The 2nd IAA Conference on AI in and for Space
SPAICE2025

Novermber 1–3, 2025, Suzhou, China

Acceptance Letter

Dear *Théo Gachet*,

It's our great pleasure to inform you that your paper, mentioned below, has been accepted for presentation at the 2nd IAA Conference on AI in and for Space (SPAICE2025), Novermber 1–3, 2025, Suzhou, China.

Paper ID: 5
Title: *Enhancing AI Trustworthiness in Rover Navigation: Risk-Aware Path Planning through Uncertainty Quantification*
Author(s): Théo Gachet, Luis Mansilla and Diane Magnin

Thank you for your great support to the SPAICE2025, and we are looking forward to seeing you in Suzhou, China.

Sincerely yours,

Dr. Olivier Contant
General Chair
The 2nd IAA Conference on AI in and for Space (SPAICE2025)

Accepted at the 2nd IAA SPAICE Conference 2025, Zhuhai, China.

Enhancing AI Trustworthiness in Rover Navigation: Risk-Aware Path Planning through Uncertainty Quantification

Théo Gachet
European Space Agency (ESA)
Noordwijk, The Netherlands

Luis Mansilla Garcia
European Space Agency (ESA)
Noordwijk, The Netherlands

Diane Magnin
European Space Agency (ESA)
Noordwijk, The Netherlands

Abstract

Autonomous Martian rovers must operate in harsh data-scarce environments under uncertainty. We propose a framework for uncertainty quantification in terrain segmentation combining Adaptive Prediction Sets with Kandinsky Calibration, yielding spatially coherent, statistically valid estimates. Integrated into a pixel-level risk map, they enable risk-aware navigation in unstructured environments. Trained on AI4Mars with a U-Net MobileNetV2 backbone, our model achieves reliable performance on Martian and lunar terrain. The UQ approach mitigates overconfident predictions in low-data regimes and supports decision-making through interpretable maps. A modified A* planner leverages these maps for risk-aware path planning across terrain and uncertainty levels. This study provides a principled UQ framework for segmentation in space robotics, enhancing safety and interpretability in autonomous navigation.

Keywords: Uncertainty Quantification, Conformal Prediction, Semantic Segmentation, Autonomous Navigation, Risk-Aware Planning, Planetary Rovers

1. Introduction

The integration of artificial intelligence (AI) into space exploration has significantly advanced planetary missions, particularly those relying on autonomous Martian rovers [1]. These systems operate in harsh, data-scarce environments under stringent constraints, demanding both accuracy and reliability in decision-making. Robust AI models that incorporate uncertainty quantification (UQ) and reliable prediction mechanisms are therefore essential [2, 3].

Trustworthy AI emphasizes reliability, safety, and explainability, all crucial for rovers making critical decisions in unpredictable conditions. On-board camera systems [4] and ML-based navigation algorithms [5] enhance autonomy and fault detection, reducing risks and costs. Yet these models often suffer from overfitting and overconfidence [6], highlighting the need for effective UQ. Conformal prediction provides a distribution-free, efficient way to generate calibrated uncertainty estimates [7], addressing key limitations of current AI reliability frameworks and offering promising directions for planetary exploration.

1.1. Objectives

This work enhances AI reliability for Martian rover navigation via advanced UQ. We integrate Adaptive Prediction Sets (APS) [8], which adjust prediction sets based on local model confidence, with Kandinsky Calibration [9], which leverages spatial correlations for coherent uncertainty maps. Together, they form a risk-aware framework combining uncertainty and terrain danger scores to guide a modified A* algorithm [10], balancing safety and efficiency.

1.2. Contributions

We present an integrated UQ framework for rover navigation with three innovations: (i) unifying APS and Kandinsky to capture model and spatial uncertainty, (ii) adapting to exploration constraints by incorporating terrain danger and addressing data scarcity, and (iii) synthesizing prediction sets and spatial uncertainties into a risk map that informs a modified A* planner. This boosts segmentation reliability and supports safe navigation.

A key theoretical contribution is the fused risk functional coupling *set-based coverage* from APS with *spatially coherent uncertainty* from Kandinsky. APS guarantees that the true class is included in the prediction set with prescribed confidence, while Kandinsky yields continuous, topology-aware estimates. The formulation preserves statistical coverage while modulating terrain risk with uncertainty.

