Aggregating XAI methods for insights into geoscience models with correlated and high-dimensional rasters

Evan Krell[1,3,5], **Hamid Kamangir**[3,5], **Waylon Collins**[4,5], **Scott A. King**[1,2,5], **Philippe Tissot**[3,5]

[1]Department of Computing Sciences, Texas A&M University - Corpus Christi, Corpus Christi, TX, USA
[2]innovation in COmputer REsearch Lab (iCORE)
[3]Conrad Blucher Institute for Surveying and Science, Texas A&M University - Corpus Christi, Corpus Christi, TX, USA
[4]National Weather Service, Corpus Christi, TX, USA
[5]NSF AI Institute for Research on Trustworthy AI in Weather, Climate and Coastal Oceanography

Key Points:

- Analyzed the influence of feature combinations on multiple XAI techniques with demonstration of how this leads to model insights
- Developed an extension to PartitionSHAP for explaining the influence of super-pixels within each raster channel toward the model's decision
- Demonstrated methods of aggregating local into global explanations that aid analysis by a domain expert weather forecaster

Corresponding author: Evan Krell, ekrell@islander.tamucc.edu

Abstract

Geoscience applications have been using sophisticated machine learning methods to model complex phenomena. These models are described as black boxes since it is unclear what relationships are learned. Models may exploit spurious associations that exist in the data. The lack of transparency may limit user's trust, causing them to avoid high performance models since they cannot verify that it has learned realistic strategies. EXplainable Artificial Intelligence (XAI) is a developing research area for investigating how models make their decisions. However, XAI methods are sensitive to feature correlations. This makes XAI challenging for high-dimensional models whose input rasters may have extensive spatial-temporal autocorrelation. Since many geospatial applications rely on complex models for target performance, a recommendation is to combine raster elements into semantically meaningful feature groups. However, it is challenging to determine how best to combine raster elements. Here, we explore the explanation sensitivity to grouping scheme. Experiments are performed on FogNet, a complex deep learning model that uses 3D Convolutional Neural Networks (CNN) for coastal fog prediction. We demonstrate that explanations can be combined with domain knowledge to generate hypotheses about the model. Meteorological analysis of the XAI output reveal FogNet's use of channels that capture relationships related to fog development, contributing to good overall model performance. However, analyses also reveal several deficiencies, including the reliance on channels and channel spatial patterns that correlate to the predominate fog type in the dataset, to make predictions of all fog types. Strategies to improve FogNet performance and trustworthiness are presented.

Plain Language Summary

Geoscience applications have been using sophisticated machine learning methods to model complex phenomena. These models are described as black boxes as it is unclear what relationships are learned. Users might not trust the models since they cannot determine how inputs influence model decisions. EXplainable Artificial Intelligence (XAI) is a growing research area investigating how models make their predictions. However, XAI methods are sensitive to how the inputs to the models interact. This makes XAI challenging to interpret models with a large number of predictors. Since many geospatial applications rely on complexity for high performance, a recommendation is grouping features so that the groups have less correlation with each other. Partitioning the input into a small number of distinct features may improve explanations, but might be too coarse for model insights. Here, we explore how grouping the inputs impacts XAI explanations. Experiments are performed on FogNet, a deep learning model for coastal fog prediction. Meteorological analyses of XAI output reveal the ability of FogNet to predict fog with skill, by accounting for physical processes related to fog development. However, there exists deficiencies in how FogNet makes predictions, which lower trustworthiness. Strategies to improve FogNet performance and trustworthiness are presented.

1 Introduction

Artificial Intelligence (AI) is increasingly used to develop high performance models that capture highly nonlinear spatial or spatial-temporal relationships. Their success often relies on complex Machine Learning (ML) architectures such as Deep Learning (DL). DL has been applied to geoscience tasks such as predicting soil temperature (Yu et al., 2021), typhoon paths (Xu et al., 2022), tropical cyclones (Lagerquist, 2020), sea surface temperature (SST) (Fei et al., 2022), traffic (Kreil et al., 2020)), and classification using multi-spectral (Helber et al., 2019) and synthetic aperture radar (Zakhvatkina et al., 2019)) imagery.

Complex ML models can be considered black boxes since their complexity obfuscates how they work. They learn a function based on associations between inputs and targets, but it is hard for humans to investigate how the data influences model output. However, we are using complex models mainly to capture nonlinear relationships. Hence, the interpretation is inherently more complex than for linear models. A global explanation of model behaviour based on, for example, the coefficients of multiple linear regression may not capture the richness of how the system behaves from case to case. Here, the challenge of model interpretability is not just due to their black box nature, but also to the challenges associated with understanding nonlinear relationships where global explanations may not provide the full understanding of how the system works. This may limit their use, since users cannot verify realistic decision-making. Researchers have demonstrated that at times models with seemingly high performance were using spurious relationships that would cause the model to fail in real-world use (Lapuschkin et al., 2019). The lack of transparency in complex ML models has motivated the rapid development of the field of eXplainable Artificial Intelligence (XAI) that includes various approaches to enhancing the ability to understand the model's decision-making strategies (Murdoch et al., 2019).

A major XAI approach is designing interpretable models that are more easily understood (Murdoch et al., 2019). However, this is typically at the cost of performance compared to more complex ML techniques. The simplest techniques, such as linear regression with a small number of features, may be trivial to explain, but without learning the nonlinear relationships needed for a particular application (Molnar et al., 2020). Here, we are interested in XAI techniques that provide insight into trained complex models to take advantage of their high performance for highly nonlinear phenomena. This class of methods is called *post-hoc* XAI techniques, and are used to generate various forms of explanations of the learned data associations.

Many post-hoc XAI techniques have been proposed. Typically, these work by evaluating the influence of each input feature towards model output. That is, how did a given feature influence the decision? Two major types of explanations are *feature importance* and *feature effect*. Feature importance methods evaluate a feature's influence by how much it impacts the model's performance. These methods are typically global; they are computed over a large number of examples to determine which features are important in general rather than for a specific case (local). A widely-used feature importance method is Permutation Feature Importance (PFI) (McGovern et al., 2019). Feature effect methods instead evaluate the contribution of each feature toward a particular model output. That is, for a given prediction, how much did the features push (or pull) the decision toward (or away from) that value.

Despite extensive research effort in developing novel XAI methods, none are guaranteed to produce an accurate explanation (McGovern et al., 2019). Techniques largely differ in how they probe the model; modifying the input in some way to assess the change in output. For example, PFI permutes a feature with other dataset values to break up the input data relationship but maintain the data distribution. If permuting this feature changes the model performance more than other features, then it is said to have higher

importance. It is challenging to select the appropriate XAI technique and rarely possible to quantitatively measure explanation accuracy since the true explanation is unknown. XAI techniques are known to be sensitive to feature correlations and interactions, despite efforts made to take these into account. A recommended approach is to run multiple methods and triangulate the results: consistencies suggest meaningful descriptions of model characteristics (McGovern et al., 2019).

Geoscience models often use high-dimensional inputs with substantial feature correlation and interaction. Consider a raster data input that represents a 3D wind field where channels are 2D spatial vector components at subsequent altitudes. Spatial autocorrelation exists within and across channels. Temporal autocorrelation could be introduced by including channels that represent the values at several time steps. A complex raster could be composed of multiple multi-channel features such as wind, turbulence kinetic energy, etc.

In this research, FogNet, a DL model for predicting coastal fog in the South Texas Coastal Bend (Kamangir et al., 2021), is used to analyze the impact of partitioning the raster elements into features for XAI. We describe some of the challenges in using XAI techniques to explain models that rely on high-dimensional spatio-temporal raster predictors. These include sensitivity to the choice of XAI method and grouping scheme, computational limitations using a high number of features, and modifying XAI software packages to support multi-channel explanations. After generating a large number of XAI outputs, we demonstrate using forecaster domain knowledge to generate hypotheses that will direct the next stages toward model improvement. This research extends preliminary XAI results we presented alongside an ablation study of the FogNet architecture (Kamangir et al., 2022).

1.1 Related Works

There are many geoscience modelling studies where complex models substantially outperformed simpler models. In the following three examples, models use DL architectures with gridded spatio-temporal predictors. In each study, comparisons with simpler alternative models highlighted high performance gains using the more complex architecture.

Yu et al. (2021) used spatio-temporal rasters to predict soil temperature. Each input raster is a sequence of 10 days of soil temperature estimates obtained from the ERA5 dataset, and each day is represented with a 20×20 spatial grid. The signal processing technique Ensemble Empirical Model Decomposition (EEMD) is used to characterize each of the 10 channels into 10 different time scales. This yields input raster predictors of size (10, 20, 20, 10). A DL architecture was developed with 3D convolution to learn spatio-temporal features. This was compared to simpler architectures (2D convolution) and lower-dimensional inputs (without EEMD), demonstrating significant performance gains using the most complex model.

Xu et al. (2022) developed a DL model for typhoon path prediction. The spatio-temporal input is created from the EAR-Interim 3D typhoon dataset. The predictors include 31×31 spatial grids at 4 isobaric planes at 4 time steps, yielding rasters of size (4, 4, 31, 31). Comparing several ML techniques and DL configurations, the best was a fusion of DL with 3D convolution and a Generalized Linear Model. Again, the more complex configurations outperformed simpler alternatives.

Fei et al. (2022) developed a hybrid model for bias correcting SST from a numerical model. The input raster has 3 time steps of 4 variables from the Hybrid Coordinate Ocean Model. With a spatial grid size of 48×48, the input data size is (3, 4, 48, 48). 3D features were learned for each time step using 3D convolution and attention blocks. Their outputs were fed into a convolutional Long Short-Term Memory (LSTM) model to learn

temporal patterns. An ablation study showed that the performance increased as additional modules were added to the DL architecture.

McGovern et al. (2019) reviewed XAI for meteorological ML which regularly use high-dimensional spatio-temporal rasters. Several techniques were used to explain a tornado prediction CNN whose input rasters are 12 channels of 32×32 spatial grids. Saliency Maps were used to highlight salient elements in each channel. Backward Optimization (BWO) was used to optimize synthetic rasters that maximize neuron activation. This is used to show what features would look like to create either optimal tornadic or non-tornadic storms, which can be used to verify that the model has a realistic understanding of these classes. Class Activation Maps (CAMs) were also used to generate heatmaps of influential elements, but do not explain individual channels: the output is a single 2D heatmap explaining a 3D raster. Many XAI techniques were made with simple RGB image models in mind, and do not separately operate on the channels. The study showed that XAI techniques can be used to gain a variety of model insights, but could not quantitatively rank explanations since the ground truth explanation is unknown. The authors warn against the potential for bias confirmation: the explanation that looks like what the modelers expected might not be the most accurate. Instead, the recommended strategy is to apply multiple XAI methods; consistencies provide evidence of the true explanation.

Gagne II et al. (2019) used a CNN to predict severe hail occurring during a storm. Given a storm-centered raster, the model predicts if the hail size will exceed 25mm. Input rasters of size (32, 32, 15) were generated from the NCAR convection-allowing NWP: 5 atmospheric variables of 32×32 gridded data, each at 3 pressure levels. PFI and BWO were used to rank important features and visualize influential variable relationships, respectively. Instead of individual raster elements, PFI ranked entire channels grouped into features such as *Geopotential height at 500 hPa*. BWO optimizes the raster such that the explanations have the same dimensionality as the input. Based on the explanations, the authors found evidence of the model learning physical relationships associated with hail.

Using storm-centered MYRORSS radar imagery and proximity soundings as predictors, Lagerquist (2020) developed a CNN to predict next-hour tornado occurrence. The input raster includes 14 feature maps, each a 128×128 spatial grid. This (128, 128, 14)-size raster goes through a number of convolution layers before combining with 4864 soundings. The combined data is passed through a dense layer to produce the tornadic probability. Several XAI methods were used to analyze how the model works. Like Gagne II et al. (2019)), PFI was used to rank the importance of channels and BWO used to show maximizing inputs. Feature effect methods Saliency Maps and CAMs were applied to generate explanation heatmaps. Since it is difficult to verify explanation accuracy, additional verification was performed to increase confidence in the explanations. Adebayo et al. (2018) observed that XAI-based heatmaps are sometimes overly influenced by discontinuities in the input raster. XAI methods could be operating more like edge detectors than model explainers. Adebayo et al. (2018) developed sanity checks that compute the likeliness that the XAI output could have been generated by a simple edge detection algorithm. Lagerquist (2020) applied these sanity checks, which suggested that the explanations are in fact based on model behavior. Saliency maps and CAMs output an explanation for each instance, making it challenging to get a global perspective of influential features. Lagerquist (2020) aggregated the explanations to examine general differences between explanations of 4 extreme cases: best hits, best correct nulls, worst misses, and worst false alarms. This was possible because the storm-centered images have a spatial consistency that allows meaningful aggregation of multiple inputs. For each class, 100 explanations were generated and combined into an explanation of that class using probability-matched means.

Hilburn et al. (2021) used Geostationary Operational Environmental Satellite (GOES) imagery to train a U-Net architecture to estimate the spatial distribution of composite reflectivity. The input is composed of 4 GOES bands, each a 256×256 image, and the output is a 256×256 spatial grid. Layer-wise Relevance Propagation (LRP) identifies influential raster elements using a backwards pass through the neural network. Influence is based on the flow of contribution from the neurons, tracing back to find which input features were most influential in activating the neurons that contributed to the prediction. LRP is computed for each output pixel. In the GLM channel of the input raster, LRP results suggest that the network focuses on lightning regions. The authors then created modified inputs, removing the lighting in the GLM channel to observe model output. The results indicated that the lightning did in fact contribute significantly. Here, an existing XAI technique was used, but with additional steps taken to increase confidence.

Beucher et al. (2022) used a CNN for probability of potential acid sulfate soil occurrence. The data is 14 gridded variables (topography, soil, and climate) over a wetlands region in Jutland, Denmark. The model outputs probabilities for a single grid cell, given that cell's 14-channel input raster of 5×5 covariate matrices. Feature effect method SHapley Additive exPlanations (SHAP) was applied with entire channels as features. Since each input is associated with a spatial location, the authors extended the SHAP analysis to the spatial region to visualize the spatial distribution of each feature's influence.

These works demonstrate XAI methods used to interpret complex models for geoscience applications that used multi-channel rasters with spatial or spatio-temporal information. Since the true explanation is unknown, none can provide a quantitative measure of explanation accuracy. Efforts to increase trust in the explanations included sanity checks (Lagerquist, 2020) and testing hand-crafted synthetic data (Hilburn et al., 2021).

