Response to the associate editor

Dear Editor and Reviewers, We appreciate your careful review and valuable comments on our manuscript, entitled “A Probabilistic Approach to Wildfire Spread Prediction Using a Denoising Diffusion Surrogate Model” (egusphere-2025-2836). We have thoroughly considered all the points and revised our manuscript accordingly. We believe the quality of the revised manuscript is improved significantly, and both goals and results are now presented in a more accurate and clearer way.

In this cover letter, we provide a detailed, point-by-point response to the reviewers’ comments, outlining the revisions we have made. For ease of reference, we have numbered the comments. To make the modifications easily identifiable, we have highlighted the changes in Blue in the revised manuscript.

Thank you for your time and consideration. We appreciate the opportunity to improve our work based on your valuable feedback.

Response to referees

Reply to referee 1

We thank the Reviewer for the comments and suggestions on our manuscript.

We have responded to most of the Reviewer’s comments in our online reply dated 9 August 2025 (at the end of this letter, for reference). Following the Reviewer’s suggestion, we conducted comprehensive numerical experiments using three different neural network architectures, namely a UNet with attention mechanism, a standard UNet, and a convolutional autoencoder with ResBlocks. Each network was trained both deterministically and using diffusion-based training. The numerical results are provided in Appendix D (p. 32) of the revised manuscript and are summarised below.