1.3. Project Architecture

The framework includes ground-based and onboard processing. Earth-side tasks comprise preprocessing, labeling, and calibration. A U-Net with MobileNetV2 is trained and calibrated with APS and Kandinsky to compute cluster curves and quantile thresholds $q_{\text{APS}}(\alpha)$. Onboard, the model processes new images into probability maps, combines them with risk parameters to compute pixel uncertainties, and builds a risk map that feeds a modified A* algorithm to compute and execute safe, efficient paths.

2. Methods

2.1. Data Preparation

The AI4MARS dataset [11] from NASA contains 35,000 images from the Curiosity, Opportunity, and Spirit rovers with 326,000 semantic segmentation labels. Each image is annotated by 10 crowd-sourced annotators, and 1,500 high-quality validation labels were curated by NASA experts. To address class imbalance, 16,064 valid image-label pairs were selected and stratified into training (60%), validation (20%), test (10%), and calibration (10%) sets, preserving terrain class distribution. Images and masks were resized to 128×128; nearest-neighbor interpolation preserved mask semantics. Pixel intensities were normalized to $[0, 1]$, and masks were manually cleaned for consistency.

2.2. Model Architecture

We use a U-Net [12] with a MobileNetV2 encoder [13], combining efficient down-sampling for context with up-sampling for spatial resolution. This architecture balances accuracy and low computational cost, making it suitable for autonomous Martian rovers. Training uses the Adam optimizer, batch size 32, early stopping (max 50 epochs), and a learning rate scheduler.

2.3. Adaptive Prediction Sets

Adaptive Prediction Sets (APS) [8, 14] enable uncertainty-aware predictions with formal statistical guarantees. By adapting the size of the predicted label sets to the model's confidence, APS ensures that the true class is included with probability at least $1 - \alpha$. In our framework, APS is adapted for pixel-wise segmentation using a single threshold $\hat{q}$ learned from a stratified calibration set of 1606 image-label pairs, totaling over 1.68 billion valid pixels across four balanced terrain classes, and applied at inference to produce prediction sets of varying size.

Algorithm 1 APS Calibration and Prediction

Require: Calibration set $\mathcal{D}_{\text{calib}}$ with labels $y^{(c)}$, model $\hat{f}$, desired coverage $1 - \alpha$, test images
Ensure: Prediction sets $C_{i,j}^{(x)}$ for all pixels (i, j) in a new image x

Calibration phase: Determine the global threshold $\hat{q}$ to guarantee marginal coverage over $\mathcal{D}_{\text{calib}}$

Initialize empty score set $\mathcal{S} \leftarrow \emptyset$
for all pixels (i, j) in calibration images $c \in \mathcal{D}_{\text{calib}}$ **do**
  Compute class probabilities $\hat{f}(c)_{i,j} \in \mathbb{R}^K$
  Define the permutation $\pi(c)_{i,j}$ such that $\hat{f}(c)_{i,j}[\pi(c)_{i,j}[1]] \geq \cdots \geq \hat{f}(c)_{i,j}[\pi(c)_{i,j}[K]]$
  Let m be the smallest index such that $\pi(c)_{i,j}[m] = y_{i,j}^{(c)}$ (true class index)
  Compute score: $s_{i,j}^{(c)} = \sum_{k=1}^{m} \hat{f}(c)_{i,j}[\pi(c)_{i,j}[k]]$
  Append $s_{i,j}^{(c)}$ to $\mathcal{S}$
end for
Compute quantile treshold: $\hat{q} = \text{Quantile}(\mathcal{S}, \frac{(n+1)(1-\alpha)}{n})$

Prediction phase: Build prediction sets for each pixel by accumulating confidence until $\hat{q}$ is reached

for all pixels (i, j) in new test image x **do**
  Compute predicted probabilities: $\hat{f}(x)_{i,j} \in \mathbb{R}^K$
  Define the permutation $\pi(x)_{i,j}$ such that $\hat{f}(x)_{i,j}[\pi(x)_{i,j}[1]] \geq \cdots \geq \hat{f}(x)_{i,j}[\pi(x)_{i,j}[K]]$
  Initialize cumulative sum $S \leftarrow 0$
  for $m = 1$ to K **do**
    $S \leftarrow S + \hat{f}(x)_{i,j}[\pi(x)_{i,j}[m]]$
    if $S \geq \hat{q}$ **then**
      Define prediction set for pixel (i, j) in image x as: $C_{i,j}^{(x)} = \{\pi(x)_{i,j}[1], \ldots, \pi(x)_{i,j}[m]\}$
      break
    end if
  end for
end for