Au et al. (2021) discuss XAI on grouped features based on 3 main motivations. First, it may be infeasible to generate explanations for a large dataset of high-dimensional inputs. Second, correlations and interactions may yield misleading explanations. Third, groups of related features may facilitate human interpretation of explanations. We are less interested in the third motivation. For many geoscience applications, we expect that expert users will be interested in the most granular explanations of the spatio-temporal variables influencing the model. However, the first two motivations are of major concern for high-dimensional geoscience ML. Au et al. (2021) discuss 3 feature importance methods on grouped features: PFI, refitting (retraining the model with the group removed), and LossSHAP.

1.2 Contributions

Our work provides the following contributions:

1. An analysis of the sensitivity of XAI methods to the granularity of the feature grouping scheme. That is, we investigate how the choice of grouping raster elements for aggregated XAI influences the output explanations.
2. Methods for aggregating local explanations into global model insights. By taking advantage of spatial consistency across all input cases, we combine cases into global explanations for each fog type and classification outcome.
3. A case study using aggregation of features and of explanations to investigate a 3D CNN for coastal fog prediction. Our XAI visualizations are produced based on feedback from a NWS meteorologist who then uses the explanations to generate hypotheses on the model's learned strategies.
4. Channel-wise PartitionShap (CwPS), a modification of the XAI software PartitionShap (Lundberg & Lee, 2017) to add support for hierarchical, recursive SHAP-based explanations on each raster channel. We introduced CwPS briefly in our FogNet

ablation study (Kamangir et al., 2022), but describe and apply it more extensively here.

5. Meteorological interpretations of XAI output were conducted in order to access the trustworthiness of FogNet to operational meteorologists with respect to fog prediction, as evidenced by the mechanisms and the ambient environmental conditions associated with coastal fog that are captured by FogNet, and the ability of FogNet to learn the unique processes associated with different fog types.

2 Methods

FogNet (Kamangir et al., 2021) is a DL architecture for predicting coastal fog. Models were trained to predict visibility at 3 thresholds (less than 1600m, 3200m, 6400m) and 3 lead times (6, 12, and 24-hours). A specific FogNet model instance is trained for a visibility threshold and lead time. FogNet models were trained for the South Texas Gulf Coast. Most input features were derived from the North American Mesoscale Forecast System (NAM), a deterministic numerical weather prediction (NWP) modeling system. An additional feature is observed SST from the NASA Multiscale Ultra-high Resolution (MUR) satellite dataset. The target visibility data is observations at Mustang Beach Airport in Port Aransas, Texas (KRAS). FogNet acts as AI-based Model Output Statistics (MOS) correcting the NWP output based on observations at the target region.

Each predictor is a raster of metocean variables. Each channel is a 32×32 grid with a spacing of 12 km. Thus, the domain is a 384 km^2 region along the Texas Gulf Coast, containing both coast and offshore. The target is a binary class representing whether or not there is visibility at the specific threshold.

Variables were selected to capture 3D spatial and temporal relationships related to fog. Most variables are included at multiple altitudes, forecasts, and time steps, for example, vertical velocity at *750 mb, 0 hours* and at *900 mb, 12 hours*. The number of channels used depends on the lead time. For 6 and 12-hour lead times, there are 288 channels so that the input raster size is (32, 32, 288). For 24-hour lead time predictions, an additional set of channels are included with a raster size of (32, 32, 384).

The FogNet architecture, based on 3D convolutions, was designed to capture relationships across both the spatial grids and the spatial-temporal channels as is often needed for geoscience applications (Kamangir et al., 2021). Dilated 3D convolution is used so that the model is not limited to convolution over adjacent elements. Instead, skip values are learned that define which pixels are involved in the convolution. This allows flexible learning of 3D features. Since the pattern of channels is that each variable is repeated at four time steps, it can be challenging for the model to separate spatial and temporal features. So, the FogNet architecture first separately learns the spatial and temporal features. They are then combined to allow the model to exploit any useful spatio-temporal relationships. Other mechanisms are the Dense Blocks that reduce the number of learnable parameters and mitigates the vanishing gradient problem, and Attention Mechanism that suppress the influence of less discriminative features.

FogNet outperforms the operational High-Resolution Ensemble Forecast (HREF) across several performance metrics (Kamangir et al., 2021). Here, we focus on FogNet trained for visibility < 1600m at 24-hour lead time. This longer lead time uses additional channels as compared to the shorter lead times, making it a better (that is, more challenging) case study for XAI on high-dimensional geoscience models. The trained weights used are those from the highest performing model in a set of experiments previously reported (Kamangir et al., 2021).

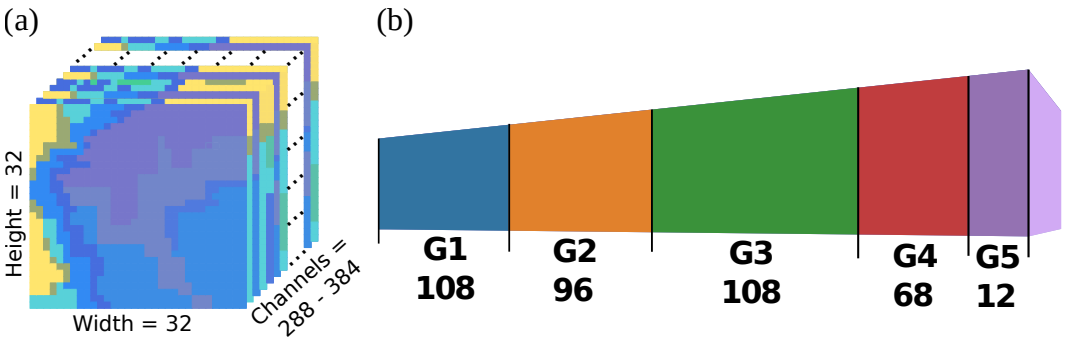


Figure 1: The FogNet input data (a) is a raster with 32 rows × 32 columns. The number of channels depends on the lead time — 288 for 6 and 12 hours, and 384 for 24 hours. The channels are divided into 5 groups based on their physical characteristics (b). `G1` is wind, `G2` is turbulence kinetic energy and humidity, `G3` is lower atmosphere thermodynamic profile, `G4` is surface atmospheric moisture and microphysics, and `G5` is sea surface temperature.

2.1 Physics-Based Channel Groups

The features originating from the NAM NWP modeling system, described in more detail in the FogNet paper (Kamangir et al., 2021), were divided into four groups (G1, G2, G3, G4). A fifth group, G5, combines the satellite sea surface temperature with outputs from the NAM NWP model. Each of the groupings combine features with a similar relationship to fog development. G1 features capture the relationship between wind and fog, and include u, v wind component profiles below 700 mb (10-meter height; isobaric levels from 975 mb to 700 mb at 25 mb increments), and the frictional velocity at the surface. Surface (10-meter) wind speed magnitudes $\geq$2.5 m/s, and surface (3-meter) frictional velocity magnitudes $\geq$0.3 m/s, can dissipate, or preclude the development of, radiation fog (Tardif & Rasmussen, 2007; Liu et al., 2011). However, wind speeds $\geq$2.5 m/s are essential for the development of advection fog (Koračin et al., 2014). Friction velocity is related to the turbulent component of the wind. In particular, friction velocity is equal to the square root of the Reynolds stress divided by air density; the Reynolds stress refers to the mean force per unit area imposed by turbulent motion on the mean flow (Glickman, 2000).

The G2 features are the turbulence kinetic energy (TKE) and specific humidity (Q) profiles below 700 mb (TKE and Q from 975 mb to 700 mb, at 25-meter increments), and captures the scenario whereby the combination of turbulence and a decrease in Q with height can dissipate or preclude fog (Toth et al., 2010). Further, radiation fog generally requires an increase in Q with height in the lower levels (Petterssen, 1940; Baker et al., 2002). In addition, TKE at 975 mb may capture near surface mechanical turbulence that contributes to advection fog formation (Huang et al., 2011).

The G3 features approximate the thermodynamic profile below 700 mb by including relative humidity (RH) and temperatures (TMP) at the 2-meter height, and at isobaric levels from 975 mb to 700 mb at 25 mb increments. Radiation fog occurrence is correlated with a thermodynamic profile characterized by a thin moist/saturated layer near the surface, followed by much drier air aloft. Yet, advection and stratus-lowering fogs are associated with deeper moist layers (Croft et al., 1997; Dupont et al., 2016). Both radiaton and advection fogs are associated with a near surface temperature inversion (increase in TMP with height) (Koračin et al., 2014; Mohan et al., 2020).

The features of G4 account for surface moisture and level of air saturation (2-meter q and dew point depression, respectively), and microphysical processes responsible for fog development. Microphysics features include the NAM surface visibility, which measures the empirical relationship between visibility reduction (owing to fog) and cloud liquid water (from the NAM microphysics parameterization scheme), and temperature at the lifted condensation level (TLCL) (inversely related to the activation of cloud condensation nuclei or CCN). Vertical velocity (VVEL) below 700 mb was included in G4 given its relationship to CCN activation (Gultepe et al., 2017). However, VVEL also relates to G1 wind features and to larger scale environmental conditions that relate to fog. For example, radiation fog occurs when VVEL magnitudes are weak (Gultepe et al., 2017). Further, advection fog tends to occur within an environment characterized by synoptic scale subsidence (negative VVEL values) below 500-mb (Huang et al., 2011; Yang et al., 2017; Mohan et al., 2020). In addition, radiation fog tends to occur during weak local subsidence below 220 meters (Liu et al., 2011; Dupont et al., 2016), and also above 220 meters (Mohan et al., 2020).

The G5 features describe sea surface temperature (SST), difference between SST and temperature (TMP-SST), and the difference between dew point temperature and SST (DPT-SST). They capture the conditions consistent with the development of marine advection fog that occurs at the KRAS target primarily during the Winter months. Further, T-SST modulates radiation fog. When T-SST$<$0, near surface upward directed sensible heat flux can counteract radiational cooling and thus either delay the onset of, or prevent, radiation fog (Liu et al., 2011). Marine advection fog tends to occur within a specific range of SST values (P. Li et al., 2016), and occurs when DPT-SST$\geq$0 or TMP-SST$\geq$0 (Koračin et al., 2014).

Although the foregoing groups contain features that relate to a particular fog generating or dissipating mechanism, or possess a statistical correlation to fog occurrence, there exists correlations across the groups. For example, Groups 1, 2, and 3 are correlated; a temperature inversion (temperature increase with height) in the lower levels (G3), which typically occurs during fog events, will result in an atmospheric condition known as positive static stability (Wallace & Hobbs, 1977) which will suppress vertical mixing of air which in turn affects surface wind velocity (G1). Furthermore, a thermal inversion can suppress turbulence (G2) (Stull, 1988). Groups 3 and 5 are related since G3 contains surface RH, which is inversely proportional to the G5 feature dew point depression. Since wind has a turbulence component, there exists a relationship between G1 wind and G2 TKE. In addition, G1 and G4 are related since surface wind divergence (convergence) results in downward (upward) VVEL immediately aloft.

2.2 Explainable Artificial Intelligence

In this research, FogNet is used as a case study to investigate using XAI on features grouped by multiple levels of granularity to aid interpretation. We apply the 3 feature importance methods discussed by Au et al. (2021), as well as the feature effect method PartitionSHAP. The goal is to analyze FogNet with relatively granular feature groups. Since larger groups of correlated features are expected to produce more accurate explanations, we will use consistency among grouping schemes to guide confidence in the XAI outputs. That is, when the more granular outputs align with coarser outputs, we have more trust that we can use the detailed output for model insights. When they disagree, we assume that the feature correlations and interactions impede accurate explanations at the finer-grained level. Section 2.1 discusses geometric feature grouping schemes. Sections 2.3 and 2.4 briefly describe the feature effect and importance methods used, respectively.

2.2.1 XAI with Grouped Features

The features for the XAI algorithms are not necessarily the atomic components of the input data, but are defined by the grouping scheme used. Thus, a feature could be the entire `G1` from the FogNet raster, or a single channel in `G1` or even an 8×8 superpixel inside that channel.

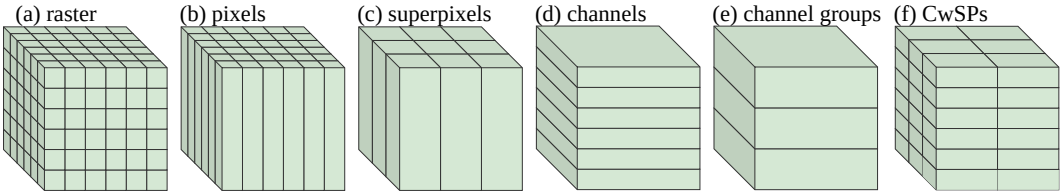


Figure 2: Various geometric partition schemes for grouping 3D raster elements. The most granular is the (a) raster itself where no grouping is applied: the features are individual (row, column, channel) elements. Each spatial element can be grouped into a (b) pixel that contains all channels at that (row, column) location. Adjacent pixels may be combined into coarser (c) superpixels. Similarly, adjacent (d) channels may be aggregated into (e) channel groups. Within each channel, the elements may be aggregated into (f) channel-wise superpixels (CwSPs).

Figure 2 shows several possible geometric partition schemes. This is far from exhaustive, and the schemes can be arbitrarily complex. For example, a superpixel could include a subset of the channels (e.g. two adjacent channels) instead of either all (*superpixels*) or one (*channel-wise superpixels*). The groups do not have to be uniform in size. For example, we could choose to group each channel individually, except to explicitly group channels of wind horizontal and vertical components together.

The most granular is to apply no grouping, using each raster element as a feature. While fine-grained explanations are ideal, there are practical limitations from the computational requirements and the sensitivity to feature correlations and interactions.

XAI software packages are often developed with RGB images in mind, treating pixels as features to generate 2D explanation heatmaps. For geoscience applications, including FogNet, the pixels often represent discrete spatial locations across a number of variable channels. So a pixel-level grouping of FogNet would only highlight influential regions, but not which of the 384 variables were influential at that location.

Pixels may be grouped into superpixels for faster computation on a smaller number of features. Groups of adjacent elements may contain sufficient information to trigger a change in model output even if single pixels couldn't with so little information individually. A drawback is that evenly dividing the space creates arbitrary superpixels with possible discontinuities, i.e. containing both land and water along the shoreline.

Entire channels can be treated as features to reveal influential variables, and is very practical for geoscience applications where channels are each a distinct spatial variable map. However, the explanations do not reveal where on the earth surface the variable is most influential. Both Gagne II et al. (2019) and Lagerquist (2020) applied PFI to channels. Channel groups are a collection of adjacent channels. When adjacency is meaningful (i.e. altitude/time), then the channel groups may capture autocorrelation to improve XAI accuracy. For example, the vertical structure/profile of the FogNet feature/channel air temperature (TMP) is more important to operational meteorologists when forecasting fog and thunderstorms, than the TMP at a particular level. Such adjacent channels can be aggregated into a single XAI feature.

