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

Théo Gachet*[1], Luis Mansilla[1], Diane Magnin[1]

[1] *European Space Agency (ESA), Noordwijk, The Netherlands*

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 produces calibrated uncertainty whose marginal coverage guarantee transfers to unseen Martian 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 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.

*Corresponding author. E-Mail: theo.gachet.pro@outlook.fr

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 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. This goes beyond method stacking: unlike variance-only estimators such as MC Dropout or Deep Ensembles, which lack coverage guarantees and ignore spatial structure, our representation is coverage-preserving and spatially coherent by construction, and directly guides path planning.

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 $q_{\mathrm{APS}}(\alpha)$. 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.

3rd Conference on AI in and for Space (SPAICE 2026)

2 Methods

2.1 Data and Model Architecture

We use the AI4MARS dataset [11] (∼35,000 Mars-rover images with 326,000 crowd-sourced labels), retaining 16,064 valid image-label pairs stratified into training, validation, test, and calibration splits. The calibration split is disjoint from training and test, as its exchangeability with the test data underpins the conformal coverage guarantee. The four terrain classes are strongly imbalanced, motivating the stratified sampling and the emphasis on calibrated uncertainty.

We segment with a U-Net [12] and a MobileNetV2 encoder [13], selected for the low computational cost required onboard, trained with standard settings (Adam, sparse categorical cross-entropy, early stopping). On the held-out test set the model reaches 0.83 pixel accuracy and 0.72 F1. We stress that the contribution is the UQ framework, not segmentation accuracy: conformal calibration guarantees valid coverage regardless of model quality, so a moderate backbone is a sufficient and honest testbed.

2.2 Adaptive Prediction Sets

Adaptive Prediction Sets (APS) [8, 14] produce uncertainty-aware predictions with a formal statistical guarantee: by sizing each predicted label set to the model's confidence, APS ensures that the true class is contained in the set with probability at least $1-\alpha$. This guarantee is distribution-free and holds for any underlying model, which makes APS well suited to the data-scarce, domain-shifting conditions of planetary exploration. We adapt APS to pixel-wise segmentation through a single threshold $\hat{q}$, learned from a stratified calibration set of 1606 image-label pairs, over 1.68 billion valid pixels across four balanced terrain classes, and applied at inference to every pixel.

The resulting sets are adaptively sized: confident pixels yield a single label, whereas ambiguous ones accumulate several. This granularity turns the marginal coverage guarantee, which holds across the image domain, into a per-pixel confidence signal, precisely what downstream risk-aware navigation requires. The procedure is summarized in Algorithm 1.

2.3 Kandinsky Calibration

Kandinsky Calibration [9] refines pixel-level uncertainty by clustering quantile-based non-conformity curves, capturing structured uncertainty without altering the model. It proceeds in two steps: it learns cluster-specific quantile functions from the same 1606-image calibration set, and the same pixel-wise logits and labels, as APS, then assigns calibrated uncertainty values at inference. Because neighbouring pixels tend to share non-conformity behaviour, clustering pools their statistics and yields stable estimates even where per-pixel calibration data are sparse.

By sharing statistics across clustered curves, the method produces a spatially coherent uncertainty map while remaining lightweight, model-agnostic, and robust to semantic and spatial variations in confidence. The full calibration and inference procedure is given in Algorithm 2.

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 $C^{(x)}_{i,j}$ 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 $\pi(c)_{i,j}[m] = y^{(c)}_{i,j}$ (true class index)
  Append score: $s^{(c)}_{i,j} = \sum_{k=1}^{m} \hat{f}(c)_{i,j}[\pi(c)_{i,j}[k]]$ to $\mathcal{S}$
Compute quantile threshold: $\hat{q} = \text{Quantile}(\mathcal{S}, \frac{(n+1)(1-\alpha)}{n})$

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: $C^{(x)}_{i,j} = \{\pi(x)_{i,j}[1], \ldots, \pi(x)_{i,j}[m]\}$ and **break**