Prediction sets are thus adaptively sized: confident pixels yield smaller sets, while uncertain ones include more classes. This adaptivity enables fine-grained uncertainty estimation, crucial for downstream modules such as risk-aware navigation. The statistical validity of APS ensures that the true class is contained in the prediction set with probability at least $1 - \alpha$ across the image domain.

2.4. Kandinsky Calibration

Kandinsky Calibration [9] enhances pixel-level uncertainty estimation by clustering quantile-based non-conformity curves. It captures structured uncertainty without modifying the model, using a two-step process: learning cluster-specific quantile functions from a calibration set of 1606 images, with the same pixel-wise logits and labels as APS, then assigning calibrated uncertainty values at inference.

Algorithm 2 Kandinsky Calibration and Inference

Require: Calibration set $\mathcal{D}_{\text{calib}}$ with logits $\hat{f}(c)$ and labels $y^{(c)}$, number of clusters n, test images
Ensure: Pixel-level uncertainty values $\{u_{i,j}\}$ for test image x, with $u_{i,j} \in [0,1]$

Calibration phase: Learn uncertainty profiles by clustering pixel-wise non-conformity curves

for all pixels (i,j) in all calibration images $c \in \mathcal{D}_{\text{calib}}$ **do**
  Compute score: $S_{i,j}^{(c)} = -\log \hat{f}(c)_{i,j}^{y_{i,j}^{(c)}}$ using the probability $\hat{f}(c)_{i,j}^{y_{i,j}^{(c)}}$ assigned to the true label $y_{i,j}^{(c)}$
  Normalize the non-conformity score $S_{i,j}^{(c)}$ to $[0,1]$ using min-max normalization
  Extract pixel-wise quantile curve: $Q_{i,j}(\tau) = \text{quantile}(S_{i,j}, \tau)$, with τ the x-axis quantile level
end for
Stack all $Q_{i,j}(\tau)$ vectors into matrix Q
Set number of clusters $n = 4$, selected via Elbow [15] and Silhouette [16] methods
Apply K-means clustering on Q into n groups
for all clusters $c = 1$ to n **do**
  Let $\mathcal{C}_c$ be the set of all pixels assigned to cluster c
  Compute the average quantile curve over $\mathcal{C}_c$: $Q_c(\tau) = \dfrac{1}{|\mathcal{C}_c|} \displaystyle\sum_{(i,j)\in\mathcal{C}_c} Q_{i,j}(\tau)$
end for

Inference phase: Estimate pixel uncertainty by matching to cluster curve and interpolating

for all pixels (i,j) in new test image x **do**
  Predict class probabilities $\hat{f}(x)_{i,j} \in \mathbb{R}^K$
  Set k^* to the index of the class with highest probability: $k^* = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ \hat{f}(x)_{i,j}^{k}$
  Compute score $S_{i,j} = -\log \hat{f}(x)_{i,j}^{k^*}$ and normalize to $[0,1]$
  Match (i,j) to the nearest cluster c based on $Q_{i,j}(\tau)$ curve similarity
  Sample points $\{(S_\ell, u_\ell)\}$ along the cluster-specific quantile curve $Q_c(\tau)$, where S_ℓ are x-axis (score) values and u_ℓ are the corresponding y-axis (uncertainty) values on the curve $Q_c(\tau)$
  Find interval $[S_i, S_{i+1})$ s.t. $S_i \le S_{i,j} < S_{i+1}$
  Interpolate: $t = \dfrac{S_{i,j} - S_i}{S_{i+1} - S_i}$, and then $u_{i,j} = (1-t)\cdot u_i + t \cdot u_{i+1}$
end for
Normalize all $u_{i,j}$ values to $[0,1]$

This algorithm yields a spatially coherent pixel-level uncertainty map. Leveraging clustered quantile curves, it remains lightweight, model-agnostic, and robust to semantic and spatial confidence variations.