Channel-wise superpixels are simply superpixels but within each channel. These features allows for explanations of *what variable* and *where*, i.e. u-component of the wind, approximately 40-km east of the target KRAS, at the 700-mb pressure level.

Here, FogNet features are grouped at 3 schemes. At the least granular, based on the 5 physics-based channel groups already defined (Section 2.1). We expect the groups to be semantically meaningful and to produce the most accurate explanations. Our ablation study (Kamangir et al., 2022) confirmed that each group contributes to FogNet's high performance. So we expect that each group should be assigned significant importance. The drawback is that we learn relatively little. If we discover that `G1` (wind) is a high-ranking feature, we still don't know which of the 108 wind-related channels are influential and at what geographic locations.

Next, each of the 384 channels are used as features. Ideally, this reveals more insight into the model's decision making: which variables, at what altitudes and time steps. However, there is already a risk of highly-correlated channels diluting the detection of each channel's true influence. To assess sensitivity, we sum the channel-wise XAI values to see if the summations achieve the same group ranking as the channel groups XAI.

To achieve spatio-temporal explanations, the lowest level of granularity is channel-wise superpixels. We will assess sensitivity by aggregating the superpixels in each channel for comparison with channel-wise XAI, and further aggregated into groups to compare to group-based XAI.

2.3 Feature Effect Methods

Feature effect methods are intended to quantify the extent that a given feature influences a specific model output, that is, each feature's contribution to the prediction. Unlike feature importance methods, feature effect reveals features being used by the model even when they have very little impact on the overall performance. Or, when the feature is used frequently for both correct and incorrect outputs. The positive and negative impact on performance may cancel out such that a very influential feature is not detected by a feature importance method.

Feature effect methods are called local methods because the explanation is for a specific model output. This is in contrast with global methods that explain how the features are used by the model across a set of instances. Thus, feature effect methods present more detailed information. It is possible to find incorrect learned strategies that occur only occasionally. A global method can average out the information across a set of examples, so these rarer feature contributions might not be detected. But they may be the the most interesting if they are most influential during extreme events like storms where the model's decisions are most critical.

However, it is challenging to obtain global model insights from a large set of high-dimensional local explanations. Each explanation is a raster of values with the same dimensions as the number of features. Thus, it is common to aggregate the local explanations into a smaller number of global explanations. Instead of a single global explanation, they can be aggregated by category. For example, Lagerquist (2020) combined XAI results by extreme cases: best hits, worst misses, etc. This aids interpretation of the set of explanations by the differences: does the model rely on different features for fog vs non-fog cases, and do these differences match forecaster knowledge?

In our case, the FogNet input rasters have consistent grid cell geography. That is, each has identical geographic extent and resolution such that (row, col) coordinates always align spatially across samples. We take advantage of this to aggregate explanations by summing values at each coordinate across a set of samples as shown in Section 2.3. At the risk of losing fine-grained information from case-by-case model behavior, this con-

verts the set of local explanations into a much smaller set of figures that can be analyzed more easily.

2.3.1 SHapley Additive exPlanations

Game-theoretic Shapley values are the fairly distributed credit to players in a cooperative game Lundberg & Lee (2017). That is, each player should be payed by how much they contributed to the outcome. In the XAI setting, the features can be considered players in a game to generate the model output. Thus, a feature that influenced the model to a greater extent is a player that should receive more payout for their contribution.

Shapley values are a feature's average marginal contribution to the output. Calculating Shapley values directly has combinatorial complexity with the number of features. Since it is infeasible for high-dimensional data, Lundberg & Lee (2017) developed a sampling-based approximation called SHapley Additive exPlanations (SHAP). Molnar et al. (2020) gives a detailed explanation of Shapley values and SHAP, including a discussion of advantages and disadvantages.

A single contribution is the difference between the model output with and without a feature x. However, models usually expect a fixed input, and do not support leaving out a feature. Feature removal has to be simulated somehow, and a variety of methods have been proposed. In canonical SHAP, the value of the removed feature x is replaced with the value from x in other dataset examples. By replacing x with many such values and averaging the result, SHAP evaluates the average difference in output between the true value of x and output without that value.

The key to Shapley values (and SHAP) is that many additional output comparisons are performed to take into account feature dependencies and interactions. In the context of a cooperative game, consider a team that has 2 high-performing players x and y. The remaining players on the team have no skill. With x and y playing, the team wins despite no help from the others. The goal is to fairly assign payout to the players based on their contribution to the game's outcome. Suppose x is removed from play and y is still able to win the game. Comparing the two games, one could conclude that x did not contribute to the win. Instead, if y were removed and x wins the game then it appears that y does not contribute. However, removing both x and y causes the team to lose the game. Thus, the change in game outcome from player x depends on player y.

The combinatorial complexity of Shapley values is because it takes the above dependency issue into account. To evaluate the contribution of x, it does more than just compare model outputs with and without x present. It repeats the comparison, but considering all possible combinations of other players being present or absent from the game. A feature's Shapley value is a weighted average of the contribution over all the possible combinations of players. SHAP approximates the Shapley values over a set of samples for performance, but still potentially requires a very large number of evaluations to converge to a close approximation. Thus, computing Shapley or SHAP values may be infeasible for more granular feature grouping schemes.

2.3.2 PartitionSHAP

In the case of FogNet, it is impractical to use SHAP for channels or channel-wise superpixels because of the large number of features. Lundberg & Lee (2017) provide an alternative called PartitionSHAP that uses a hierarchical grouping to recursively approximate Shapley values for superpixels with a significantly reduced number of calculations. The computational complexity of PartitionSHAP is quadratic with the number of raster elements instead of SHAP's exponential complexity.

Given a partition tree that defines a hierarchy of feature groups, PartitionSHAP recursively traverses the tree to calculate Owen values. Owen values are equivalent to Shapley values for a linear model, but otherwise have their own game-theoretic properties that are useful for dealing with correlated features. Unlike Shapley values, the recursive Owen values are able to correctly assign to feature groups credit even if the correlated features are broken while perturbing those features. However, this is only true if the partition tree groups correlated features.

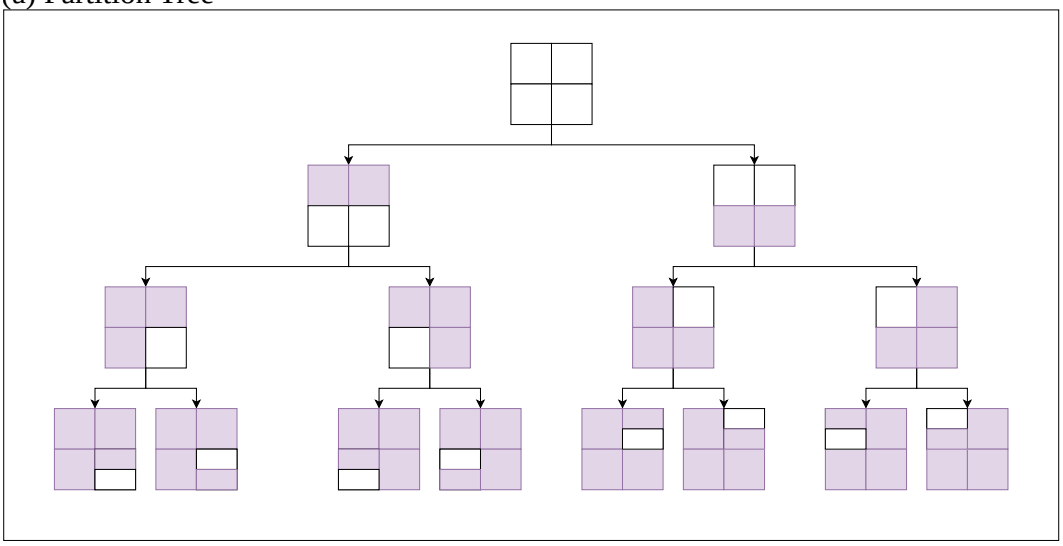
For tabular data, PartitionSHAP uses a clustering algorithm to define the partition tree in a data-driven fashion. For rasters, PartitionSHAP partitions by recursively dividing the data into 2 equal-size superpixels. Figure 3a shows the first 4 levels of a partition tree constructed for a single-channel raster. The root is the largest group, the entire image. Each node's children represent splitting it into 2 superpixels. PartitionSHAP's image partitioning algorithm is illustrated in Figure 4.

Figure 3b illustrates how the Owen values are calculated based on the recursively defined feature hierarchy. First, consider calculating the Owen value of the root node's left child. This is the superpixel representing the bottom half raster elements. The Owen value is the weighted sum of multiple model evaluations that represent the change in output with and without the superpixel present. The left-hand operation is the difference in model output with no information (all values removed) and with the superpixel's values added. The right-hand operation is the difference between the model output with all values present and with the superpixel's values removed. Together, these describe the contribution of the superpixel. Below is the calculation for the bottom-right quadrant superpixel. This example more clearly shows how the hierarchy reduces the number of required computations compared to SHAP. There are four comparisons. First, the difference in output when only the superpixel is present. Then, the output when the group is present but the superpixel is removed. Next, the group is absent, except for the superpixel (and the parent's sibling is present). Finally, the group is present (sibling absent), and the superpixel removed. All four evaluations are with respect to the superpixel being evaluated and its parent group. With SHAP, evaluating this bottom-right superpixel would have required evaluating the model with all other quadrants being present or absent. Here, there is no evaluation of the top-left and top-right quadrants since they are not part of the bottom-right's feature hierarchy. Since the image-based partitioning is performed by arbitrarily splitting the raster elements by the image size, there is no guarantee that the partition hierarchy captures correlated feature groups. Thus, Owen value's game-theoretic guarantees are violated. Regardless, Hamilton et al. (2021) applied PartitionSHAP and described the explanations as high quality and outperforming several other XAI methods including Integrated Gradients and LIME. Even without partitioning the raster into optimally correlated clusters, the superpixels contain spatially-correlated elements and might cause an appreciable change in the model output compared to a single raster element.

Our main motivation to use PartitionSHAP is efficiency. Shapley-based channel-wise superpixel explanations are feasible because of 2 properties. First, the recursive scheme that lowers the number of required evaluations already described. Second, PartitionSHAP selectively explores the tree to calculate more granular superpixel values based on the magnitude of the Owen values: a superpixel with higher Owen values is prioritized such that its children superpixels will be evaluated before those with lower magnitude values. Given a maximum number of evaluations, PartitionSHAP generates explanations with more influential raster elements at increased granularity.

PartitionSHAP divides by rows and columns, and only by channels when at a single (row, col) pixel. Here, we are interested in superpixels inside each of the channels. These represent windows of spatial regions within a single feature map. We added an additional partition scheme option to Lundberg & Lee (2017)'s SHAP software. This par-

(a) Partition Tree



(b) Owen value calculation

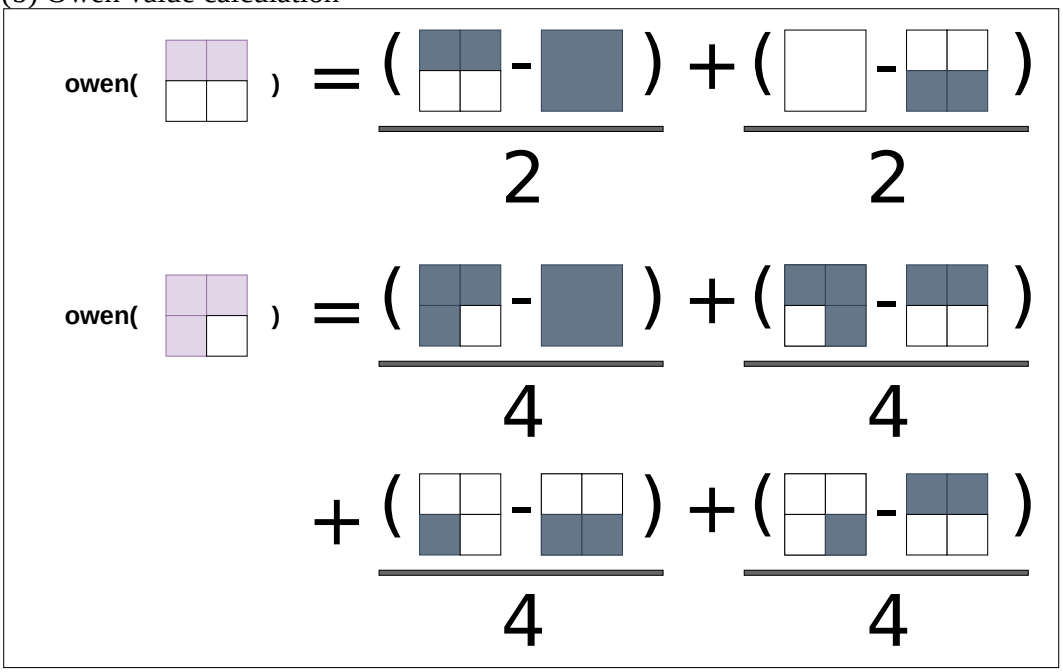


Figure 3: The hierarchy is defined by the partition tree that is generated by recursively splitting the raster. An example partition tree for a single channel, shown to a depth of 4, is given in (a). The white elements indicate the superpixel at that node. The tree continues until the leaf nodes are single (row, col) elements. Owen values (b) are calculated recursively, where each superpixel is evaluated based on comparisons with the elements in its larger group either present or absent. The number of comparisons required to calculate Owen values doubles at each level, starting with 1 at the root. The computational speedup compared to conventional SHAP is that the evaluations are limited to within a branch of the tree. That is, to evaluate the Owen value of the lower-right quadrant, it does not evaluate the situations where only the upper-left quadrant is present or absent.

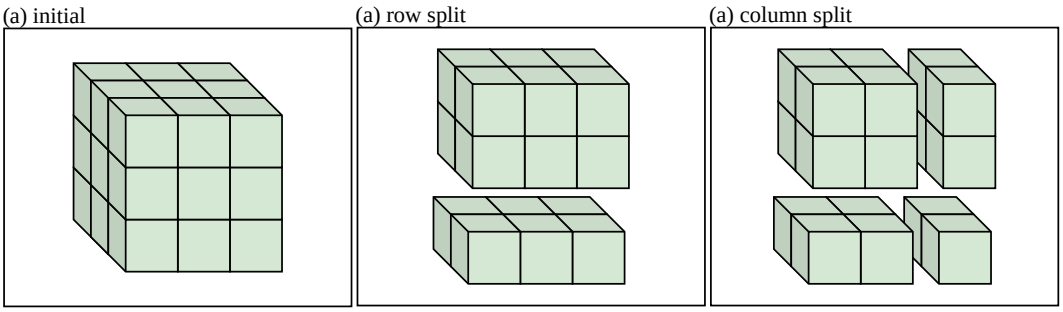


Figure 4: PartitionSHAP's default scheme for dividing raster elements into a partition tree. Given an input raster (a), the rows and columns are alternatively halved. (b) demonstrates a row split that divides vertically into two groups. This is followed by a column split (c) further dividing each horizontally. This process continues, recursively building tree where each group is a node whose children are the two groups formed by splitting it.

tition scheme, illustrated in Figure 5, splits along the channels first, then into superpixels within each channel.