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
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 $\mathcal{C}_c$: $Q_c(\tau) = \frac{1}{|\mathcal{C}_c|} \sum_{(i,j)\in\mathcal{C}_c} Q_{i,j}(\tau)$

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: $k^* = \arg\max_{k\in\{1,\dots,K\}} \hat{f}(x)_{i,j}^k$
  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: $t = \frac{S_{i,j} - S_i}{S_{i+1} - S_i}$, and then $u_{i,j} = (1-t)\cdot u_i + t \cdot u_{i+1}$
Normalize all $u_{i,j}$ values to $[0,1]$

2.4 Risk Map Construction

The framework assesses terrain safety by fusing model confidence with spatial uncertainty at every pixel. At pixel (i,j), APS return a set $C_{i,j}$ of plausible terrain classes whose cumulative predicted probability exceeds the calibrated threshold, each class k carrying a user-defined danger score d_k, reflecting its intrinsic traversal hazard, and a model-predicted probability $P_{i,j,k}$. Kandinsky Calibration contributes a complementary signal: a normalized uncertainty estimate $U_{i,j}$ derived from cluster-wise non-conformity curves and spatial patterns. The final risk score $F_{i,j}$ combines the two, amplifying danger wherever the model is uncertain:

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

The bracketed term is a coverage-aware expected danger: summing over the APS set $C_{i,j}$ rather than the top-1 class alone, every plausible class of an ambiguous pixel contributes its own risk, weighted by its likelihood and hazard. The factor $(1 + \lambda U_{i,j})$ then inflates this danger where Kandinsky reports high spatial uncertainty, with $\lambda \ge 0$ controlling how strongly ambiguity is penalized. The score is thus high in regions that are dangerous, uncertain, or both, supporting risk-aware navigation. Its novelty lies in explicitly fusing APS coverage guarantees with the continuous, spatially coherent uncertainty of Kandinsky.

3 Results

We evaluate whether the calibrated uncertainty is trustworthy enough to drive navigation, along three axes: the statistical validity of the marginal APS coverage, the robustness of both uncertainty signals to input degradation, and the safety gain the fused risk map brings to path planning. A new image is processed into pixel-wise probabilities and predicted labels. In parallel, APS turns these probabilities into prediction sets whose cardinality encodes confidence (Figure 1), while Kandinsky Calibration produces a spatially coherent uncertainty map $U_{i,j}$ from cluster-specific non-conformity curves (Figure 2). The two signals are fused by the risk functional of Equation (1) into the map of Figure 3. APS contributes calibrated class coverage, Kandinsky adds continuous spatial structure, and their product yields fine-grained, interpretable risk. A modified A* planner then consumes this map to steer paths away from hazardous or uncertain regions, unlike classical planners that reason from geometry alone.