2.4.1. Risk Map Construction

The framework combines model confidence and spatial uncertainty to assess terrain safety. For each pixel (i,j), APS define a set $C_{i,j}$ of likely terrain classes whose cumulative predicted probability exceeds the calibrated threshold. Each class k is associated with a user-defined danger score d_k, and its predicted probability $P_{i,j,k}$ is given by the segmentation model. Kandinsky Calibration provides a normalized uncertainty estimate $U_{i,j}$, derived from cluster-wise non-conformity curves and spatial patterns.

The final risk score $F_{i,j}$ integrates both sources of information, amplifying danger in uncertain regions:

$$F_{i,j} = \left(\sum_{k \in C_{i,j}} P_{i,j,k} \cdot d_k \right) \cdot (1 + \lambda \cdot U_{i,j})$$

The parameter $\lambda \ge 0$ adjusts the influence of uncertainty; higher values penalize ambiguous predictions more strongly. This formulation yields high scores in both dangerous and uncertain regions, enabling robust risk-aware navigation across Martian terrain. This risk formulation is novel in that it explicitly fuses coverage guarantees from APS with continuous, spatially coherent uncertainty from Kandinsky, rather than relying on heuristic or variance-only estimates.

2.5. Results

We evaluate the effect of uncertainty quantification on risk-aware planning by analyzing the generation of risk maps. A new image is processed by the segmentation model, yielding pixel-wise probabilities and predicted labels (Figure 1). For each pixel, APS computes a prediction set with $\alpha = 0.05$, whose cardinality encodes model confidence (Figure 2). Kandinsky Calibration then derives pixel uncertainties from cluster-specific quantile curves, highlighting ambiguous regions (Figure 3).

Combining APS sets and Kandinsky uncertainties produces a spatially coherent risk map (Figure 4), which constitutes the core contribution: APS provides calibrated class-level coverage, Kandinsky adds continuous spatial coherence, and their integration yields fine-grained, interpretable risks. A modified A* planner uses these maps to account for epistemic uncertainty, unlike classical planners based only on geometry. The planner runs on an 8-connected grid with edge cost:

$$g((i,j)\to(i',j')) = w_d\, d((i,j),(i',j')) + w_r\, \frac{F_{i,j}+F_{i',j'}}{2},$$

where d is Euclidean distance and $F_{i,j}$ the APS–Kandinsky risk. Pixels above F_{obs} are obstacles, and the heuristic is Euclidean distance to the goal. This preserves admissibility and allows balancing efficiency (w_d) against safety (w_r, λ). Beyond qualitative maps, we position our framework against MC Dropout, Deep Ensembles, and Bayesian SegNet, evaluated on segmentation (mIoU, accuracy), calibration (ECE, NLL, Brier), APS metrics (coverage, set size), and planning (path length, cumulative risk, collisions, goal reach). Our approach uniquely combines APS coverage guarantees with Kandinsky spatial coherence, yielding calibrated, interpretable risk maps that better support downstream planning.

The confidence level α sets APS strictness, with smaller sets at lower values. The uncertainty weight λ penalizes ambiguous areas, and the risk–distance weights w_r, w_d tune safety–efficiency trade-offs. These parameters steer the planner without retraining. Evaluation on lunar-like terrains shows that, despite domain shifts, the system yields coherent risk maps, confirming robustness and transferability.

