Interpreting a Mirage: Lessons from a Design Study Toward Synthetic Weather Visualizations

Steven R. Gomez
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
steven.gomez@mit.edu

Kevin K. Nam
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
kevin.nam@mit.edu

CHI'22, April 30–May 5, 2022, New Orleans, LA, USA
ACM .

Abstract
Despite rapid advances in Explainable AI (XAI) techniques in recent years, there exist few guidelines for designing interfaces that help users interpret complex Machine Learning (ML) outputs—like synthetic images—in operational settings. In this paper, we present preliminary lessons from a design study toward helping operators interpret ML-generated weather-radar visualizations for forecasting tasks. We find that these synthesized image outputs create unique interpretability challenges and opportunities compared to other ML outputs like classifier predictions, where more conventional XAI tools could be applied.

Author Keywords
Explainable AI; interpretability; visualization; weather.

CCS Concepts
•Human-centered computing → Human computer interaction (HCI);

Introduction
Explainable AI (XAI) aims to provide insights about the inner workings and trustworthiness of predictive models that are otherwise opaque to users. Earlier studies have demonstrated the benefits of XAI to improve model debugging tasks, where the intended end user has significant knowledge about Machine Learning (ML), or for helping

Weather Radar ML Tasks

We considered uses of ML to synthesize VIL visualizations for two kinds of tasks.

Precipitation nowcasting: This task involves synthesizing an accurate short-term forecast [4], for example by extrapolating VIL measurements using optical flow. Datasets: `rainymotion` optical flow library and data (github.com/hydrogo/rainymotion)

Image-to-image translation: This task involves synthesizing a VIL map at a time point using other non-radar signals (for example, if the traditional radar is unavailable). Recently this has been studied as an image-to-image style transfer problem [12] from image sets like optical and infrared satellite series to corresponding VIL maps. Datasets: SEVIR data with registered spatiotemporal modalities (sevir.mit.edu).

data-domain decision makers reason about classifiers in areas like financial decision making or medical diagnostics. However, other ML models that produce high-dimensional outputs—like image regression—can be useful to decision makers but do not easily lend themselves to common explainer techniques or off-the-shelf XAI tools like LIME [8] or SHAP [7]. In designing interpretability tools for these models, we must understand what kinds of questions and goals domain end users ("operators") have about these systems and how they assess the outputs.

We explored these questions in the context of ML-based weather radar synthesis, which is a potentially impactful tool for forecasters. In this paper we share our top lessons discovered through conversations about tools and tradecraft with subject-matter experts in weather forecasting and ML, and through design activities toward user interfaces that help operators interpret the reliability of ML outputs. Three lessons illustrated by this work include:

1. **Do** help the operator use her mental model of data relationships to judge the trustworthiness of the ML prediction.
2. **Do** communicate new explanatory information (e.g., uncertainty) about the ML outputs at a mission-specific resolution.
3. **Do not** impose user interactions for XAI that require the operator to possess extensive new ML or data-domain knowledge.

Synthetic Weather Radar

A promising use of ML is estimating or augmenting traditional sensors for decision support in data domains where performance and availability are critical. In weather forecasting, vertically integrated liquid water (VIL) is an indication of the amount of precipitation in a vertical column, usually produced from radar networks (e.g., Next-Generation Radar, or NEXRAD). VIL map visualizations (e.g., Figure 1, box A) are widely used by aviation weather forecasters [9] to understand the intensity and trajectory of weather systems. Other observations, like lightning strikes in the area and infrared or optical satellite data, are complementary to the radar data and may be correlated with the VIL estimates. As a result, recent work has looked at using neural networks to synthesize radar outputs like VIL using spatiotemporal inputs from these other sensors, which is useful when radar readings are unavailable at a location or time [12].

Some benefits of using ML in synthesis tasks (see "Weather Radar ML Tasks" sidebar) are faster updates or improvements to short-term forecast accuracy, or estimating and visualizing VIL when it is otherwise unavailable. The ability to produce these images with visual encodings that are nearly identical to traditional radar VIL maps is a double-edged sword: weather forecasters already know how to interpret the encodings so there is no additional learning curve, but that familiarity could mask new risks (sources of uncertainty in the ML, assumptions about the origin of the data, etc.).