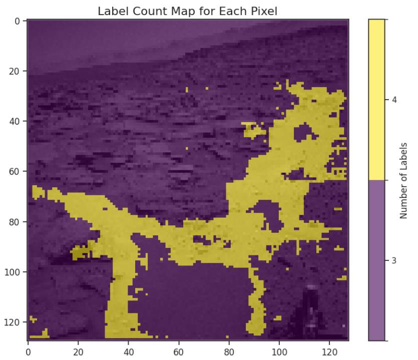

Figure 1: APS label count map

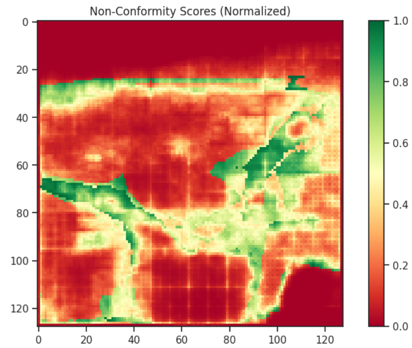

Figure 2: Kandinsky scores

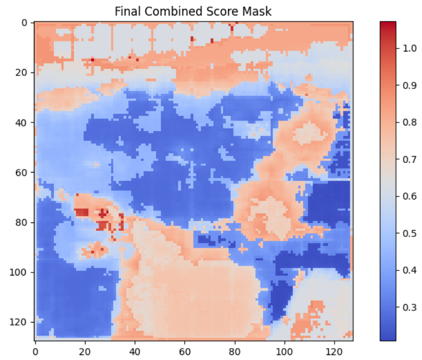

Figure 3: Final risk map

Sensitivity analysis of the APS threshold

APS is calibrated on the stratified calibration split to guarantee a marginal coverage $P(y \in C(x)) \geq 1 - \alpha$ through a single global threshold $\hat{q}$. We verify that this guarantee, fixed on the calibration data, transfers to the held-out test split by sweeping the cumulative-probability threshold and measuring both the empirical marginal coverage and the mean prediction-set size $|\mathcal{S}|$ (Figure 4). Coverage is strictly monotone and near-linear in the nominal operating regime $q \in [0.85, 0.99]$: the calibrated threshold $\hat{q} = 0.889$ (for a target $\alpha = 0.10$) yields an empirical coverage of 93.4%, formally satisfying the safety condition on unseen data. Crucially, this coverage is obtained at a near-unitary mean set size ($|\mathcal{S}| \approx 1.06$): for the vast majority of pixels the model is confident enough to commit to a single class, and only genuinely ambiguous pixels expand the set. APS therefore delivers a valid coverage guarantee without diluting spatial precision, exactly the fine-grained confidence signal the risk map exploits.

Sensitivity to input degradation

A trustworthy uncertainty signal must grow when the input degrades. We probe this by injecting additive Gaussian noise $\mathcal{N}(0, \sigma^2)$ of increasing intensity into the test images and tracking, in parallel, the two signals: the APS mean set size $|\mathcal{S}|$ (model ambiguity) and the mean Kandinsky nonconformity score $U \in [0, 1]$ (spatial uncertainty), shown in Figure 5. Both respond monotonically to the perturbation. APS ambiguity rises from a near-unitary set to $|\mathcal{S}| \approx 3.1$ at $\sigma = 1.0$, with a Spearman correlation to noise intensity of $\rho \approx 0.97$ ($p \ll 10^{-3}$), confirming a near-deterministic coupling between corruption and predictive ambiguity. The mean Kandinsky score rises in step, from 0.35 on clean images to 0.85 under heavy corruption: after a resilience plateau ($\sigma \lesssim 0.4$), where the network absorbs high-frequency noise without semantic drift, it enters a breakdown regime and climbs sharply as the scene becomes unreliable. The two modules thus degrade gracefully and in agreement, each through its own metric.

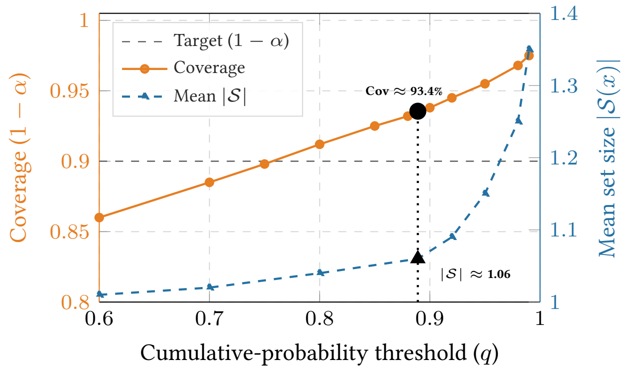

Figure 4: APS sensitivity on the test split. Empirical marginal coverage (orange, left axis) and mean prediction-set size (blue, right axis) as functions of the cumulative-probability threshold q. The calibrated threshold $\hat{q} = 0.889$ (black dotted, target $\alpha = 0.10$) attains 93.4% coverage at a near-unitary set size ($|\mathcal{S}| \approx 1.06$), confirming that the conformal guarantee generalizes to unseen data without inflating the sets.