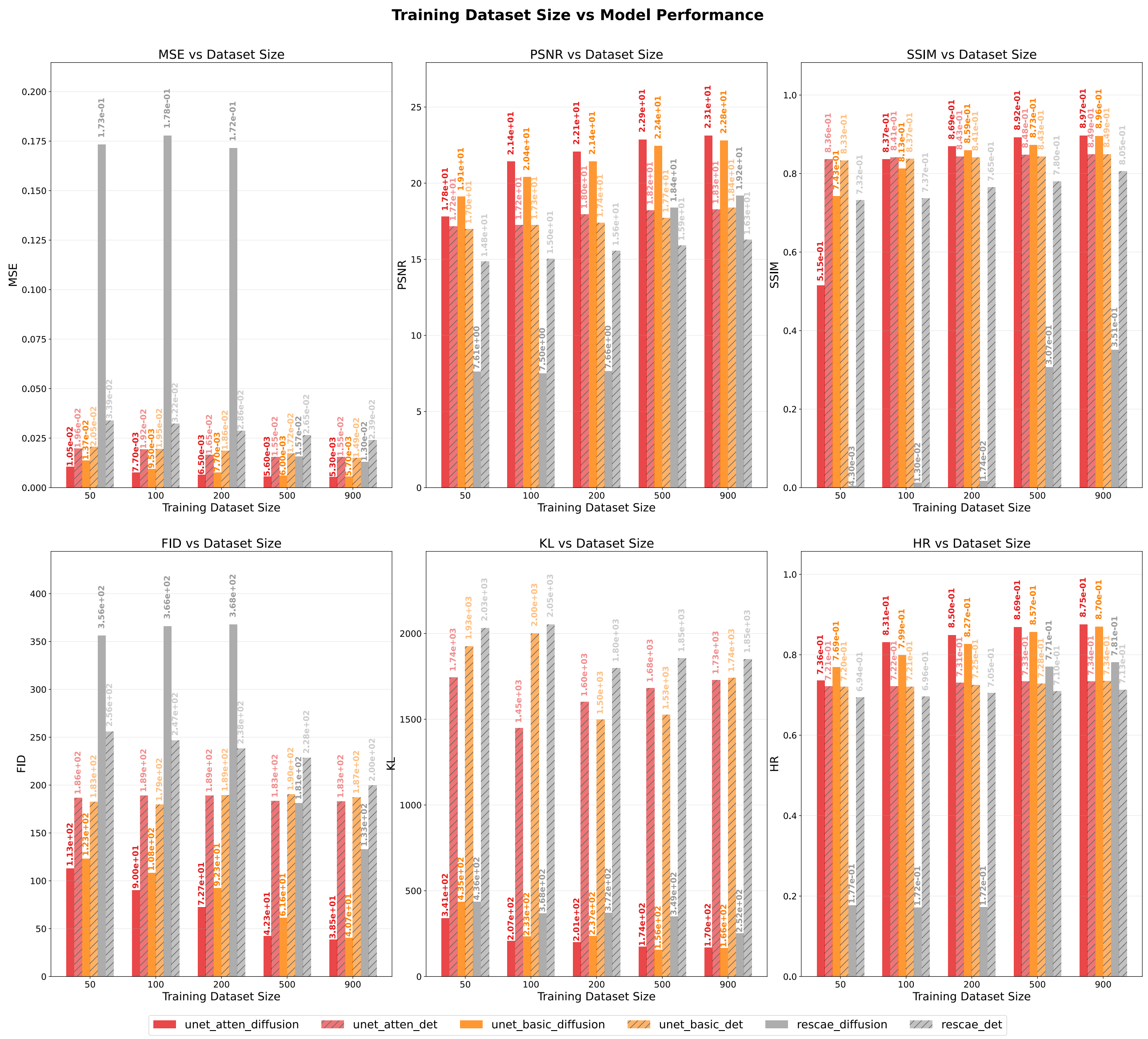

Figure 1: Ablation study comparing model performance across different training dataset sizes and architectures.

Our ablation study evaluates three distinct neural network architectures across varying training dataset sizes {50, 100, 200, 500, and 900 samples}, comparing stochastic diffusion models against deterministic approaches, as shown in Figure 1. The architectures under scrutiny include: (1) a UNet architecture with attention mechanisms (unet_atten) as employed in our main study; (2) a UNet basic architecture (unet_basic), representing a simplified variant without attention blocks; and (3) a residual AutoEncoder architecture (rescae), which maintains a similar network structure to the UNet but removes skip connections to assess their contribution to model performance. All models were trained using identical hyperparameter configurations, including a learning rate of $1e{-5}$ (selected from candidates $1e{-3}, 1e{-4}, 1e{-5}$ as the optimal choice), 200 training epochs, and the AdamW optimiser with weight decay of $1e{-4}$. All models are evaluated on the ensemble test dataset of the Chimney fire event.

The results in Figure 2 demonstrate that the UNet Architecture with attention mechanisms generally outperforms the UNet basic architecture without attention blocks, though the improvement varies across metrics and dataset sizes. For instance, at training dataset size 500, the attention-enhanced UNet achieves superior performance in MSE (5.60×10^{-3} vs 6.00×10^{-2}), SSIM (8.92×10^{-1} vs 8.73×10^{-1}), and FID (4.23×10^{1} vs 6.16×10^{1}) whilst showing comparable performance in other metrics such as PSNR. However, at smaller dataset sizes like

100, the performance differences become less pronounced, with some metrics showing marginal improvements whilst others exhibit comparable or slightly inferior performance. In contrast, the comparison between UNet architectures and the residual AutoEncoder architecture reveals more substantial performance differences. The UNet structures consistently demonstrate significant improvements across multiple metrics, highlighting the importance of skip connections in preserving fine-grained information throughout the encoding-decoding process.

The related source code, scripts, data, and experimental results have been uploaded to Zenodo (`https://zenodo.org/records/15699653`) (Yu et al., 2025). The experimental results can be found in the `out` directory.

Following our previous response, it is important to clarify that the primary focus of this study is not to establish the superiority of various model architectures, but rather to investigate the advantages of diffusion-trained stochastic models over deterministic models in simulating wildfire uncertainty. This finding is further supported by our new numerical experiments: across all three neural network architectures, diffusion-based ensemble predictions (bar chart with slashes) substantially outperform their deterministic counterparts (bar chart without slashes) as shown in Figure 1.

Our previous response (copied here for reference)

We thank the reviewer for the detailed comments and suggestions on our manuscript. However, we believe that some key aspects of our work may have been overlooked.

1. The reviewer has repeatedly suggested that a benchmark comparing "DDPM vs. ConvLSTM (or other NNs)" is necessary for the manuscript. We would like to first clarify that sampling algorithms (e.g., DDPM or DDIM) and neural network architectures (e.g., ConvLSTM or U-Net) are fundamentally two different things. Sampling algorithms define how noise is added during the forward diffusion process and removed during the reverse (denoising) process (e.g., markovian in the case of DDPM and deterministic non-markovian in the case of DDIM, see Austin et al. (2021); Song et al. (2022), whereas different neural network architectures (such as U-Net or Transformer) could be chosen to train this denoising procedure. We believe the reviewer may be confused about this fundamental concept. We can not compare a sampling/denoising method against a neural network structure.