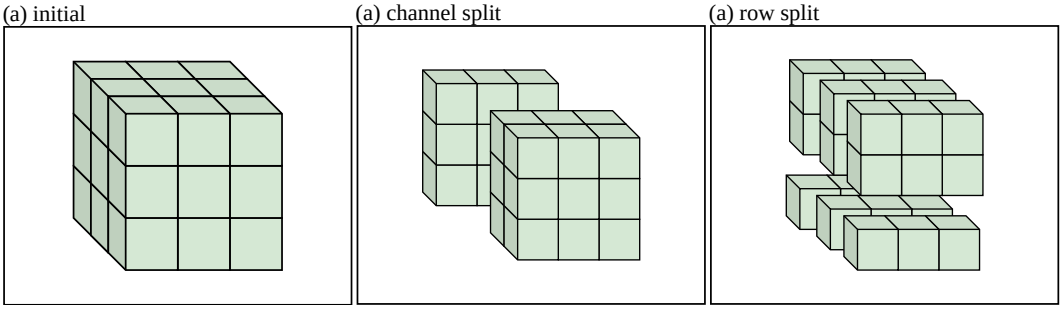


Figure 5: An alternative partition scheme used to define feature groups as Channel-wise Superpixels (CwSP). The input raster (a) is initially divided along the channels. (b) shows the result of a single channel split, dividing the raster into two halves. When the partitioning reaches a single-channel group, it begins recursively diving along the rows and channels as before. (c) shows the result of row splits performed on all three channels. We call this scheme Channel-wise PartitionSHAP (CwPS).

2.4 Feature Importance Methods

Feature importance methods are intended to quantify the extent that a given feature influences model performance. These are called global XAI methods since they describe the feature's influence based on a set of instances. With a large and representative set of examples, perhaps the entire test dataset, the explanation is expected to reflect overall model characteristics. Here we report the change in the Peirce Skill Score (PSS) to determine a features importance, however other performance metrics were used, including the Heidke Skill Score (HSS) and the Clayton Skill Score (CSS) for separate FogNet feature importance experiments. Since fog events are rare compared to non-fog, it is trivial to achieve high accuracy yet with limited forecasting skill. However, the HSS,

PSS, and CSS performance metrics measure skill. Furthermore, PSS and CSS measure economic value. Thus, it is non-trivial to achieve high skill when using these 3 metrics.

Feature importance methods differ mainly in how feature removal is simulated. Again, the fixed-length model input prevents actually removing the feature. A trivial example is to replace the feature's value with a random value, breaking the pattern originally present. A potential issue is that random values could create unrealistic input samples well outside the domain of the training data. The model's output may reflect the use of unrealistic data rather than properly simulating the removal of that feature (Molnar et al., 2020). Alternatively, the replacement value could be randomly selected from that feature's value in other dataset samples to ensure realistic values. However, the combination with other features could still be unrealistic, again risking model evaluation with out-of-sample inputs. Molnar et al. (2020) describes the problem of feature replacement in detail. Given the lack of a completely satisfying solution, it is (again) recommended to run multiple techniques to find consistent explanations.

2.4.1 Refitting Methods

Instead of simulating feature removal, an alternative is to retrain the model with a resized input that does not include that feature (Au et al., 2021). Feature importance is the difference in model performance trained with and without that feature. The new model has to learn data associations without that feature available, so it may be a very revealing assessment of that feature's influence.

Training the model from scratch for each feature requires substantial computing resources. Refitting methods are infeasible for explaining models with high-dimensional inputs. Requiring >2 hours to train, it would take >786432 hours to explain each element of FogNet's (32, 32, 384)-size raster. Because of stochasticity in training, each should really be done multiple times to compute an average. However, refitting may be applied to coarser groups such as the 5 physics-based channel groups. Here, we refer to the refitting method as `Group-Hold-Out` (GHO) and use is to explain the 5 channel groups.

While refitting methods avoid the problem of out-of-sample feature replacement, they do not entirely mitigate feature correlation concerns discussed in Section 2.3.1. If features x and y provide strong discriminative information, but are highly correlated with another, then retraining with only x or y removed might have negligible impact on model performance. One could imagine retraining the model with each group of features removed, like SHAP, but with combinatorial model retraining.

Another issue is that the explanation is technically not for the model originally to be explained, since each refitting generates a new model. If the model behavior is narrowly constrained, then the explanation should be valid. But if each model is learning unique strategies (i.e. many equally valid data associations can predict the target), then it may be misleading to rely on this as an explanation of the specific model.

2.4.2 Permutation Feature Importance

PFI simulates feature removal with permutation to replace the values of the feature of interest (Breiman, 2001). This is done over a set of samples to produce a global explanation. The following is a brief summary of the PFI algorithm used to calculate the importance of a single feature $x_i \in X$ where X is the set of all features. This is repeated for each feature. Without feature grouping, every (row, col, channel) is a distinct feature.

First, for every sample in a set of samples, permute the value of feature x_i and compute the model output with the modified input. This yields a set of model outputs. Then, compute the model performance using a chosen metric (e.g. the loss function). Next, cal-

culate the difference between the model's original (base) performance and that of the modified input data. The mean difference is the importance score. If the model performance drops significantly, then x_i is considered an important feature. If there is minimal performance change, then x_i is either unimportant or has information that is redundant with other features (McGovern et al., 2019). Alternatively, the performance could actually increase which indicates that the feature was in fact hurting performance (negative importance).

There are 2 main ways to perform the permutation with grouped features. One is a joint permutation where all the values of the feature being permuted move together into another instance. The goal is to maintain a valid relationship to avoid replacing the feature with out-of-sample values. Another approach is a completely random permutation where each value could be any other value from the permuted features. Since many of the FogNet data samples are similar to each other, we chose the latter to avoid replacing a feature with very similar values to what it originally had.

2.4.3 LossSHAP

LossSHAP is a SHAP variant for global feature importance (Covert et al., 2020). Au et al. (2021) provide a complete description of using Shapley-based XAI algorithms for grouped feature importance. LossSHAP can be described as a hybrid of SHAP and PFI. Instead of calculating the average marginal contribution (change in output) like SHAP, LossSHAP calculates the average marginal importance (change in performance).

Like PFI, the importance is based on the difference in performance with and without the feature across a set of samples. Like SHAP, the importance is the weighted average of this performance difference, considering all possible combinations of other features being present or absent.

3 Results

The XAI methods described were applied to explain FogNet. Feature importance methods were applied to the entire test dataset of 2229 cases. PFI was applied to three feature grouping schemes: channel groups, channels, and channel-wise super pixels (CwSPs). Because of their substantial computational requirements, LossSHAP and GHO were applied only to the 5 channel groups.

Feature effect methods were applied to 293 cases taken from both the test and validation datasets. This includes all 67 hits, 64 misses, and 78 false alarms, as well as 84 randomly selected correct rejections. The hits and misses are further broken down by fog type. Here, we are most interested in *advection fog* (A) and the combined category *radiation and advection-radiation fog* (R/A-R). To clarify, unless the phrase *radiation and advection radiation fog* is used, whenever *radiation fog* appears alone in this paper, it refers to both radiation and advection-radiation fogs. Both radiation and advection-radiation fogs are grouped together because the same mechanism is responsible for the formation of both: radiational cooling. The mechanism governing *advection fog* is fundamentally different. Advection fog is the majority fog case in the data set, where FogNet was shown to perform well. Radiation and advecton-radiation fogs are highly underrepresented in the data. For advection fog, there are 50 hits and 34 misses. For radiation and advection-radiation fog, there is only 1 hit and 10 misses. The environmental conditions that contribute to radiation fog are more difficult to predict via numerical weather prediction (NWP) models than those governing advection fog (Stull, 1988; Gultepe et al., 2007). Thus, the rarity of radiation and advection-radiation fog, combined with the difficulty of forecasting this fog type, represents a major challenge. The other fog types include precipitation, frontal and cloud base lowering (CBL) fog. Frontal fog can be divided into frontal passage, warm front pre-frontal, and cold front post-frontal fog types.

Table 1: Summary of XAI methods applied to explain FogNet[a]

Explanation Type	Technique	Channel Groups	Channels	Super Pixels
Feature importance	PFI	✓	✓	✓
	GHO	✓	×	×
	LS	✓	×	×
Feature effect	SHAP	✓	×	×
	CwPS	×	×	✓

[a]To analyze the sensitivity to the choice of grouping scheme, methods are applied to Channel groups, channels, and Channel-wise Superpixels. Because of computational limitations, not all methods are applied to all schemes.

Table 1 summarizes which XAI methods were applied to each grouping scheme and for each type of explanation.

3.1 Feature Effect

The two methods used to study feature effect were SHAP and Channel-wise Partition Shap (CwPS) and they were applied to the entire dataset. SHAP could not be applied to the 384 channels directly due to its complexity, however we did apply it to the 5 channel groups. Here, SHAP values are determined by the contribution to each group towards FogNet's output fog predictions, and when a group is not used, the values for all features within the group are set to zero. The result is a SHAP value for every case in the test dataset. Figure 6a shows the distribution of the group SHAP values for all 2228 test cases. The most common case (96%) is no fog and although we use a threshold value of .8, the mean value is .048. Thus, the SHAP values overall tend to be quite small for most cases in the test dataset, and Figure 6a shows that the groups all have a very similar impact, since their SHAP value distributions are similar. However, Figure 6b shows the distributions of the SHAP values broken out by outcome (hit: 37, miss: 30, correct-reject: 2126, false-alarm: 35). Here we see a different story. Group four plays a bigger role in moving the decision of FogNet towards one of fog. The other four groups also contribute to a decision of fog, but their distributions are very close to each other showing a a somewhat similar contribution for when the model predicts fog.

CwPS was used to determine SHAP values for super pixels. CwPS creates a local explanation and needs to be performed on each case individually. Since it is quite slow, we use only 293 FogNet cases (a sample of the correct rejects, plus all the hits, false alarms, and misses) from the validation and test datasets to get use the local explanations to get a sense of a global explanation. The validation data contains fog and non-fog cases from 2009-2012, and the test data from 2018-2020. While Molnar (2022) generally recommends performing XAI on the test data, we combined it with validation because of the highly imbalanced dataset having very few fog cases.

The hits and misses are further broken down by fog type. Here, we focus on 2 categories: (1) advection fog and (2) combined radiation & advection-radiation fog. The latter are grouped because they are driven by the same process: radiational cooling. Forecasters rely on distinctly different conditions to predict these fog categories, so it is of interest to compare their XAI outputs. Does the model learn different strategies to predict broad fog types? Based on our previous FogNet analysis (Kamangir et al., 2021), we have found that FogNet performs poorly for radiation & advection-radiation fog. We

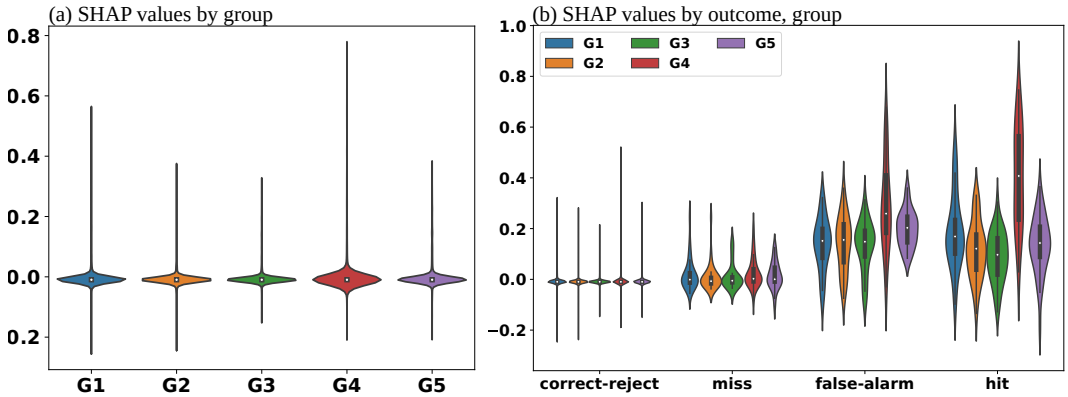


Figure 6: SHAP feature effect results for the 5 groups. SHAP values for each group are calculated for each of the 2228 cases and the violin plots represent the distribution of the SHAP value for those cases. (a) lists all 2228 cases, while (b) aggregates based on the outcome.

hypothesize that FogNet is mainly learning to predict the dominant fog type, advection fog. But, without XAI, we do not know if FogNet is simply applying advection fog strategies to all fog types, or if it is learning different strategies ineffectively.

CwPS yields a high-volume output: 293 explanations, each a (32, 32, 384)-size SHAP value raster. It is challenging to manually inspect these for a general understanding of model behavior. Here, we are interested in broadly characterizing the model's strategies for the outcome categories. That is, we chose CwPS because it calculates feature effect values not because it produces local explanations. So, we aggregate local explanations for each outcome category. We used three aggregation schemes to analyze the results in terms of spatial-channel, spatial, and channel SHAP values.

The spatial-channel aggregations are the summation of the CwPS outputs within each outcome category: 8 aggregate explanations of size (32, 32, 384). While there is some risk of positive and negative SHAP values cancelling out, this highlights the dominant sign of the SHAP values. This enables seeing which CwSPs are consistently influential toward or away from the category's prediction. A cursory manual inspection showed that the relatively high-magnitude SHAP values are confined to a small number of channels. So, we ranked the channels by their maximum absolute superpixel SHAP value to focus on the more influential subset. The ranked spatial-channel-wise aggregation results are shown in Figure 7. This figure only shows the top ranked channels that correspond to timestep t_3 which are the 24-hour lead time NAM outputs. The decision to highlight t_3 channels is because they support a meteorological analysis of the XAI results. Specifically, to examine if the t_3 features that are detected as being important based on XAI techniques correspond to a forecaster's knowledge of fog conditions which here are predicted for a 24-hour lead time. The meteorological interpretation of these figures is included in Section 4.2. However, all XAI outputs are available online (see Section 6).

To highlight influential spatial regions, the SHAP values of each channel were summed at each (row, col) location. This procedure converted the 8 spatial-channel aggregates into the 8 (32, 32)-size rasters shown in Figure 8. An interpretation of this figure from a meteorological perspective is provided in Section 4.2.