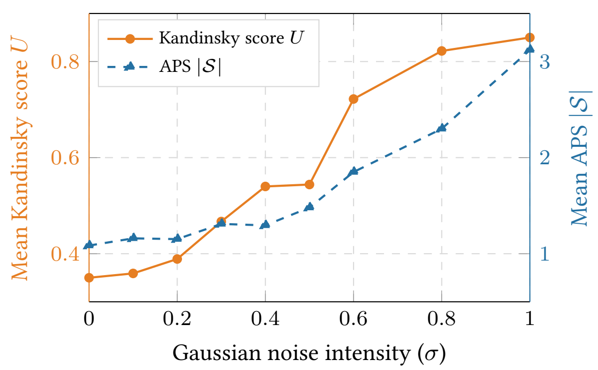

Figure 5: Uncertainty response to input degradation. As Gaussian noise intensity σ increases, both the APS mean set size (blue, right axis) and the mean Kandinsky non-conformity score $U \in [0, 1]$ (orange, left axis) rise monotonically. APS ambiguity correlates with noise at $\rho \approx 0.97$; the mean Kandinsky score climbs from 0.35 on clean images to 0.85 under heavy corruption, past a resilience plateau near $\sigma \approx 0.4$.

Risk-aware planning

A risk map is only useful if it changes what the planner does. We therefore run the same modified A* on three inputs of increasing information (no risk, a hardened geometric margin, and our continuous risk field) and compare the resulting trajectories. The planner searches an 8-connected pixel grid, and the edge between adjacent cells (i, j) and (i', j') costs $g = w_d\, d + w_r\, \frac{1}{2}(F_{i,j} + F_{i',j'})$, where d is their Euclidean separation, F the fused risk of Equation (1), and $w_d, w_r \geq 0$ two operator-set weights balancing travel distance against traversed risk ($w_r{=}0$ for the distance-only baseline). Since the Euclidean-to-goal heuristic bounds only the distance term, it stays admissible and A* remains optimal. Per-class danger scores rank terrain by traversability, $d_{\text{bedrock}}{=}0.1 < d_{\text{soil}}{=}0.2 \ll d_{\text{sand}}{=}0.8 < d_{\text{big rock}}{=}0.9$: firm ground is cheap to cross, sand and rock are expensive.

We draw 500 random start–goal pairs per map and score each path against the AI4Mars ground-truth masks, used as an *obstacle oracle* (a cell is a true obstacle when its class is high-danger, i.e. *sand* or *big rock* with $d_k \geq 0.8$), so that scoring never leaks into planning. Each path yields four numbers. Goal reach is the fraction of queries that return a navigable path. The critical collision rate T_C is the share of goal-reaching paths that cross a true obstacle (it captures the silent failure a rover must avoid, where the planner reports success yet the route would strike a real hazard). The mean traversed risk $\bar{F}$ is the risk of Equation (1) averaged over the path's cells. Finally, the path-length ratio $\eta \geq 1$ is the planned length divided by the straight-line distance, i.e. the detour price of safety.

Planner	Goal reach	T_C (%)	$\bar{F}$	η
Distance-only A*	0.99	41.7	0.63	1.02
Geometric inflation	0.86	22.9	0.48	1.31
Risk-aware (ours)	0.97	**9.4**	**0.27**	1.18

Table 1: Planner comparison over 500 Monte-Carlo start–goal queries, scored against the AI4Mars obstacle oracle. Goal reach, critical collision rate T_C, mean traversed risk $\bar{F}$, and path-length ratio η for a distance-only baseline, geometric inflation, and our risk-aware planner.

The three planners trace a safety–efficiency frontier on which our method dominates. Distance-only A* is fast and near-geodesic ($\eta{=}1.02$, 99% goal reach) but terrain-blind, driving through a true obstacle in 41.7% of its successful runs. Geometric inflation fails on both sides at once. By dilating obstacles by a fixed radius, it condemns genuinely free space, disconnecting 14% of queries (86% goal reach) and stretching paths by 31% ($\eta{=}1.31$), a spatial false-positive blocking akin to the freezing-robot problem. Yet it still collides in 22.9% of cases, because a binary margin cannot express that a cell is only *mildly* risky or uncertain. Our continuous, coverage-aware field keeps goal reach at 97%, more than halves the traversed risk ($\bar{F}$ from 0.63 to 0.27), and cuts critical collisions to 9.4%, a factor of 4.4 below distance-only and 2.4 below inflation, for an 18% detour. Safety here comes from *grading* risk continuously, not from *hardening* geometry.