The main objective of our experiments here is to compare a diffusion-based generative training algorithm with the deterministic training method (based on MSE) for wildfire prediction, rather than to evaluate different neural network architectures. Therefore, we compared the performance of a conditional diffusion model based on U-Net to that of the same U-Net trained using a deterministic approach. In addition, using a different neural network architecture might improve the accuracy of deterministic training, but it would not provide probabilistic predictions or capture the uncertainty of fire propagation. And also, the new network architecture will likely improve the diffusion model's performance as well. This does not qualitatively affect our comparison of diffusion and deterministic training.

We thank the reviewer for this question and will clarify the differences between neural network architectures and diffusion sampling algorithms for non-expert readers in ML.

2. Regarding the novelty of our work, although we agree that diffusion models have recently been applied in geoscience, to the authors' knowledge, this is the first study to apply diffusion-based generative AI to wildfire spread prediction (see a recent review paper by Xu et al. (2025)). In fact, to our knowledge, only one previous GMD publication (Elena Tomasi et al., April 2025) has applied a latent diffusion model to a downscaling task. Therefore, we believe that our paper is the first to use a conditional diffusion model for dynamical-system prediction in GMD.

More importantly, our diffusion model is trained using data generated from a stochastic simulator of wildfire. Therefore, we examine if the ensemble generated by the diffusion model could represent the stochasticity of the original physics model, which brings a unique contribution and insight to the community. We have also designed a specific validation procedure to compare the two ensembles generated by the stochastic physics model and the diffusion AI model, as described in Section 2.2.2 and illustrated in Figures 3 and 7 of our manuscript.

We believe that developing a surrogate model using diffusion-based generative method to capture uncertainties in stochastic physics simulators is novel within geoscience, if not in the broader computational physics field.

Following the reviewer's suggestion, we will perform additional hyperparameter tuning in the revised manuscript to improve our diffusion model's performance. However, as noted, our primary objective is to demonstrate a generative diffusion model's ability to capture the stochasticity of the physics-based model, which our current results already successfully achieve.

3. The reviewer repeatedly refers to DDPM as our denoising approach and points out its computational inefficiency. However, in our manuscript we employ the DDIM algorithm, as clearly stated in the first sentence of Section 3.1.4, in Equation 9 on page 14, and in Algorithm 2 on page 15. We also explain our choice of DDIM over DDPM, indeed specifically for its superior computational efficiency, in Section 3.1 on page 15. Thus, we believe the reviewer may have overlooked some important statements in our methodology section.

Reply to referee 2

0.1 General Comments:

In this paper, a probabilistic method using a denoising diffusion surrogate model is applied to study the wildfire spread prediction, which has the advantage of quantifying the uncertainty. The study focuses on synthetic wildfire data generated by a probabilistic cellular automata-based simulator. The study is systematic, and the presentation of the results is detailed. I have a few minor suggestions, especially several clarification questions.

We thank the Reviewer for the positive comments and the appreciation of our work. We have carefully addressed all the comments with additional numerical experiments, and we revised the manuscript accordingly. A point-to-point response is listed here below.

0.2 Specific Comments:

1. The authors highlighted that "this study seeks to address the limitations of traditional deterministic wildfire forecasting methods." What about the existing stochastic or probabilistic models?

   We agree with the Reviewer that stochastic models are widely used in wildfire modelling to capture extreme fire behaviour. In fact, the training and test data used in this work is generated via a physics-based stochastic cellular-automata fire simulator Alexandridis et al. (2008). Our goal here is not to compare conventional stochastic fire predictors with deep learning–based ones but rather to investigate whether a generative machine learning model can effectively simulate wildfire propagation dynamics by learning from and reproducing the stochastic behaviour of a physics-based CA model.

   Accordingly, we thoroughly compare the outputs of the proposed diffusion-based wildfire predictor against the original CA model and show that, with the diffusion approach, we can numerically represent the probability density function of the CA outputs. By contrast, conventional deterministic models typically predict only the mean of possible scenarios and therefore lose the ability to capture extreme fire events. Following the Reviewer's suggestion, we have highlighted this aspect in the introduction of the revised manuscript (page 4) "*Evaluation uses data from a probabilistic cellular-automata emulator incorporating canopy cover, canopy density, and slope. We analyse the stochastic outputs to assess whether the diffusion model captures the physics model's uncertainties.*"

2. The authors may add more explicit statements to highlight the novelty of this work. Is this just an application or are there existing improvements in the techniques?