Finally, CwPS outputs were aggregated into 384 channel explanations for each of the 8 categories to highlight influential spatial-temporal metocean variables, i.e. *Vertical velocity at 950mb,* t_1. The straightforward approach is to simply sum or average the

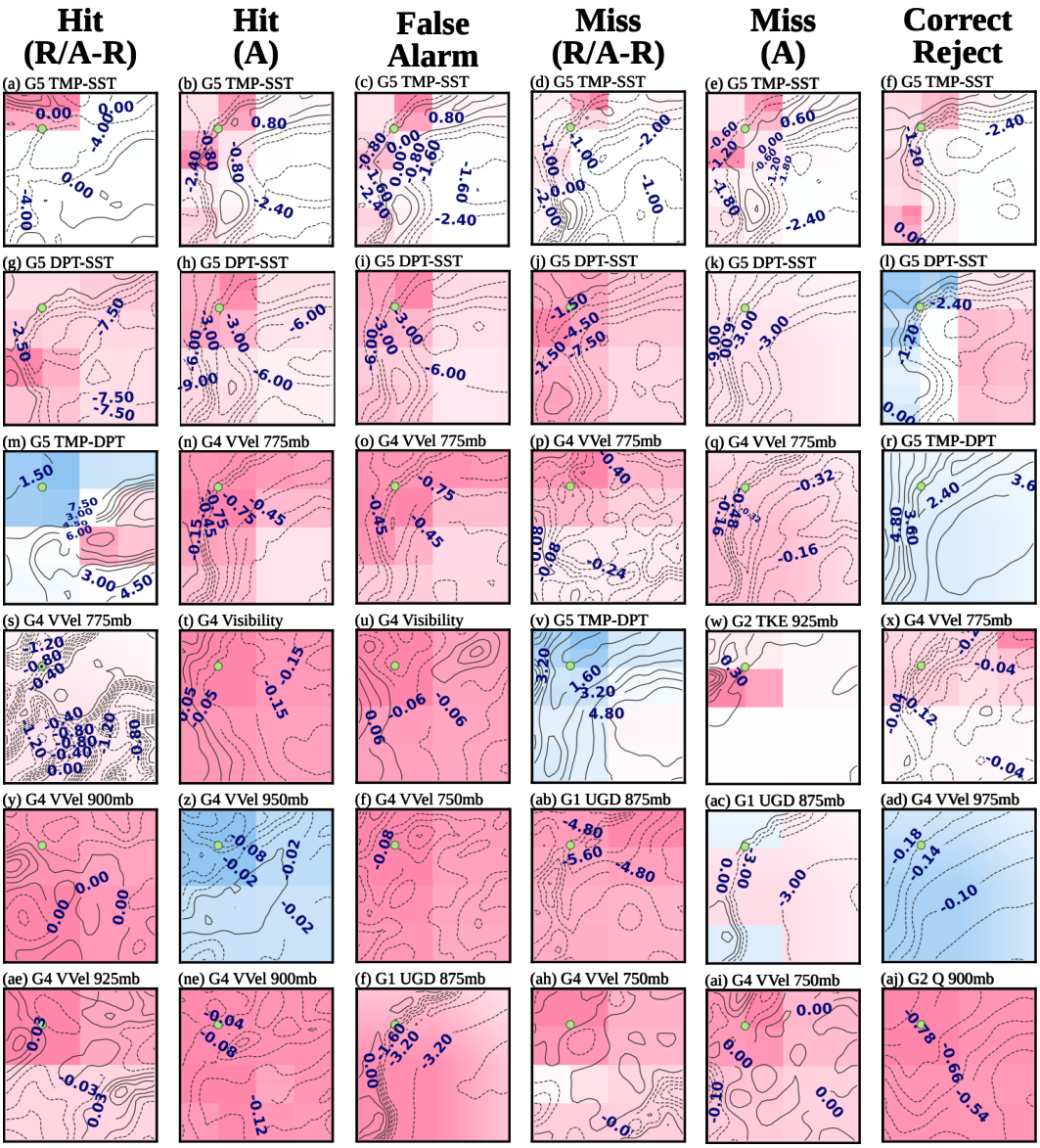


Figure 7: CwPS spatial aggregates for T_3. Shown are the top ranked channels (top to bottom) based on the absolute maximum SHAP value in the superpixels, organized by FogNet prediction outcome. Red means the feature influenced the model toward the outcome while blue indicates away from. R/A-R (A) refer to radiation and advection-radiation (advection) fog cases

superpixels within each channel. By performing XAI at the superpixel level, we expect that some influential features will not be detected as such because of the correlation issues discussed. Thus, we want to draw out even low-magnitude values to compare the relative channel influence to channel-wise feature important results (Figures 10b, 10c). This motivated a counting-based aggregation, instead of summation. First, the channels were ordered by the maximum absolute value of their superpixels. Then, for each channel we counted the number of times that it appeared in the top N channels. Intuitively, if a channel frequently occurs in the top N then it suggests that the channel is overall influential. We also counted the number of occurrences of each channel in the *bottom* N

channels, which would suggest relatively less influence. The results of counting the occurrences of each channel in the top and bottom 50 channels are given in Figure 9. Note that Figure 9a shows that all G4 and G5 features are amongst the most influential with respect to radiation and advection-radiation fog for cases where FogNet successfully predicts fog or mist with 1600 meter or less visibility. The G4 features include the NAM visibility and vertical velocities 700 mb and below. Negative vertical velocities tend to occur below the 220 meter height level during radiation and advection-radiation fog (Dupont et al., 2016; Liu et al., 2011). The G5 features include TMP-DPT, which must be less than 2 degrees Celsius to facilitate saturation necessary for radiation fog, and TMP-SST which modulates fog development, as mentioned earlier; if TMP-SST is negative, an upward-directed sensible heat flux will counteract radiational cooling and either delay fog onset, or prevent fog.

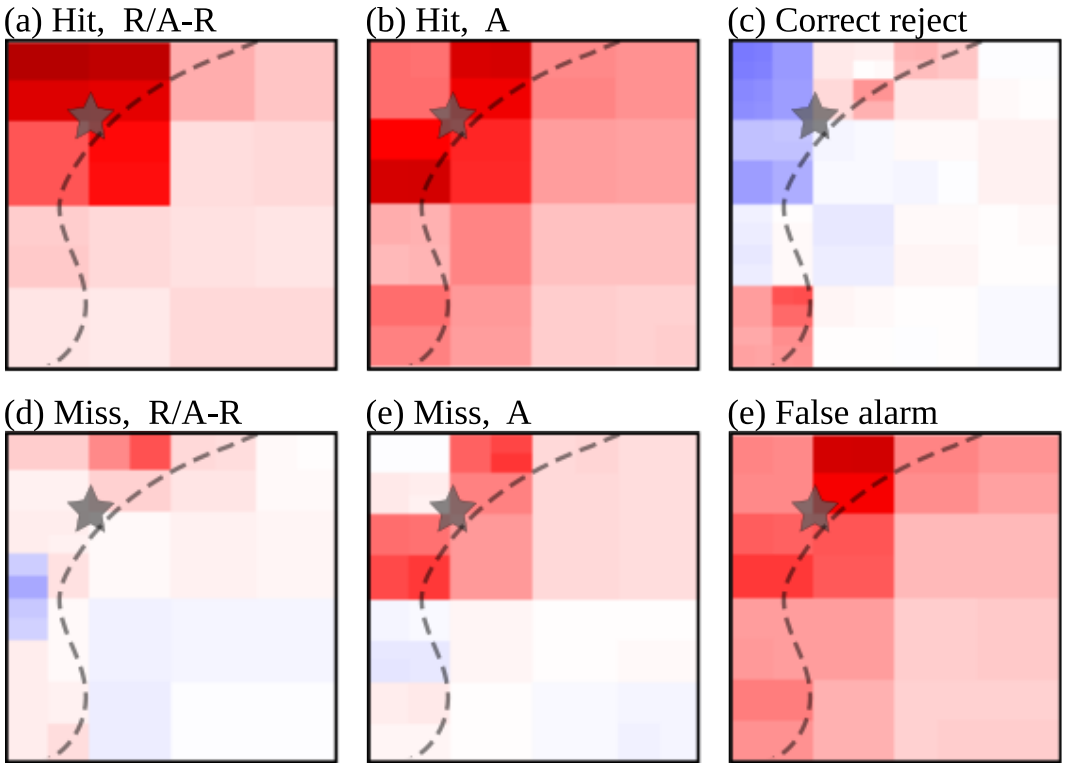


Figure 8: Spatial aggregates of CwPS results based on classification outcome and fog type. To convert a set of local explanations to global, the set of (32, 32, 384)-sized raster explanations have been aggregated into 6 (32, 32)-size explanations. To highlight influential spatial regions, the spatial-channel aggregates are summed along the channels to yield a single 2D spatial explanation. The dotted curve represents the shoreline, with land to the left and water to the right. The star indicates the location of airport KRAS, the source of the fog observations. Red means the feature influenced the model toward the outcome while blue indicates away from.

3.2 Feature Importance

Three feature importance methods were applied to three feature grouping schemes. Because of the computational complexity, only PFI was applied to more granular channel-wise and CwSP schemes. For detailed model insights, granular explanations are preferred. But because of the correlation issues discussed, coarser explanations are expected to be

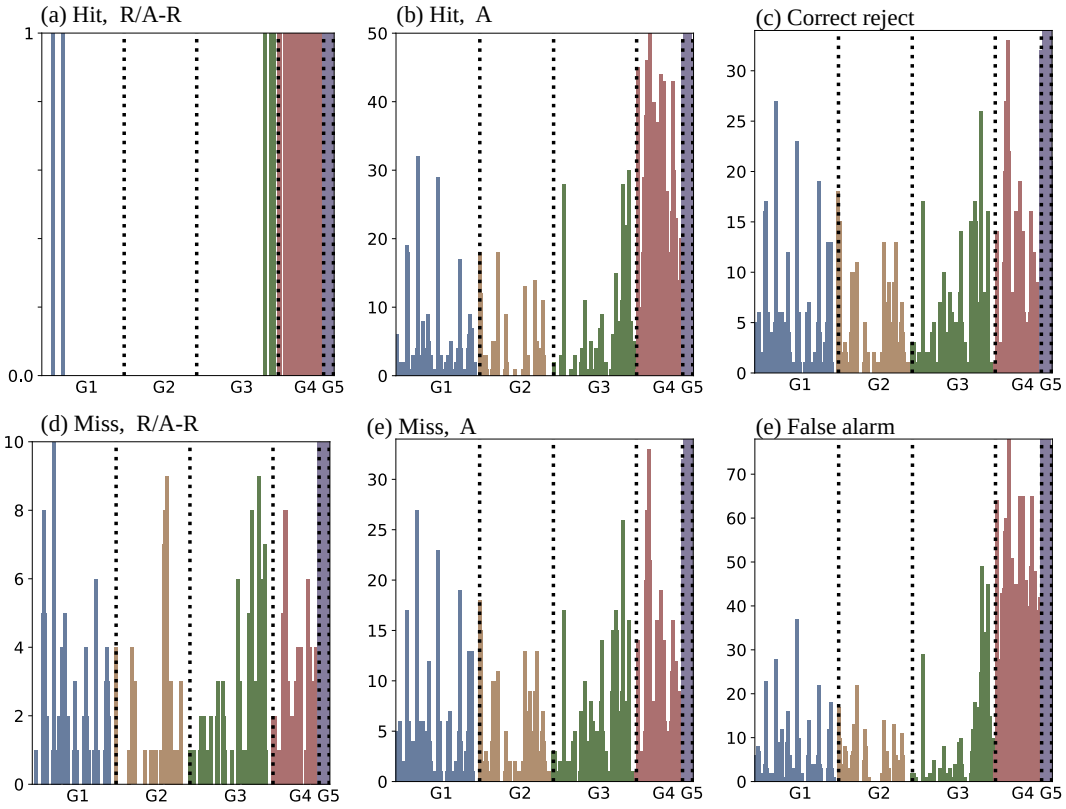


Figure 9: CwPS channel rankings based on classification outcome and fog type. When summing the superpixel SHAP values, the disproportionate influence of **G4** and **G5** channels causes channels in other groups to virtually disappear. We instead ranked channels based on the number of times that a channel appears within the top 50 channels.

more accurate. However, when granular explanations agree with the coarser explanations, there is increased confidence in the accuracy of the more granular explanations. Some groups may be more sensitive to the feature grouping scheme than others. It is possible that the sensitivity comparison will suggest that we can use the coarser explanations for a subset of groups that show greater consistency. However, it is not straightforward to directly compare the importance values at different groups; at each level of feature grouping granularity, we sum PFI values into the coarser groups for a comparison of the rankings.

Another sensitivity check is that of the different XAI methods. Since we are only applying XAI methods to the coarser grouping schemes with PFI, we cannot compare them to other XAI methods directly. However, we can check PFI's consistency with GHO and LS. The latter methods are much more complex and expected to be more robust to issues of correlation and out-of-sample inputs. So if PFI performed on channel groups reaches similar relative feature importance rankings to the others, then we have additional confidence in using the (consistent) more granular PFI explanations.

Figure 10 gives all feature importance results, aligned in a table to assist comparison. Each column corresponds to the grouping scheme used to generate importance values. Column 1 is CwSPs, column 2 is channels, and column 3 is channel groups. Rows correspond to the aggregation level. The top of each column is without any aggregation.

The second row is for channels, and third for channel groups. This figure is discussed in detail in Section 4

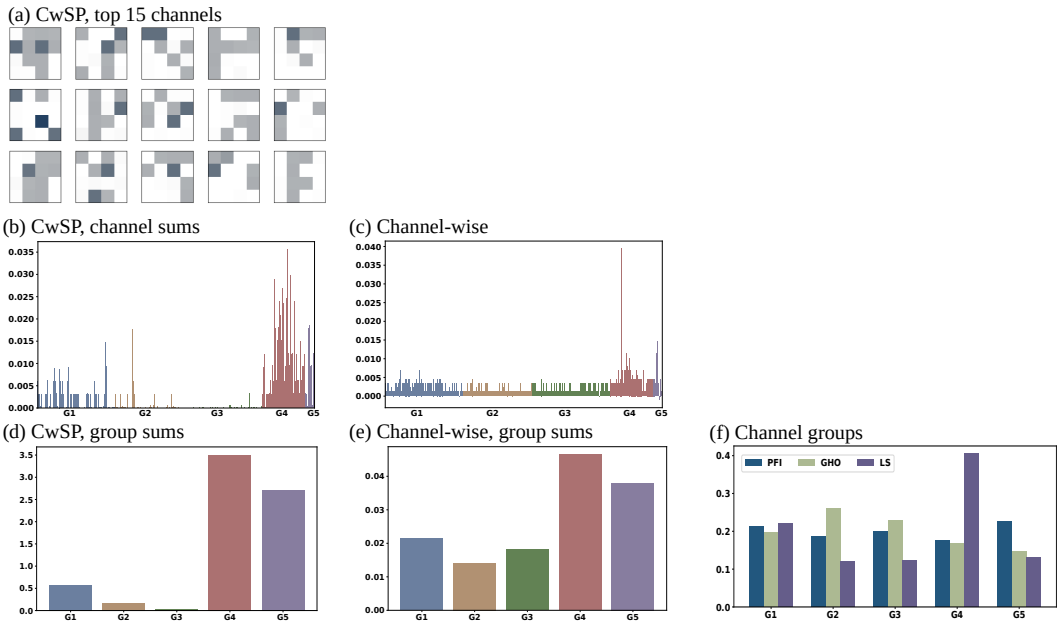


Figure 10: Feature importance XAI methods at three levels of granularity. To compare consistency across grouping schemes, the more granular grouping scheme's explanations are aggregated into coarser ones. a) Each column corresponds to the grouping feature used for the XAI method, and each row corresponds to an aggregation granularity. a) shows the top 15 channels based on PFI performed on CwSP (left-to-right, top-to-bottom). b) and c) show individual features. d) and e) show the prior row aggregated by group, and e shows methods used on the five groups.