4 Discussion

Combining APS with Kandinsky yields reliable pixel-wise uncertainty: APS bounds class-wise ambiguity with a finite-sample coverage guarantee, Kandinsky adds spatially coherent structure, and their fusion supports fine-grained risk assessment. The APS sweep (Figure 4) confirms that the marginal guarantee transfers to unseen data at near-unitary set size, and under injected noise (Figure 5) both signals grow monotonically and in agreement, evidence that the estimated uncertainty tracks genuine input degradation rather than a modelling artefact. Fed to the modified A*, this risk field cuts the critical collision rate several-fold over distance-only and inflation baselines (Table 1) at a modest detour. The fusion thus retains the conformal coverage guarantee while adding spatial coherence, two properties that variance-only estimators such as MC Dropout or Deep Ensembles do not jointly provide.

Several limitations remain. The danger scores and the weights (λ, w_r, w_d) are set offline and held fixed during traversal, so the planner cannot yet adapt them to conditions encountered en route. The APS coverage guarantee is marginal over the image domain and assumes exchangeability between calibration and test pixels, which domain shift to lunar or terrestrial scenes may weaken even where the maps stay qualitatively coherent. Finally, runtime, though acceptable for the rover's slow stop-and-image duty cycle (under one minute per risk map on a standard laptop without GPU, dominated by corridor width), has not been validated on flight-representative embedded hardware.

5 Conclusions

We presented a pixel-level UQ pipeline for planetary navigation that fuses APS coverage with Kandinsky's spatial coherence into a single risk functional driving a risk-aware A*. Across our experiments the calibrated uncertainty proved trustworthy: the marginal guarantee held on unseen data at near-unitary set size, both signals responded monotonically to input degradation, and the resulting risk field reduced critical collisions several-fold against distance-only and inflation baselines at a modest path-length cost, behaving predictably under every operator parameter and yielding coherent maps across Martian, lunar and Earth-like terrain. The framework thus offers the coverage guarantees and spatial coherence that variance-only baselines lack, and transfers across domains without modification. Future work targets online adaptation of the danger and weighting parameters, deployment on flight-representative hardware, and validation across broader robotic perception systems under uncertainty.

References

1. Verma, V., Hartman, F., Rankin, A., *et al. Robotic Operations During Perseverance's First Extended Mission* in *2025 IEEE Aerospace Conference* (2025). https://www-robotics.jpl.nasa.gov/media/documents/2025_IEEE_Aero_RO_ops.pdf.
2. Wing, J. M. *Trustworthy AI* 2020.
3. Delseny, H., Gabreau, C., Gauffriau, A., Beaudouin, B. & Ponsolle, L. *White Paper: Machine Learning in Certified Systems* 2021.
4. Panicucci, P. *Autonomous vision-based navigation and shape reconstruction of an unknown asteroid during approach phase* 2021.
5. Ibrahim, S. K., Ahmed, A., Zeidan, M. A. E. & Ziedan, I. E. Machine Learning Techniques for Satellite Fault Diagnosis. *Ain Shams Engineering Journal* **11,** 45–56 (2020).
6. Shamsi, M. H., Ali, U., Mangina, E. & O'Donnell, J. A framework for uncertainty quantification in building heat demand simulations using reduced-order grey-box energy models. *Applied Energy* (2020).
7. Shafer, G. & Vovk, V. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research* (2008).
8. Angelopoulos, A. N., Bates, S., Malik, J. & Jordan, M. I. *Uncertainty Sets for Image Classifiers Using Conformal Prediction* in *International Conference on Learning Representations* (2021).
9. Brunekreef, J., Marcus, E., Sheombarsing, R., Sonke, J.-J. & Teuwen, J. Kandinsky Conformal Prediction: Efficient Calibration of Image Segmentation Algorithms. *arXiv preprint arXiv:2311.11837* (2023).
10. Hart, P., Nilsson, N. & Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on Systems Science and Cybernetics.* https://doi.org/10.1109/tssc.1968.300136 (1968).
11. Swan, R. M. *et al. 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), 1982–1991.
12. Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation* 2015. arXiv: 1505.04597 [cs.CV].
13. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. *MobileNetV2: Inverted Residuals and Linear Bottlenecks* 2019. arXiv: 1801.04381 [cs.CV].
14. Romano, Y., Sesia, M. & Candès, E. J. *Classification with Valid and Adaptive Coverage* arXiv:2006.02544. 2020.
15. Bholowalia, P. & Kumar, A. EBK-means: A clustering technique based on elbow method and k-means++. *International Journal of Computer Applications* **105,** 17–24 (2014).
16. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65 (1987).