   This study is, to our knowledge, the first to apply diffusion-based generative AI to wildfire spread prediction (Xu et al., 2025). More importantly, our model is trained on data from a stochastic wildfire simulator, allowing us to test whether the diffusion model's ensemble reproduces the stochasticity of the original physics model. We designed a dedicated validation procedure to compare ensembles from both models, as detailed in Section 2.2.2 and shown in Figures 3 and 7. We believe that developing a surrogate model using diffusion-based generative methods to capture uncertainties in stochastic physics simulators is novel in computational science. Following the Reviewer's suggestion, we have added a paragraph in the introduction of the revised manuscript (page 1) to highlight it "*To the authors' knowledge, no existing work has used diffusion-based generative models to predict fire spread in the literature. Furthermore, we believe that employing such surrogates to capture uncertainties in stochastic physics simulators is novel in computational science.*"

3. The interpretability of probabilistic forecasting needs more discussion. These forecasts indeed provide UQ. But is such a UQ reliable and accurate?

   We have thoroughly compared the estimated probability distribution of our generative model with that of the physics-based model to ensure the diffusion model provides accurate uncertainty quantification of possible fire-spread scenarios. Following the Reviewer's suggestion, we have highlighted the comparison between the ensemble diffusion model's UQ and the numerical UQ of the original physics-based CA model in the revised manuscript (page 22) "*In particular, the strong FID and KL results indicate that the diffusion model's estimated probability distribution is reliable, as it closely matches that of the original physics-based CA model.*".

4. The physical mechanism is quite complicated, and therefore several variables are involved in the models. How sensitive is the diffusion emulator with respect to the perturbation of each parameter/input?

   In fact, to generate different scenarios for the dataset, we randomised the initial parameters following our previous work (Cheng et al., 2022). Consequently, both the training and test datasets contain fire scenarios generated with different initial parameters. We have added a description in section 2.1 (page 7) of the revised manuscript to clarify this. "*The operational parameters p_h, a, c_1, and c_2 influence the fire forecast, where a is the slope effect coefficient and c_1, c_2 are the wind effect coefficients. The detailed formulations of the slope and wind effects are described in* Cheng et al. *(2022)...Training data is generated via Latin Hypercube Sampling (LHS) within the range of an ensemble of perturbed parameter sets:*"

   $$p_h \in [0.00,\ 0.35],\ a \in [0.00,\ 0.14],\ c_1 \in [0.00,\ 0.12],\ c_2 \in [0.00,\ 0.40] \quad (1)$$

   where the parameter ranges are based on the previous study by Cheng et al. *(2022).*

   Following the Reviewer's question, we have also added an analysis in page 23 of the manuscript "*It is also worth mentioning that the CA model parameters p_h, a, c_1, and c_2 are randomly perturbed when generating the training and test datasets. The numerical results presented in Table 3 further demonstrate the robustness of the proposed diffusion model against variations in fire modelling parameters.*"

5. The role of some of the details of the emulator's components needs to be discussed. For example, what if the attention mechanism is removed?

We thank the Reviewer for pointing out this question regarding the architectural components of our model. Following the Reviewer's suggestion, we have conducted an ablation study examining the role of attention mechanisms and other architectural details, which we present in the Appendix D: Ablation study on model architecture of our paper and illustrate in Figure D1 (page 32) of the revised manuscript, summarised here below.