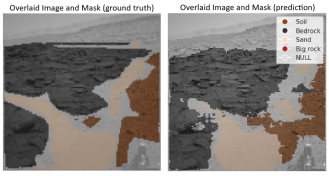

Fig. 1. Predicted classes

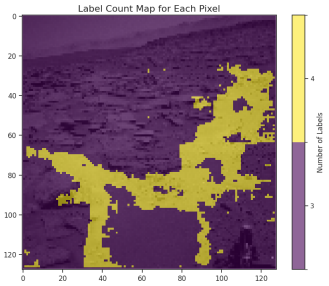

Fig. 2. APS label count map

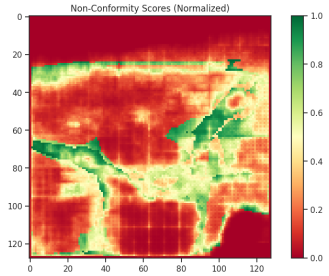

Fig. 3. Kandinsky scores

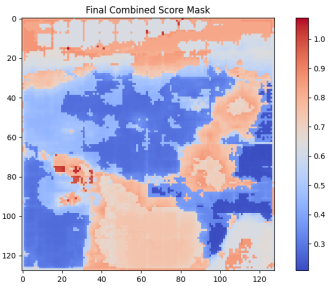

Fig. 4. Final risk map

2.6. Discussion

Combining APS with Kandinsky produces reliable pixel-wise uncertainties: APS captures class-wise ambiguity, Kandinsky exploits spatial regularities, and together they support fine-grained risk assessment. The planner adapts to mission profiles via α, d_k, λ, w_r, and w_d. Compared to MC Dropout, Deep Ensembles, and Bayesian SegNet, our framework unifies statistical validity and spatial coherence, producing calibrated and interpretable risk maps that better support planning. The modified A* integrates geometry and risk, ensuring paths account for terrain danger and uncertainty. The framework generalizes to non-Martian terrains, though parameters are static and chosen offline. Runtime remains acceptable (<45 seconds per risk map on a standard laptop without GPU acceleration), but embedded or time-critical deployment needs further validation.

2.7. Conclusions

This work introduced a pixel-level UQ pipeline for planetary navigation that fuses APS coverage with Kandinsky coherence into a novel risk functional. Beyond qualitative results, it was positioned against MC Dropout, Deep Ensembles, and Bayesian SegNet, showing its distinct advantage in producing calibrated and interpretable risk maps that directly support path planning. The integration with a modified A* demonstrates how uncertainty-aware risk maps guide safe and efficient navigation. The framework is modular, efficient, and generalizes beyond Mars, making it a strong candidate for future missions. Future work should focus on online parameter adaptation, hardware deployment, and broader validation across robotic perception systems under uncertainty.

References

[1] V. Verma, F. Hartman, A. Rankin *et al.*, "Robotic operations during perseverance's first extended mission," in *2025 IEEE Aerospace Conference*. IEEE, 2025. [Online]. Available: https://www-robotics.jpl.nasa.gov/media/documents/2025_IEEE_Aero_RO_ops.pdf

[2] J. M. Wing, "Trustworthy ai," 2020.

[3] H. Delseny, C. Gabreau, A. Gauffriau, B. Beaudouin, and L. Ponsolle, "White paper: Machine learning in certified systems," 2021.

[4] P. Panicucci, "Autonomous vision-based navigation and shape reconstruction of an unknown asteroid during approach phase," 2021.

[5] S. K. Ibrahim, A. Ahmed, M. A. E. Zeidan, and I. E. Ziedan, "Machine learning techniques for satellite fault diagnosis," *Ain Shams Engineering Journal*, vol. 11, pp. 45–56, 2020.

[6] M. H. Shamsi, U. Ali, E. Mangina, and J. O'Donnell, "A framework for uncertainty quantification in building heat demand simulations using reduced-order grey-box energy models," *Applied Energy*, 2020.

[7] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, 2008.

[8] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," in *International Conference on Learning Representations*, 2021.

[9] J. Brunekreef, E. Marcus, R. Sheombarsing, J.-J. Sonke, and J. Teuwen, "Kandinsky conformal prediction: Efficient calibration of image segmentation algorithms," *arXiv preprint arXiv:2311.11837*, 2023.

[10] P. Hart, N. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics*, 1968. [Online]. Available: https://doi.org/10.1109/tssc.1968.300136

[11] R. M. Swan, D. Atha, H. A. Leopold, M. Gildner, S. Oij, C. Chiu, and M. Ono, "Ai4mars: A dataset for terrain-aware autonomous driving on mars," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 1982–1991.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019.

[14] Y. Romano, M. Sesia, and E. J. Candès, "Classification with valid and adaptive coverage," arXiv:2006.02544, 2020.

[15] P. Bholowalia and A. Kumar, "Ebk-means: A clustering technique based on elbow method and k-means++," *International Journal of Computer Applications*, vol. 105, no. 9, pp. 17–24, 2014.

[16] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.