability. Future work should prioritize enhancing model explainability and applying uncertainty calibration techniques to improve trustworthiness and support eventual certification by analysis.

References

[1] Federal Aviation Administration, "Reduction of Fuel Tank Flammability in Transport Category Airplanes," , 07 2008. URL https://www.govinfo.gov/content/pkg/FR-2008-07-21/pdf/E8-16084.pdf.

[2] Federal Aviation Administration, "Fuel Tank Flammability Assessment Method," Tech. rep., FAA, 2008. Available at: https://www.faa.gov/regulations_policies/rulemaking/committees/documents/media/FTFAM.pdf.

[3] European Union Aviation Safety Agency, "CS-25, Appendix M, Fuel Tank Flammability Reduction Means (FRM)," , 2023. Available at: https://www.easa.europa.eu/en/document-library/certification-specifications/cs-25-amendment-26.

[4] Mucsányi, B., Kirchhof, M., Nguyen, E., Rubinstein, A., and Oh, S. J., "Trustworthy Machine Learning," *arXiv preprint arXiv:2310.08215*, 2023.

[5] Shen, Y., Patel, H. C., Xu, Z., and Alonso, J. J., "Application of multi-fidelity transfer learning with autoencoders for efficient construction of surrogate models," *AIAA SCITECH 2024 Forum*, 2024, p. 0013.

[6] Yang, G., Allen, C. B., Markesteijn, A. P., Abid, H. A., Karabasov, S. A., and Toropov, V. V., "Surrogate Model-Based Acoustic Optimisation of Jet Nozzle Exit Geometry," *AIAA SCITECH 2022 Forum*, 2022, p. 0683.

[7] Odisho, E. V., Truong, D., and Joslin, R. E., "Applying Machine Learning to Enhance Runway Safety Through Runway Excursion Risk Mitigation," *Journal of Aerospace Information Systems*, Vol. 19, No. 2, 2022, pp. 98–112. https://doi.org/10.2514/1.I010972, URL https://doi.org/10.2514/1.I010972.

[8] Kim, J., Justin, C., Mavris, D., and Briceno, S., "Data-Driven Approach Using Machine Learning for Real-Time Flight Path Optimization," *Journal of Aerospace Information Systems*, Vol. 19, No. 1, 2022, pp. 3–21. https://doi.org/10.2514/1.I010940, URL https://doi.org/10.2514/1.I010940.

[9] Fala, N., Georgalis, G., and Arzamani, N., "Study on Machine Learning Methods for General Aviation Flight Phase Identification," *Journal of Aerospace Information Systems*, Vol. 20, No. 10, 2023, pp. 636–647. https://doi.org/10.2514/1.I011246, URL https://doi.org/10.2514/1.I011246.

[10] Anhichem, M., Timme, S., Castagna, J., Peace, A., and Maina, M., "Bayesian Approaches for Efficient and Uncertainty-Aware Prediction of Pressure Distributions," *AIAA SCITECH 2024 Forum*, 2024, p. 0253.

[11] Heidari, A., Werthen-Brabants, L., Dhaene, T., Couckuyt, I., Onur, C., van Gils, P., and Jojic, I., "Data-Driven Surrogate Modeling for the Flammability Reduction System," *AIAA SCITECH 2024 Forum*, 2024, p. 0785.

[12] Geragersian, P., Petrunin, I., Guo, W., and Grech, R., "Uncertainty-based Sensor Fusion Architecture using Bayesian-LSTM Neural Network," *AIAA SCITECH 2023 Forum*, 2023, p. 0193.

[13] Tong, H., Hauth, J. M., Huan, X., Zhou, B. Y., Gauger, N. R., Morelli, M. C., and Guardone, A., "Bayesian Recurrent Neural Networks for Monitoring Rotorcraft Icing from Aeroacoustics Time-Series Data," *AIAA Scitech 2022 Forum*, 2022, p. 2358.

[14] Pang, Y., and Liu, Y., "Probabilistic aircraft trajectory prediction considering weather uncertainties using dropout as Bayesian approximate variational inference," *AIAA Scitech 2020 Forum*, 2020, p. 1413.

[15] Memarzadeh, M., Matthews, B., and Templin, T., "Multiclass Anomaly Detection in Flight Data Using Semi-Supervised Explainable Deep Learning Model," *Journal of Aerospace Information Systems*, Vol. 19, No. 2, 2022, pp. 83–97. https://doi.org/10.2514/1.I010959, URL https://doi.org/10.2514/1.I010959.

[16] European Union Aviation Safety Agency, "Artificial Intelligence Roadmap 2.0: Human-centric approach to AI in aviation," , May 2023. Available at: https://www.easa.europa.eu/en/document-library/general-publications/artificial-intelligence-roadmap-20.

[17] European Union Aviation Safety Agency, "Concept Paper: Guidance for Level 1&2 Machine Learning Applications. Concept paper for consult." , Feb 2023. Available at: https://www.easa.europa.eu/en/document-library/general-publications/concept-paper-guidance-level-1-2-machine-learning-applications.

[18] SAE G-34/EUROCAE WG-114, "Artificial Intelligence in Aviation," , 2022. Available at: https://www.sae.org/works/committeeHome.do?comtID=TEAG34.

[19] Lakshminarayanan, B., Pritzel, A., and Blundell, C., "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, 2017, pp. 6402–6413.

[20] Gal, Y., and Ghahramani, Z., "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *international conference on machine learning*, PMLR, 2016, pp. 1050–1059.

[21] Koenker, R., and Bassett, G., "Regression Quantiles," *Econometrica*, Vol. 46, No. 1, 1978, pp. 33–50. URL http://www.jstor.org/stable/1913643.

[22] Steinwart, I., and Christmann, A., "Estimating conditional quantiles with the help of the pinball loss," *Bernoulli*, Vol. 17, No. 1, 2011. https://doi.org/10.3150/10-bej267, URL http://dx.doi.org/10.3150/10-BEJ267.

[23] Flach, P., and Kull, M., "Precision-recall-gain curves: PR analysis done right," *Advances in neural information processing systems*, Vol. 28, 2015.