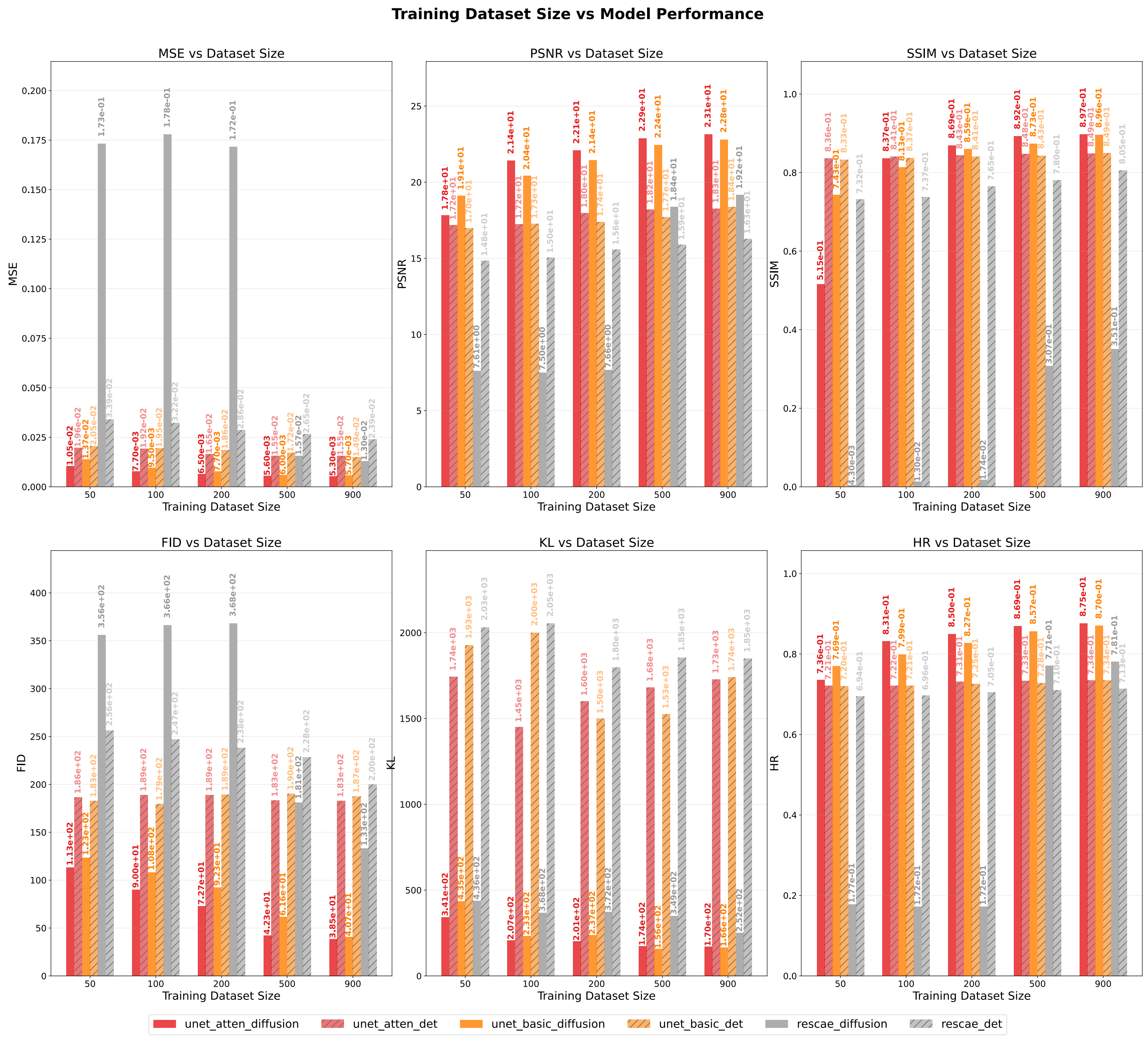

Figure 2: Ablation study comparing model performance across different training dataset sizes and architectures.

Our ablation study evaluates three distinct neural network architectures across varying training dataset sizes {50, 100, 200, 500, and 900 samples}, comparing stochastic diffusion models against deterministic approaches, as shown in Figure D1. The architectures examined include: (1) a UNet architecture with attention mechanisms (unet_atten) as employed in our main study; (2) a UNet basic architecture (unet_basic), representing a simplified variant without attention blocks; and (3) a residual AutoEncoder architecture (rescae), which maintains a similar network structure to the UNet but removes skip connections to assess their contribution to model performance. All models were trained using identical hyperparameter configurations, including a learning rate of $1e{-}5$ (selected from candidates $1e{-}3, 1e{-}4, 1e{-}5$ as the optimal choice), 200 training epochs, and the AdamW optimiser with weight decay of $1e{-}4$. All models are evaluated on the ensemble test dataset of the Chimney fire event.