Developer and Operator Needs

To better understand needs for interpretability with these models, we spoke with a small group of potential users—weather forecasters whose role is primarily to advise pilots in navigating around weather—as well as developers of ML-based synthetic radar tools. Our goal was to get high-level information about the tools, tradecraft, and how tasks are performed typically. While not exhaustive, this section lays out several important considerations we took away.

First, the forecasters we spoke with need to know whether radar outputs (ML-based or traditional) are reliable at oper-

Target Analysis Questions

- **Q1** *What other weather observations would support or call this radar image into question?*
- **Q2** *Is this radar consistent with what the system predicted in the past under similar conditions? (Was it correct then?)*
- **Q3** *How confident is the AI that the weather will interfere with a given flight path?*

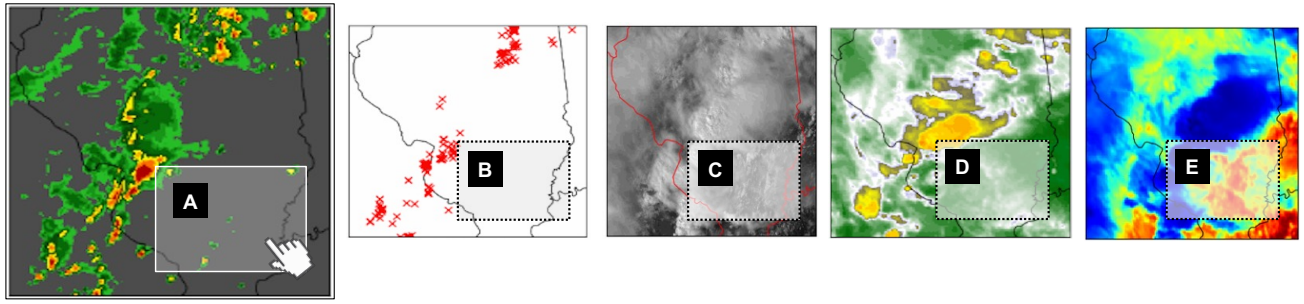

Figure 1: A concept for linked, multi-view visualizations of the predicted VIL (A) and the model inputs. Input modalities include (B) lightning strikes, (C) optical satellite, (D) infrared satellite (6.9 μm), and (E) infrared satellite (10.7 μm) imagery. Map thumbnails are modified from examples in the SEVIR Tutorial [1].

ation time. They have some bandwidth for additional verification of model outputs that might contain errors. Helping operators interpret reliability and build trust is a higher priority than understanding the mechanisms of the underlying model that computes the radar image. Suresh et al. note that these are two top-level goals for XAI (trust building and model understanding) that can be difficult to distinguish precisely [11], but may be influenced by the operators' ML knowledge and primary decision tasks.

Second, ML developers and forecaster operators have different measures of success for these tools. Developers said they evaluated a VIL synthesis model by calculating differences between the image outputs and VIL data from a controlled dataset, which they treated as ground truth. Forecasters interpret radar images in operational settings where ground truth does not exist, and were concerned with ensuring that pilots have the necessary weather information to safely navigate an aircraft. As a result, errors in the imagery farther away from the flight path may be less of a concern than ones near it. In general, forecasters want explanations

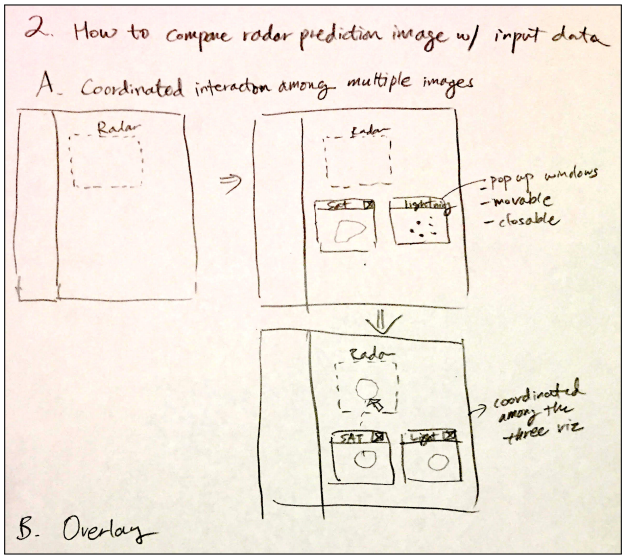

Figure 2: Early paper sketch showing linked views of the input data modalities and the radar output.