The 2$^{\text{nd}}$ 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 2$^{\text{nd}}$ 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 $q_{\mathrm{APS}}(\alpha)$. 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 ∈ D_calib do
    Compute score: S_{i,j}^{(c)} = −log f̂(c)_{i,j}^{y_{i,j}^{(c)}} using the probability f̂(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}(τ) = quantile(S_{i,j}, τ), with τ the x-axis quantile level
end for
Stack all Q_{i,j}(τ) 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 C_c be the set of all pixels assigned to cluster c
    Compute the average quantile curve over C_c: Q_c(τ) = 1/|C_c| Σ_{(i,j)∈C_c} Q_{i,j}(τ)
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 f̂(x)_{i,j} ∈ ℝ^K
    Set k* to the index of the class with highest probability: k* = argmax_{k∈{1,...,K}} f̂(x)_{i,j}^k
    Compute score S_{i,j} = −log f̂(x)_{i,j}^{k*} and normalize to [0, 1]
    Match (i, j) to the nearest cluster c based on Q_{i,j}(τ) curve similarity
    Sample points {(S_ℓ, u_ℓ)} along the cluster-specific quantile curve Q_c(τ), where S_ℓ are x-axis
    (score) values and u_ℓ are the corresponding y-axis (uncertainty) values on the curve Q_c(τ)
    Find interval [S_i, S_{i+1}) s.t. S_i ≤ S_{i,j} < S_{i+1}
    Interpolate: t = (S_{i,j} − S_i)/(S_{i+1} − S_i), and then u_{i,j} = (1 − t) · u_i + t · 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 $\lambda \geq 0$ 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.

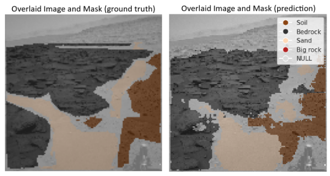

Fig. 1. Predicted classes

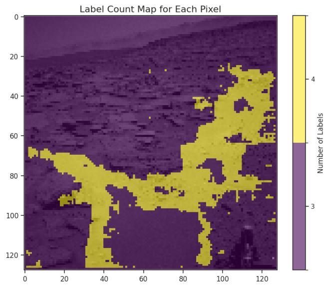

Fig. 2. APS label count map

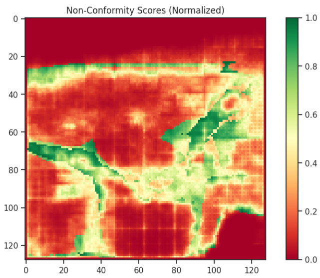

Fig. 3. Kandinsky scores

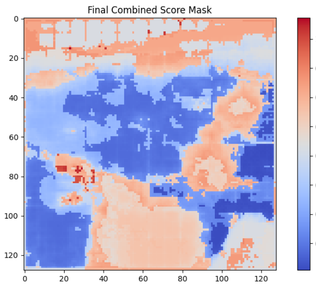

Fig. 4. Final risk map

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.