The related source code, scripts, data, and experimental results have been uploaded to Zenodo (`https://zenodo.org/records/15699653`) (Yu et al., 2025). The experimental results can be found in the `out` directory.

Turning to the specific question about attention mechanisms, our experimental results demonstrate that the attention-enhanced UNet generally outperforms the UNet basic architecture without attention blocks, though the benefits vary across metrics and dataset sizes, as shown in Figure 1. It is worth noting that at smaller dataset sizes (e.g., 100 samples), the performance differences between architectures with and without attention become less pronounced. This observation may partially be attributed to the specific characteristics of our experimental data, including its moderate spatial resolution (64×64), which might not fully utilise the representational capabilities of attention mechanisms. We suggest that in scenarios involving more complex spatial patterns or higher-resolution data, the benefits of attention mechanisms could potentially be more pronounced.

Our ablation study demonstrates that whilst architectural improvements (such as attention mechanisms) provide meaningful enhancements, the more substantial performance gains arise from the fundamental shift from deterministic to stochastic modelling approaches. The consistent superiority of diffusion models across all architectural variants reinforces our central ideas that stochastic models are inherently better suited for capturing the uncertainty and variability characteristic of wildfire behaviour.

6. Some details about the background should be added. For example, subsampling frames at 20-hour intervals is used for training. Why is such a specific number chosen? There are a lot of mathematical details, but some of the physics or reasoning are missing.

   We thank the Reviewer for pointing this out. The 20-hour interval was chosen to ensure sufficiently large time steps for observing meaningful differences across fire-spread stages. This time interval is consistent with recent research works (Kondylatos et al., 2022; Huot et al., 2022). Following the Reviewer's suggestion, we have added a paragraph in the revised manuscript (page 9) to explain the reason of this choice "*To enlarge the prediction window and maintain substantial differences between successive fire states, frames are subsampled from each simulation at intervals of 10 time steps (20 hours), yielding six frames per wildfire event.*"

Bibliography

Alexandridis, A., Vakalis, D., Siettos, C. I., and Bafas, G. V.: A Cellular Automata Model for Forest Fire Spread Prediction: The Case of the Wildfire That Swept through Spetses Island in 1990, Applied Mathematics and Computation, 204, 191–201, https://doi.org/10.1016/j.amc.2008.06.046, 2008.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and van den Berg, R.: Structured Denoising Diffusion Models in Discrete State-Spaces, in: Advances in Neural Information Processing Systems, vol. 34, pp. 17 981–17 993, Curran Associates, Inc., 2021.

Cheng, S., Jin, Y., Harrison, S. P., Quilodrán-Casas, C., Prentice, I. C., Guo, Y.-K., and Arcucci, R.: Parameter Flexible Wildfire Prediction Using Machine Learning Techniques: Forward and Inverse Modelling, Remote Sensing, 14, 3228, https://doi.org/10.3390/rs14133228, 2022.

Huot, F., Hu, R. L., Goyal, N., Sankar, T., Ihme, M., and Chen, Y.-F.: Next day wildfire spread: A machine learning dataset to predict wildfire spreading from remote-sensing data, IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13, 2022.

Kondylatos, S., Prapas, I., Ronco, M., Papoutsis, I., Camps-Valls, G., Piles, M., Fernández-Torres, M.-Á., and Carvalhais, N.: Wildfire danger prediction and understanding with deep learning, Geophysical Research Letters, 49, e2022GL099 368, 2022.

Song, J., Meng, C., and Ermon, S.: Denoising Diffusion Implicit Models, https://doi.org/10.48550/arXiv.2010.02502, 2022.

Xu, Z., Li, J., Cheng, S., Rui, X., Zhao, Y., He, H., Guan, H., Sharma, A., Erxleben, M., Chang, R., et al.: Deep learning for wildfire risk prediction: Integrating remote sensing and environmental data, ISPRS Journal of Photogrammetry and Remote Sensing, 227, 632–677, 2025.

Yu, W., Ghosh, A., Finn, T., Arcucci, R., Bocquet, M., and Cheng, S.: A Probabilistic Approach to Wildfire Spread Prediction Using a Denoising Diffusion Surrogate Model, https://doi.org/10.5281/zenodo.15699653, 2025.