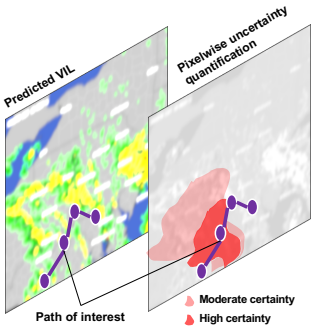

Figure 3: Pixelwise uncertainty quantification could be visually encoded with the VIL to help operators understand the likelihood of weather near paths of interest.

that help them reason about potential mission outcomes more so than assessments about the AI quality, which follows previous findings about the questions ML non-experts want to answer about system outputs [6].

Another meaningful difference between developer and operator groups is task response time and cognitive load constraints while interpreting the radar images. In an R&D setting there may be few constraints on the developers debugging their code. Aviation forecasters, on the other hand, need to interpret these images then feed information to pilots or other forecasters who roll up information into other weather products. They may spend time verifying a radar image or other modeling output by comparing it with other surface observations or models. Building this mental model of the weather system requires experience, attention to detail, and consumes cognitive bandwidth.

Interface Design Ideation and Lessons

After these initial conversations, we used affinity diagramming to organize common themes in needs, then did paper sketching to ideate on user interface layouts and interaction concepts that can address needs. Some concepts included using other data or metrics that are realistically feasible, even if they did not exist at the time of our exercise. This section describes a sample of the design concepts we selected for further exploration. We then share preliminary lessons (in bold) that could be useful in developing other XAI applications.

The most promising design concepts were well-aligned with analysis questions we expected operators would ask about the radar outputs, synthesized from our discussions (see “Target Analysis Questions" sidebar).

For Q1, we sketched out different ways that the spatiotemporal weather inputs to the ML radar model could be visualized alongside the output (Figure 2). During our conversations with weather experts, some noted they could possibly envision the weather radar image for an area if they had optical satellite or other weather observations there. In other words, they might have a pre-existing mental model of how these spatiotemporal data modalities are correlated. The user’s mental model of the data domain is complementary to their mental model of the AI system itself [10]. In fact, a user interface prototype that supports an operator in contrasting their expectation of the output (based on this mental model) with the prediction could help them gauge the trustworthiness of a prediction or the model as whole. Surfacing these differences or similarities could be thought of as a *key component* for a new explanation interface to communicate, using the framework of Eiband et al. [5], that in turn helps calibrate the user’s mental model of the system.

For Q2, helping the operators assess the radar output based on similar instances could also support trust building; for example, retrieving the top k samples of weather inputs from the ML training data that are most similar to the operational task at hand, and displaying the corresponding outputs. The operator can again leverage their mental model of how much change in the VIL prediction from these cases seems reasonable given the differences in inputs, and make a judgment about how stable the model seems to be. In both cases, we find it could be valuable to **help the operator use their mental model of the domain data to assess the ML prediction** with additional context.