4 Discussion

We will use Figure 10 to analyze the sensitivity of explanations to the choice of feature aggregation granularity. Results from PSS-based PFI are shown, but computations were made based on HSS and CSS as well.

Figure 10a shows the results of PFI applied to CwSP features. To assess the consistency of CwSP explanations to channel-wise, Figure 10b shows the summations of absolute PFI values in each channel. This can then be directly compared to Figure 10c, the PFI values computed when PFI is applied directly to channel-wise features. When considering super pixels at the channel level (Figure 10b), the most important channels tend to be within G4 and G5. The top channel occurs in G4: *Vertical velocity at 950mb, t_1*. Sparse G1 and G2 channels have some importance, with practically no importance for G3. When considering individual channels using PFI, Figure 10c, we observe considerable influence from G4 and G5 as we did with the CwSP-based results in Figure 10b. Again, *Vertical velocity at 950mb, t_1* in G4 has the highest importance. But, otherwise, the exact rank order does differ between CwSP and channel-wise explanations.

Two differences stand out between Figure 10b and Figure 10c. In the channel-wise results (Figure 10c), two G4 and G5 channels have such high importance that others are relatively suppressed. This does not occur in the CwSP results (Figure 10b), where some channels in G1 and G2 are shown to be comparable to G4 and G5 channels. Another

Table 2: Top 15 t_3 channels ranked with channel-wise (Cw) and CwSP schemes

Cw (PSS)	CwSP (PSS)	Cw (HSS)	CwSP (HSS)	Cw (CSS)	CwSP (CSS)
G5 TMP-DPT	**G4** VVel 850mb	**G5** TMP-DPT	**G4** VVel 850mb	**G5** SST	**G4** VVel 825mb
G4 VVel 925mb	**G4** VVel 825mb	**G5** SST	**G4** VVel 925mb	**G4** VVel 825mb	**G4** VVel 850mb
G4 VVel 775mb	**G4** VVel 925mb	**G4** VVel 925mb	**G4** VVel 900mb	**G5** TMP-SST	**G4** VVel 800mb
G4 VVel 900mb	**G4** VVel 950mb	**G1** UGD 875mb	**G4** VVel 825mb	**G1** UGRD 950mb	**G4** VVel 925mb
G4 VVel 800mb	**G4** VVel 900mb	**G1** VGD 875mb	**G5** SST	**G1** VGRD 775mb	**G4** VVel 975mb
G4 VVel 875mb	**G4** VVel 800mb	**G4** VVel 900mb	**G4** VVel 950mb	**G4** VVel 850mb	**G4** VVel 900mb
G1 VGRD 875mb	**G4** VVel 700mb	**G2** Q 875mb	**G4** VVel 725mb	**G1** VGRD 10m	**G5** TMP-DPT
G4 VVel 850mb	**G4** VVel 975mb	**G1** UGD 750mb	**G4** VVel 775mb	**G4** VVel 775mb	**G4** VVel 950mb
G4 VVel 975mb	**G5** TMP-SST	**G1** VGRD 10 meter	**G4** VVel 800mb	**G5** TMP-DPT	**G4** VVel 700mb
G4 VVel 700mb	**G4** Q surface	**G4** VVel 700mb	**G4** Surface vis	**G1** UGRD 875mb	**G4** VVel 725mb
G1 UGRD 825mb	**G5** TMP-DPT	**G4** VVel 800mb	**G1** UGD 875mb	**G1** VGRD 800mb	**G5** TMP-SST
G4 VVel 725mb	**G4** VVel 725mb	**G4** VVel 725mb	**G1** UGD 850mb	**G4** VVel 950mb	**G2** Q 950mb
G2 TKE 900mb	**G4** VVel 750mb	**G5** TMP-SST	**G4** LCLT	**G3** TMP 800mb	**G1** VGRD 925mb
G1 VGRD 900mb	**G4** VVel 775mb	**G1** VGD 775mb	**G3** RH 850mb	**G3** RH 2m	**G4** VVel 750mb
G2 TKE 850mb	**G4** VVel 875mb	**G1** UGD 900mb	**G4** VVel 975mb	**G3** DPT 2	**G4** Q surface

difference is that in the CwSP results (Figure 10b), G3 channels are considered to have practically no importance while in channel-wise results (Figure 10c), G1-G3 are approximately uniform in average importance. At the superpixel level, importance scores mean that the specific superpixel (a spatial region within a variable) had influence on the model. At the channel level, importance scores mean that at least some spatial region within the variable had influence. When a channel is important according to Figure 10c but not in Figure 10b, it may suggest that the model is learning a large-scale feature in that channel such that no individual superpixels are important in isolation. By comparing Figure 10b and Figure 10c, we get some insight into the scale of the features learned by the model.

It is also possible that the difference is due simply to randomness in the permutations. However, there is evidence that suggests that, at least to some extent, the difference between Figure 10b and Figure 10c reflects the scale of the learned features. In general, the importance scores are smaller when summing superpixels which suggests that the importance becomes diluted at the smaller scale. Also, the dilution is prominent in G1 - G3 which are vertical profiles where it is expected that granular features will have minimal information about fog in isolation.

The difference between CwSP and channel-wise PFI is further emphasized when summing both to the group level as shown in Figures 10d and 10e. Comparing G4 and G5 across the 2 figures, we observe that their relative importance is consistent. However, we observe G1-G3 importance drop considerably from the transition to the more granular CwSPs. Figure 10f shows PFI, along with LS and GHO, applied directly to the channel groups. Using this grouping scheme, the importance of the groups is more uniform. G4 is now the least important group, instead of the most as it is in Figures 10d and 10e. The manner in which G1-G3 drop in importance as the granularity of the partitions increases suggests that the model is learning large-scale patterns for those groups. G4 and G5, however, are less influenced by the change in granularity which suggests that the important learned information is generally smaller scale so that granular perturbations of the model still influence model performance.

Based on how PFI works and the dilution of some groups across granularities, there is evidence that the differences are due at least in part to the scale of the learned features. It is actually encouraging that XAI provides evidence that the model is able to learn large scale features that take advantage of spatio-temporal relationships across the channels of the high-dimensional input. We argue that our PFI interpretation relies heavily on having performed XAI on three different feature groups. If we were to compare CwSPs to channel groups (skipping channels), we would have less confidence in the interpretation that the discrepancy between them reflects characteristics of the learned patterns. Since we would have only the 2 examples, we would be less confident that the difference is not merely due to randomness or inherent inaccuracy due to correlations. But by including the channel-wise output, we observe G1-G3 reduce in importance in relation to the increase in granularity, which increases our confidence that the explanations reflect reality.

However, the explanation is not entirely satisfying. For example, we can see that G1 is important but we do not really know which parts of the raster compose its learned features. Even if we get the sense that, broadly, across-channel relationships are involved, we don't know if these are spatial, temporal or spatio-temporal. Are there very important channel sequences? With the present computational efficiencies, it would be too computationally complex to perform PFI on all combinations of channels, much less all possible voxels within the raster.

Another concern is the overall accuracy of PFI itself. In addition to PFI, Figure 10f includes group results using 2 other XAI methods: GHO and LS. While we are minimally concerned about the differences in exact magnitudes between the three methods, we are concerned with their disagreement in the group rankings amongst themselves and the aggregation of the CWSP and channel-wise PFI results.

In the case of GHO, some degree of disagreement is expected. As discussed, each feature's importance is based on refitting the model so that the explanation is based on a set of models. This means that the model is given the chance to learn other relationships within the data in the absence of the removed group. This is quite different from the other two XAI methods that are based on models that had access to the removed group during training. Even if a particular model placed high emphasis on particular features, that does not mean that other features could not be used instead to achieve similar performance. Compared with the PFI and LS, the GHO results show relative uniformity in the importance of the groups. This suggests that the model is still able to learn fog prediction strategies by using different feature relationships. The FogNet ablation study (Kamangir et al., 2022) has already shown that the best performance occurs when all groups are present.

The comparisons lead to concerns about the substantial disagreement between PFI and LS. This means that if we were able to use LS for the more granular groups, we might have a very different interpretation of the model. This is concerning because there is evidence that LS results would more accurately reflect how the model works. First, the game theoretic guarantees suggest that SHAP-based methods might have greater reliability. Second, by averaging over the marginal distribution, LS importance scores are based on several comparisons of perturbed features instead of one. Finally, we observed that CwSP PFI results are unreliable: the spatial maps change dramatically from run to run. This is discussed in Section 4.1. On the other hand, the CwPS results are very stable across repetitions.

4.1 Unreliable Spatial Distribution of PFI CwSP Output

By repeating the PFI computations multiple times, very high variance is observed in the output for CwSP results. This was not the case for channel-wise PFI. The output of each CwSP PFI repetition does not significantly alter the ranking of the chan-

nels, when summing the importance scores. But the distribution of those scores among the superpixels is inconsistent. Each complete repetition of CwSP PFI requires 80 hours of computation, so we are unable to run extensive repetitions with present computational capabilities. But among three runs, we observe little similarity among the spatial maps. Since their summed channel-wise rankings are relatively consistent, we choose to analyze the top channels as determined by CwSP PFI to that of channel-wise PFI. But, unlike the CwPS output that produces stable explanations across repetition, we choose to not analyze the spatial distribution of the importance scores.

While feature effect and feature importance methods measure different aspects of the model, is it of interest to compare the PFI values from Figure 10 to the top channels based on CwPS shown in Figure 9. Across all cases, the overall shape of the channel rankings is not unlike that in Figure 10c. Expect that every G5 channel is consistently of very high influence according to CwPS. There is an explanation that makes it consistent with the PFI measures. Recall that feature effect includes when the model uses a feature for incorrect decisions. Comparing Advection fog hits to misses, G5 features have very high influence in both. This means that G5 channels are being used both for decisions that improve and decrease performance. The net effect could be a lowered importance compared to G4. Comparing hits to misses, G4 values have more influence on the hits than misses which would increase G4 importance.

4.2 Meteorological Interpretation

The following meteorological interpretation will involve the following analyses: 1) A description of high-ranking feature effect FogNet features (Figure 7), and an assessment of the physical processes responsible for, and the environmental conditions associated with, fog that are accounted for based on these features. 2) An assessment of the usefulness of features by comparing feature effect contribution to FogNet prediction, a type of analysis similar to that performed by (Clare et al., 2022). 3) A spatial analysis of feature effect output over the FogNet model domain (Figure 7), including an evaluation of the spatial orientation/distribution of the features that correspond to FogNet performance outcome (Hit, Miss, False Alarm, Correct Rejection in Figure 7), similar to the analysis performed by (Lagerquist, 2020). 4) An evaluation of the feature importance XAI output (Table 2, Figure 10) from a meteorological perspective. 5) A meteorological interpretation of aggregate feature effect output over the FogNet model domain in Figure 8 (influential spatial regions.) Finally, 6) A summary of the results in this section and suggestions on how to improve FogNet performance based on XAI results. In this section, the terms `feature` and `channel` will be used interchangeably.

4.2.1 Physical Mechanisms and Environmental Conditions Related to High-Ranking Features Based on Feature Effect

The following is a description of FogNet features considered high-ranking, based on feature effect XAI output of FogNet predictions, and an assessment of the physical mechanisms and environmental conditions associated with fog that are accounted for by these features.

This paragraph describes the features depicted in Figure 7, which details the 6 highest-ranking sets of features associated with FogNet prediction outcome (Hits and Misses for advection and radiation fog, and False Alarms and Correct Rejections for all fog predictions). These featuers are `TMP-SST`, `DPT-SST`, `TMP-DPT`, `Visibility`, `UGRD 875-mb`, `Q 900-mb`, `TKE 925-mb`, and `VVEL` (`various levels within the 975-750 mb layer`). FogNet feature `SST` is the temperature of the earth's surface, defined more specifically as follows: Over water surfaces, a satellite-derived estimate of the temperature of the sea surface. Over land surfaces, the NAM skin (radiometric) temperature, which is a satellite-derived land surface temperature of the top few millimeters (Z.-L. Li et al., 2013). (See section

2.4.7 of Kamangir et al. (2021)). TMP-SST is the difference between the 2-meter air temperature and the temperature of the earth's surface. This difference (TMP-SST≠0) results in a transfer of heat (sensible heat flux) (Taylor, 2015). The feature DPT-SST is the difference between 2-meter dew point temperature and the temperature of the underlying surface. Feature TMP-DPT is the NAM 2-meter dew point depression (difference between the 2-meter air temperature and the 2-meter dew point temperature.) The FogNet feature Visibility is the diagnosis of visibility which accounts for the extinction of light by hydrometeors generated by the NAM microphysics scheme (details in the FogNet paper by (Kamangir et al., 2021)). Feature UGRD 875-mb represent the U wind component at the 875-mb pressure level. Feature Q 900-mb is the specific humidity (mass of atmospheric water vapor to the total mass of air) at the 900-mb pressure level. TKE 925-mb is turbulence kinetic energy at the 925-mb level. Finally, feature VVelX is the vertical velocity at the X pressure level.

Table 3 is a reorganization of Figure 7 to reflect the relationship between high-ranking FogNet features (based on superpixel absolute maximum SHAP values), and corresponding FogNet prediction and feature effect output, organized by advection fog, radiation fog, and no-fog cases. The specific FogNet predictions analyzed are the 24 hour binary predictions of whether the visibility (in mist or fog) is ≤1600 meters. Although the 24 hour FogNet predictions are based on 0, 6, 12, and 24 hour NAM predictions (Kamangir et al., 2021), the feature maps in Figure 7 depict the composite of 24 hour NAM predictions only (T3), which time matches the FogNet prediction/corresponding visibility observation.