For Q3, we recognized early on that quantifying model confidence or uncertainty in broad ways (one score for a large region) is less helpful to decision makers than fine-grained approaches. The forecasters we spoke with were interested in evaluating the safety of flight paths that are more precise than coarse tiles in VIL maps. Recent work toward pix-

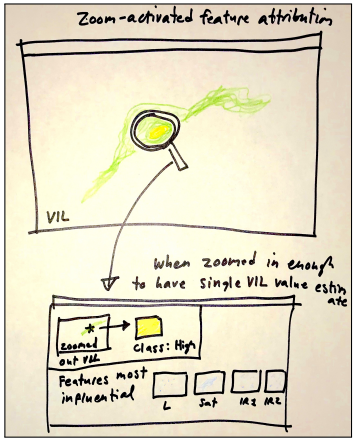

Figure 4: Zoom interactions could enable simpler predictions for XAI tools to explain (i.e., single pixel values), but it is not clear if operators discover mission-relevant insights about flight paths this way.

elwise uncertainty quantification for image regression [3, 2] could enable scalar-valued uncertainty at the pixelwise level, so we considered concepts for visually encoding both uncertainty and predicted VIL in the output colors, as well as overlays or linked views that trade-off compactness for legibility (see Figure 3).

It became apparent to us that some interface concepts we sketched were too divergent from typical operator workflows to be practical. For example, we envisioned a way to reduce the VIL synthesis problem to classification in order to leverage existing saliency techniques (see Figure 4). Class labels describing phenomena in whole weather radar patches were not available to us, but individual VIL pixels could be quantized or segmented. However, interpreting explanations at the pixelwise level may not directly inform forecasters' decision tasks. As in uncertainty quantification, we find that **new explanatory information should be communicated at the resolution of the decision task**, or mission. We also considered tools to let operators make counterfactual perturbations to spatial inputs (e.g., draw alternative lightning strikes) to evaluate what-if scenarios with the ML model. One concern that emerged was that operators could make physically unlikely or impossible perturbations when editing one input modality but not the others in corresponding ways. An operator's ability to evaluate inconsistencies between the input modalities could be strained by dramatic changes that create unfamiliar weather scenarios, even if they have a sense of these data relationships under typical conditions. User testing with visualizations of the model inputs is needed to better understand the opportunities for interactive perturbation. In general, we suggest **not imposing workflows that require extensive *new* knowledge about the ML system or underlying domain physics**, which could be error prone or invite misinterpretation.

Conclusion

This work led to three lessons we took from early human-centered design activities toward more interpretable ML outputs in a weather radar application. We focused on domain operators (aviation forecasters) as the primary stakeholders who must interpret the reliability and usefulness of synthetic visualizations. Next steps for this work include implementing the design concepts and performing empirical user studies, which could help us refine and develop new guidelines for human-centered XAI. Insights from this work could apply in other design studies where the target users are not ML experts but must gauge the reliability of complex predictions using visualization.

Acknowledgements

The authors would like to thank Larry Marine, Greg Bennett, Daniel Padgett, and others at AF CyberWorx for their contributions to this initial design study. We thank Mark Veillette and Jay Roberts for their feedback, as well as those who who graciously shared their time and insights with us. Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] 2022. SEVIR Tutorial. Website. (2022). Retrieved March 1, 2022 from `https://nbviewer.org/github/MIT-AI-Accelerator/eie-sevir/blob/master/`

`examples/SEVIR_Tutorial.ipynb`.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, and others. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297.

[3] Anastasios N Angelopoulos, Amit P Kohli, Stephen Bates, Michael I Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. 2022. Image-to-Image Regression with Distribution-Free Uncertainty Quantification and Applications in Imaging. *arXiv preprint arXiv:2202.05265* (2022).

[4] Georgy Ayzel, Maik Heistermann, and Tanja Winterrath. 2019. Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0. 1). *Geoscientific Model Development* 12, 4 (2019), 1387–1402.

[5] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design Into Practice. In *23rd International Conference on Intelligent User Interfaces*. 211–223.

[6] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[7] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[9] Michael Robinson, James Evans, and Bradley Crowe. 2002. En route weather depiction benefits of the NEXRAD vertically integrated liquid water product utilized by the corridor integrated weather system. In *10th Conference on Aviation, Range and Aerospace Meteorology, American Meteorological Society, Portland, OR*.

[10] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the Human? Bridging the Gap Between AI and HCI*. Association for Computing Machinery, Inc.

[11] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[12] Mark Veillette, Siddharth Samsi, and Chris Mattioli. 2020. SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems* 33 (2020), 22009–22019.