FogNet features in Table 3 contribute greatly to FogNet predictions, regardless of FogNet performance (feature effect). An assessment of the physical mechanism and/or environmental conditions associated with fog that are inferred from these features are explained as follows: The feature Visibility is used by FogNet when making fog predictions that include radiation fog, advection fog, and no fog (Figures 7ak, t, and u, respectively.) The use of this feature suggest that FogNet recognizes a relationship between fog and microphysical processes. The feature TMP-SST has high feature effect with respect to radiation, advection, and no-fog FogNet predictions (Figures 7a-f). The condition TMP-SST>0 implies a downward-directed near surface sensible heat flux to the sea, which results in a corresponding heat loss or cooling of the near surface air temperature to the dew point temperature (TMP-DPT=0) resulting in saturation and subsequent fog development (subject to a cloud drop-size distribution that favors the extinction of light and subsequent visibility reduction). Thus, the condition TMP-SST≥0 directly contributes to marine fog development. However, with respect to radiation fog, radiation (not downward-directed sensible heat flux) directly contributes to the cooling of air to saturation and TMP-SST modulates fog development (has a secondary or indirect effect). The feature TMP-DPT has high feature effect with respect to radiation fog (Figures 7m,v) and no fog (Figure 7r), yet not with advection fog. A possible explanation for why TMP-DPT does not appear with respect to advection fog is as follows: The direct influence of TMP-SST to marine advection fog development and the more indirect effect of this feature to radiation (as discussed above), suggests that TMP-SST may have obscured the effect of TMP-DPT for advection fog cases, yet not for radiation fog cases. As mentioned earlier, additional features with high feature effect include DPT-SST, UGRD 875mb, TKE 925mb, and Q 900mb. The likelihood of sea fog increases as DPT-SST increases positively (Cho et al., 2000). In fact, the condition DPT-SST≥0 is used operationally at the U.S. National Weather Service Weather Forecast Office in Corpus Christi Texas, which oversees the study area, to predict marine advection fog along the Middle Texas coast during the cool season. Wind directly influences fog development and thus it is not surprising that UGRD 875-mb is influencial for radiation, advection and no-fog cases (Figures 7ab,7ac, and 7ag).) Feature UGRD 875mb likely has high feature effect given the preference for lighter winds aloft during radiation fog events and significant onshore wind associated with advection fog at KRAS (Koračin et al., 2014; Mohan et al., 2020). Note that TKE 925mb appears only for

advection fog cases, which may reflect the enhanced thermal (mechanical) turbulence above (below) the thermal turbulence interface (TTI) associated with marine advection fog (Huang et al., 2011, 2015). In particular, mechanical turbulence below the TTI (caused by vertical wind shear within the near surface temperature inversion) transports warm and moist/saturated air toward the cooler surface, contributing to fog development (Huang et al., 2011). It must be emphasized that the role of turbulence to fog is complex. Some turbulent mixing is essential for the vertical extension of the fog layer (Stull, 1988; Dupont et al., 2016; Price, 2019). However, high turbulence within an environment where `Q` decreases with height can dissipate fog (Toth et al., 2010; Price, 2019). Note that feature `Q 900mb` is listed in both the radiation and no fog cases. A review of the SkewT-logP profiles at the Corpus Christi International Airport in Corpus Christi Texas (USA), located approximately 41 km west of the target KRAS, corresponding to the time of radiation and advection fog events at KRAS in the 2018-2020 period (not shown), revealed that the mean moist layer (based on the dew point depression or RH) associated with advection (radiation) fog extended from the surface to around 875mb (990mb). These results are consistent with previous research which indicates that SkewT-LogP diagrams associated with radiation fog are characterized by a thin near surface moist layer, followed by much drier air aloft, and that the moist layer is typically deeper during advection fog events (Croft et al., 1997; Mohan et al., 2020). It is possible that the difference between `Q 900-mb` magnitudes when comparing advection and no-fog cases is not significant enough to generate a high feataure effect. Finally, Table 3 indicates that vertical velocities (`VVEL`) within the 975-750mb layer have high feature effect for both fog and non-fog cases. Both radiation and advection fogs are typically associated with synoptic scale subsidence (negative vertical velocities) below the 500mb level (Huang et al., 2011; Yang et al., 2017; Mohan et al., 2020). Thus, XAI results suggest that FogNet (when making predictions) used features that appear to account for microphysical processes, the possible contribution of the TTI to marine advection fog, the near saturated condition at the surface associated with fog, the effect of near surface sensible heat flux to fog development, and synoptic scale air motions during fog development, consistent with domain knowledge. Thus, with respect to coastal marine fog prediction, it is logical to assume that the trust in FogNet by operational meteorologists would be significant.

4.2.2 Assessment of the Utility of Feature Effect Contributions to FogNet Predictions

As mentioned earlier, each row in Figure 7 depicts a set of FogNet features (and corresponding feature map equal to the size of the FogNet domain), organized as a function of the corresponding FogNet prediction outcome (Hit, Miss, False Alarm, Correct Rejection), and fog type (radiation versus advection fog). Each feature map contains the composite isolines of either the raw values of the feature (`TMP-SST`, `DPT-SST`, `VVelX`, `UGRD`, `VGRD`) or the standardized values (all other features). The features chosen are determined based on the CwPS feature effect method and thus these are features used by FogNet when making predictions, regardless of effect on FogNet performance. The regions covered by red (blue) colors are regions where the feature contributed toward (away from) the FogNet prediction outcome (Hit, Mass, False Alarm, Correct Rejection). For example, the red color in Figure 7g means that the feature (`DPT-SST`) pushes FogNet toward the `Hit` outcome and thus toward the prediction of fog occurrence. However, the blue color in Figure 7r means that the feature (`TMP-DPT`) pulls FogNet away from the Correct Rejection outcome and thus away from a prediction of fog non-occurrence, or toward a prediction of fog occurrence. The 7 rows are ranked (top to bottom) based on the absolute maximum of superpixel SHAP values.

Table 3 also includes a comparison between the feature effect output and the FogNet prediction, for the 42 feature maps in Figure 7. In accordance with the type of XAI analysis performed by Clare et al. (2022), if a given feature pushes FogNet toward a positive fog prediction and FogNet actually predicted fog, or if the feature pulls FogNet away

from a positive fog prediction and FogNet predicted no-fog, then that feature was helpful to FogNet (possess utility). For the vast majority (≈83%) of the 42 feature maps in Figure 7, the feature was helpful to FogNet. It is unclear what insight can be derived from the minority of unhelpful cases since these corresponding features were helpful at other times, and span across the advection fog, radiation fog, and no fog cases.

Organized by

4.2.3 Feature Spatial Analysis and Corresponding Feature Effect with Respect to Radiation and Advection Fogs

The following is an analysis of the Figure 7 feature map patterns of the same feature, coincident with the corresponding feature effect output, as a function of FogNet fog prediction outcome (Hits, False Alarms, Misses, Correct Rejection) similar to the analysis performed by (Lagerquist, 2020) with respect to the XAI analysis of a tornado prediction model. However, the only Figure 7 features that appear in all columns (Hits, False Alarms, Misses, and Correct Rejections) on the same row (constant absolute maximum SHAP value), for both radiation and advection fog, are `TMP-SST` and `DPT-SST`, corresponding to the 2 highest-ranked rows (based on the absolute maximum SHAP value). Thus, the following Lagerquist (2020) style analysis is restricted to these 2 rows.

With respect to the highest ranked row (Figures 7a-f), note that for the radiation fog cases, `TMP-SST`<0 at KRAS in Figure 7a corresponds to an upward-directed sensible heat flux, which can delay the onset of radiation fog. Since fog occurred (`Hit` column), it is likely that either the heat flux was insufficient to delay or prevent fog, or `TMP-SST` was inaccurate (recall that the feature maps in Figure 7 are 24-hour NAM predictions.) However, note the region in the northwest corner of Figure 7a, where `TMP-SST` transitions to positive values (supportive of radiation fog) coincident with `TMP-SST` pushing FogNet toward a fog prediction (red color), as expected via domain knowledge. The pattern of the `TMP-SST` feature for the corresponding `Miss` (Figure 7d) differs somewhat, yet the values over KRAS are also negative and the red-color over KRAS means this `TMP-SST` pushes FogNet toward a no-fog prediction, thus away from a fog prediction, which is plausible since `TMP-SST`<0 is detrimental to fog development. With respect to advection fog corresponding to Figures 7b and 7e, note that the `TMP-SST` values at KRAS are slightly negative, which can delay the onset of advection fog. However, the region of `TMP-SST`≥ 0 along the nearshore waters (within 20 nautical miles offshore) in both figures; this pattern is typical of advection fog events (as discussed in Section 4.2.1). Note that the nearshore waters and KRAS are coincident with the region where `TMP-SST` pushes FogNet toward prediction of fog (red colored CwPS output.) Although the False Alarms and Correct Rejections for this first row (Figures 7c,f) should be interpreted with respect to all fog types, the patterns of these 2 feature maps can easily be understood in the context of advection fog (the predominate fog type in the dataset of this study.) Note that for the False Alarms, the `TMP-SST` pattern is similar to that from the corresponding `Hit` feature map in that the weakly positive values over the nearshore waters are retained. Since the vast majority of fog cases were of the advection fog type, we speculate that FogNet learned that such a pattern is consistent with advection fog, hence the prediction of fog, even for cases when conditions were otherwise not conducive to fog (hence the False Alarms.) Extending this logic of the propensity for FogNet to predict advection fog, rather than fog of any type, to Correct Rejection, it is clear that FogNet did not predict fog since the `TMP-SST` advection fog pattern (`TMP-SST`≥ 0 over the nearshore waters) does not exist in Figure 7f. This apparent overreliance of the advection fog strategy to predict all fog types erodes trust in the use of FogNet to predict fog regardless of fog type.

Regarding the second highest ranked row (Figures 7g-l), note that in all columns, the `DPT-SST` values are slightly negative over KRAS, yet should be positive for the advection fog Hit prediction outcome (Figure 7h). `DPT-SST`>0 is essential for advection fog

Table 3: Forty-two of the highest ranking FogNet features based on feature effect[a]

Figure 7 suffix	Feature	Toward/Away from fog prediction at KRAS (CwPS)	Did FogNet predict fog at KRAS?
RADIATION FOG CASES			
a	TMP-SST	Toward	Yes
d	TMP-SST	Away	No
g	DPT-SST	Toward	Yes
j	DPT-SST	Away	No
m	TMP-DPT	Away	Yes
p	VVel 775mb	Away	No
s	VVel 775mb	Toward	Yes
v	TMP-DPT	Toward	No
y	VVel 900mb	Toward	Yes
ab	UGRD 875mb	Away	No
ae	VVel 925mb	Toward	Yes
ah	VVel 750mb	Away	No
ak	Visibility	Toward	Yes
an	Q 900mb	Away	No
ADVECTION FOG CASES			
b	TMP-SST	Toward	Yes
e	TMP-SST	Away	No
h	DPT-SST	Toward	Yes
k	DPT-SST	Toward	No
n	VVel 775mb	Toward	Yes
q	VVel 775mb	Away	No
t	Visibility	Toward	Yes
w	TKE 925mb	Away	No
z	VVel 950mb	Away	Yes
ac	UGRD 875mb	Away	No
af	VVel 900mb	Toward	Yes
ai	VVel 750mb	Away	No
al	VVel 750mb	Toward	Yes
ao	VGRD 775mb	Away	No
NO FOG			
c	TMP-SST	Toward	Yes
f	TMP-SST	Away	No
i	DPT-SST	Toward	Yes
l	DPT-SST	Toward	No
o	VVel 775mb	Toward	Yes
r	TMP-DPT	Toward	No
u	Visibility	Toward	Yes
x	VVel 775mb	Away	No
aa	VVel 750mb	Toward	Yes
ad	VVel 975mb	Toward	No
ag	UGRD 875mb	Toward	Yes
aj	Q 900mb	Away	No
am	VVel 925mb	Toward	Yes
ap	URGD 875mb	Away	No

[a]Feature effect explanations generated using the CwPS partition scheme (Figure 7) organized by CwPS contribution, the corresponding FogNet prediction, fog type (radiation and advection) and fog occurrence. Radiation fog cases include both radiation and advection-radiation fog cases since they are caused by a common mechanism: radiational cooling

(Cho et al., 2000). These NAM-12 24 hour DPT-SST values may be inaccurate when considering the possibility that TMP-SST in Figure 7a is inaccurate, as mentioned before. Yet despite the inconsistency between the sign of DPT-SST and what would be expected based on domain knowledge, there is consistency between the CwPS output and the FogNet predictions (Figure 7h). If we again assume that FogNet has a tendency to predict advection fog rather than fog of any type, the False Alarm feature map (Figure 7i) is easily explainable since the DPT-SST feature map pattern is very similar to the corresponding Hit feature map (Figure 7h).

Additional insights beyond the type of provided by Lagerquist (2020) can be extracted from the spatial feature effect XAI output. Recall that for radiation fog cases (Figure 7a), the TMP-SST feature pushes FogNet toward a prediction of fog over a small region in the northwest corner of the domain. This suggests that the influence of sensible heat flux to radiation fog is local. With respect to advection fog, the strongest influence of sensible heat flux (the darkest red color shading in Figure 7b) is approximately local, with a secondary influence over the nearshore waters (weaker red-color shading in Figure 7b). This is reasonable since advection fog can be local (warm, moist air moving over a cooler surface and resulting in fog development locally) or fog can form over the nearshore waters (TMP-SST>0 in Figure 7b) and advect onshore (and lower the visibility at KRAS). Also, note that the influence of TKE is local (Figure 7w) which is reasonable since TKE is not a conservative quantity and thus cannot be advected from another location. Another insight is provided by the VVEL CwPS feature effect output. Note that the VVEL channel CwPS patterns for fog occurrence (all Hit and Miss cases) indicate that the influence of the feature is uniform (same color) and of similar strength (approximately the same shading) across the domain. This feature effect output likely reflects the fact that advection and radiation fogs generally occur in environments characterized by synoptic scale subsidence of air; since the synoptic scale is much larger than the FogNet model domain, the influence of VVEL would be somewhat uniform across the domain.

4.2.4 Evaluation of Feature Importance XAI Results

In this section, we assess features that strongly contribute to FogNet model performance (feature importance) and the associated physical relationships captured. Thus, we are focused here on features that contribute to FogNet performance. Table 2 depicts the top 15 PFI features ranked separately by the channel-wise (Cw) and CwSP methods, as a function of the following 3 separate performance metrics: Peirce Skill Score (PSS), Heidke Skill Score (HSS), and the Clayton Skill Score (CSS). Note that the list of the top 15 features vary as a function of both PFI method and performance metric. With respect to performance metrics, Murphy (1993) identifies the following 3 types of forecast goodness, when attempting to answer the question, "What is a good forecast?": consistency (correspondence between forecasts and judgments), quality (correspondence between forecasts and observations), and value (incremental benefits to users of the forecasts). Murphy (1993) defines judgments as those "recorded only in the forecasters mind", thus less applicable to FogNet (a non-human). HSS, PSS, and CSS measure quality (all three measure skill, which is an aspect of quality). PSS and CSS also measure value. In particular, the PSS represents the maximum potential economic value realized from forecasts, but only for users with cost/loss ratios that equal the base-rate. The CSS represents the range of cost/loss ratios for which users gain economic value from the forecasts. Thus, by using performance metrics that assess both forecast quality and value, we broaden the list of features that possess high feature importance, thus allowing for the discovery of a more comprehensive/exhaustive list of the features most responsible for FogNet's performance. This feature list may provide value to operational meteorologists.

From a meteorological perspective, the PFI channel-wise and CwSPs strategies in Table 2 appear to capture several mechanisms and environmental conditions associated

with fog development. Note that amongst the highest ranked PFI features with respect to all 3 performance metrics, and for the Cw and/or CwPS PFI strategies, features dew point depression (`TMP-DPT`), `TMP-SST`, and various vertical velocities (`VVEL`) below 700mb appear. These features also possess high feature effect (Figure 7) and are physically related to fog development (Section 4.2.1). Although, `VVEL` appears in the microphysics group (G4), it is likely that the PFI strategies captured subsidence in the ambient environment associated with advection and radiation fogs (Huang et al., 2011; Yang et al., 2017; Mohan et al., 2020). In other words, the kinematic effect of `VVEL` to fog is likely greater than the microphysical effect. Note that the U and V wind components (`UGRD` and `VGRD`) and specific humidity (`Q`) at the 875mb level are amongst the highest ranking features, yet for only the HSS performance metric and Cw PFI strategy. Given that `Q` is a conservative quantity, it is likely that the combination of the these features reflect moist advection associated with advection fog (the vast majority of the fog types represented in the FogNet data set.) The advection of warm moist air maintains the flow of water vapor and the near surface thermal inversion which helps to maintain advection fog (Koračin et al., 2014; Huang et al., 2011, 2015). When examining the results from the CwSP PFI method under HSS only, the microphysical based features `Visibility` and `LCLT` appear. The physical relationship between the NAM visibility to fog was discussed in Section 4.2.1. From a microphysical perspective, the activation of CNN is a necessary condition for fog formation. The likelihood of CCN activation is inversely related to the temperature at the lifted condensation level (`TLCL`). Advection fog tends to be associated with a narrow range of SST values (P. Li et al., 2016), which may explain the importance of SST magnitude to FogNet performance (HSS and CSS; Cw and/or CwPS). Lastly, for both the PSS and CSS (CwPS PFI method only), the specific humidity at the 2 meter level (`Qsfc`) appears, which appears to capture the importance of sufficient moisture content to fog formation. Thus, the combination of channel-wise and CwSP PFI methods to determine the features important to FogNet performance (based on HSS, PSS, and CSS) reveal features that account for both mesoscale and synoptic scale environmental conditions conducive to fog, surface heat fluxes that influence fog development, and microphysical contributions to fog, all in alignment with domain knowledge.

Note from Figure 10 that the relative importance of a feature/channel or group is a function of granularity. A comparison between the coarse channel grouping methods (Figure 10f) to the more granular CwSP scheme (Figure 10b), reveals a major difference in feature importance with respect to G3, which contains channels `TMP` and `RH` at 2-meters, and from 975-mb to 700-mb (at 25-mb increments.) Note the near zero importance of individual G3 features yet the significant importance of G3 as a whole. This disparity is reasonable from a meteorological perspective. Each `TMP` channel has no significant relationship to fog development, however the increase in `TMP` with height (temperature inversion) is critical to fog development (Price, 2019; Koračin et al., 2014; Huang et al., 2015). Except for the 2-meter `RH`, individual `RH` channels from 975-mb to 700-mb are less important to fog. Yet, if `RH`=100% at the 2-meter level, with a much smaller magnitude at 975-mb (e.g. `RH`=20%) and above, the surface to 700-mb `RH` vertical structure/profile (the group of `RH` channels) would be conducive to radiation fog (Mohan et al., 2020; Koračin et al., 2014; Huang et al., 2015) Hence, the negligible feature importance of G3 channels in Figure 10b and the much stronger feature importance of the G3 group in Figure 10f. These results are consistent with those found in Kamangir et al. (2022), where XAI output revealed the collective importance of the channels in G3. Extending this reasoning to the other channels and groups, G4 channels `VIS`, `TMP-DPT`, `Qsfc`, and to a lesser extent, `VVEL`, are individually important to fog development (as discussed earlier), which explains the strong granular importance, and logically a strong group importance. Similarily, each of the channels `TMP-SST`, `DPT-SST`, and `TMP-DPT` in G5 are very important to radiation and/or advection fog formation, and hence a significant granular importance. As a group, G5 is extremely influential, since `TMP-SST`>0, `DPT-SST`>0, and `TMP-DPT`$\rightarrow 0$ are essential for fog development. The importance of wind channels (G1) to FogNet performance is significant, both individually and in groups. With respect to the importance

of individual wind velocities, such velocities near the surface are important to fog (radiation fog requires light wind magnitudes near the surface; advection fog at KRAS tends to require stronger onshore flow). With respect to the influence of the vertical wind profile, the vertical wind shear (change in wind velocity with height) contributes to turbulence which strongly relates to fog; some turbulence is essential for shallow fog to increase in vertical dimension, yet high turbulence can preclude fog by forcing water droplets to collect at the surface in the form of dew, or allow for drier air aloft to mix downward resulting in unsaturated near surface conditions (Toth et al., 2010; Price, 2019). With respect to G2, the feature `TKE 975-mb` is individually important since it may capture thermal and/or mechanical turbulence associated with marine advection fog (as discussed earlier). In addition, `TKE 975-mb` can individually modulate stratus-lowering fog development. However, `TKE` values above 975-mb are less likely to contribute individually to fog. Also, each `Q` channel in the 975 to 700-mb layer is not likely to successfully predict fog, suggestive of limited importance on a granular scale. However, the profile of `TKE` and `Q` are very important to fog development (Baker et al., 2002; Toth et al., 2010).

4.2.5 Feature Effect of Influential Spatial Regions

Based on the output depicted in Figure 8, an assessment of the regions of influence within the domain is provided, for all features used by FogNet to make the 24 hour predictions. Note that when FogNet correctly predicts both radiation and advection fog (Figures 8a,b), the strongest influencial region is near the target KRAS. This suggests that mechanism(s) most responsible for FogNet's prediction performance of both fog types is (are) primarily local. In principle, mechanisms responsible for each fog type can be local and/or non-local. For example, radiation fog can occur in response to nighttime radiational cooling of local near surface moisture to saturation. However, advection-radiation fog initially involves the non-local process of moisture advecting from a maritime source during the daytime, followed by local nighttime radiational cooling (recall that both radiation and advection-radiation fog are caused by the same mechanism: radiational cooling). Advection fog can occur as warm moist air is advected over a cooler surface resulting in saturation and subsequent condensation/fog formation over the cooler surface (local), or the developing fog can be transported from one location to another (non-local). Note that for the advection fog cases (Figure 8b), the nearshore coastal waters serve as secondary region of influence (lighter red color shading over the nearshore waters). Based on the spatial analyses of feature patterns and feature effect contributions from Section 4.2.3, it is speculated that this secondary region refers to the region of cooler SST values over the nearshore waters that facilitates advection fog formation that subsequently moves onshore and lowers the visibility at KRAS. Further, since the advection fog type is the predominate type in the dataset, we argue that FogNet essentially learned advection fog at the expense of other fog types. In addition, note the comparison between the advection fog Hits to the False Alarms (Figures 8b,f). Since the vast majority of fog cases are of the advection type, the False Alarm can be interpreted as associated with advection fog. The pattern associated with False Alarms is nearly identical to the advection fog pattern for Hits. This suggests that during False Alarms, FogNet had a tendency to predict Fog when the influential spatial pattern was similar to that of successful advection fog predictions, even when atmospheric conditions were not conducive to fog development. Note also from Figure 8e that FogNet tends to miss advection fog cases when the influence pattern is local (similar to radiation). We speculate that since FogNet correlates an advection fog pattern (Figure 8b) with a fog prediction, FogNet would thus miss cases where such pattern does not exist (Figure 8e.)

The overreliance of the advection fog strategy to predict all fog types provides additional evidence that FogNet is less trustworthy with respect to the prediction of fog types other than advection fog. This is because each fog type is associated with a unique mechanism. Radiational cooing is the primary mechanism responsible for radiation and advection-radiation fog (Gultepe et al., 2007); advection fog occurs when air advects over

a warmer or cooler surface, resulting in near surface vertical exchanges of heat and moisture that result in the saturation of near surface air (Koračin et al., 2014); stratus-lowering fog involves a pre-existing stratus or stratocumulus cloud $\leq$ 1-km in height that steadily lowers to the surface, initiated by vertical mixing of radiation-cooled air at cloud top (Dupont et al., 2016); frontal passage fog occurs owing to the mixture of air masses of different temperature within the frontal zone (Glickman, 2000); both cold front post-frontal fog, and warm front pre-frontal fog occurs when rain falls into the cold, unsaturated, and stable sub-cloud layer, evaporates and moistens the near surface layer to saturation (Glickman, 2000). It is expected that a fog model that adequately accounts for these unique processes is more likely to skillfully predict fog regardless of fog type, and thus beneficial to operational meteorologists.

4.2.6 Summary and Suggestions for Improving FogNet Based on XAI

In summary, the highest ranking FogNet features demonstrating high feature effect and feature importance appear to capture critical mechanisms responsible for, and environmental conditions associated with, advection and radiation fog at the target. Based on comparisons between feature effect output and corresponding FogNet predictions, it was revealed that in ≈83% of the top 42 SHAP-based feature maps evaluated, the feature effect output was helpful to FogNet with respect to fog prediction. Furthermore, an analysis of the spatial patterns of feature maps as a function of FogNet prediction outcome suggests that FogNet predicts fog occurrence based on a `TMP-SST` feature map pattern consistent with advection fog development (Figure 7b), and that FogNet appears to predict fog whenever this pattern appears (likely because FogNet was trained on a data set where advection fog was the predominate fog type), at the expense of other fog types (associated with different fog-generating mechanisms), thus contributing to False Alarms (similarity in `TMP-SST` spatial patterns in Figures 7b,c). Further, spatial analysis of the strongest feature effect in Figures 7a,b illustrates the local effect of near surface sensible heat flux (using `TMP-SST` as a proxy) to fog development. Additional credence for this lcoal influence is provided by an assessment of the spatial pattern of the aggregate of the influence of all FogNet features (Figure 8), which reveals that the strongest feature effect occurs in a much smaller subsection of the FogNet model domain that surrounds the target (KRAS), suggesting that the primary mechanisms responsible for fog development are local in nature. This pattern analysis also adds credence to the spatial pattern analysis in Figure 7 by illustrating that FogNet tends to predict fog based on whether or not an advection fog spatial pattern exists in the model domain. The reliance on an advection fog pattern to predict all fog types reflects a lower trustworthiness in the use of FogNet by operational meteorologists to predict fog of types other than advection fog, good overall performance notwithstanding. An additional insight is provided. In Section 3.1 (feature effect) of this paper, we suggest that XAI output might explain whether FogNet is applying the strategy to predict advection fog to all fog types, or whether FogNet is learning unique strategies to predict each fog type inefficiently. We argue that both may be occurring. As previously mentioned, results from this study and from (Kamangir et al., 2022) suggest that FogNet is essentially predicting fog based primarily on the patterns it learned from advection fog cases. Also, the coarse resolution or lack of sufficient detail in the surface to 975mb layer renders FogNet less efficient in identifying the patterns of features below 975mb that relate to both radiation and advection fog. For example, having a temperature value at the surface (2-meters), 80-meters, 1000-mb, and at 975-mb will allow for a more accurate assessment of the existence and strength of the low-level temperature inversion than if only the surface and 975mb temperature were available (the current FogNet). A proper assessment of the height and strength of the lower level temperature inversion is critical to the accurate prediction of advection and radiation fog. Finally, the difference in the coarse and more granular feature importance methods with respect to groups reveal the greater importance of the group of channels (the vertical structure of G1-G3 features, and the collection of G4-G5 features) over

individual importance with respect to certain meteorological channels, consistent with domain knowledge and previous research (Kamangir et al., 2022).

The accurate/skillful prediction of radiation fog is extremely difficult; small errors in the representation of the complex radiative and turbulence processes responsible for fog development can result in large prediction errors (Stull, 1988; Gultepe et al., 2007). Conversely, it is relatively simple to predict marine advection fog since it generally requires knowledge of the less complex larger scale dynamics (Gultepe et al., 2007) or wind and sea surface temperature (Stull, 1988). Nevertheless, we argue that the following two (2) actions will improve FogNet's fog prediction performance, regardless of fog type.

First, make adjustments to the features used, and feature grouping, to optimize the benefit of both. Adjustments to the features could involve the addition of some of the same features currently used in FogNet (`TMP`, `Q`, `TKE`, `RH`, `UGRD`, `VRGRD`, and `VVel`), yet at the 1000 mb pressure level, to better capture processes that occur below ≈250 meters that directly influence radiation fog development (Liu et al., 2011; Dupont et al., 2016; Price, 2019) and thus improve FogNet performance (FogNet features did not include NAM output at the 1000 mb level.) Additional features such as the Richardson Number, which measures the ratio of the turbulence suppressing effect of atmospheric stability to the turbulence generating effect of vertical shear (Glickman, 2000), modulates radiation fog (Baker et al., 2002). The mean sea level pressure (`MSLP`) could be added since radiation fog tends to occur in the vicinity of a synoptic scale anticyclone (high-pressure system) (Meyer & Lala, 1990). In addition, the surface equivalent potential temperature (theta-E) could be added, which in combination with `MSLP`, can identify fronts associated with frontal fog (another less frequent fog type that affects KRAS.) Further, an expansion of the groups may be warranted to optimize FogNet performance. For example, the new features `MSLP` and `theta-E` could serve as a separate group to account for the synoptic scale pattern, the transfer of the `VVel` features from the current G4 to G1 given the strong correlation between horizontal and vertical air motion per the continuity relationship, and the transfer of `Qsfc` from G4 to G2 to account for moisture gradients in the surface to 1000mb and 1000mb to 975mb layers. Essentially, the current FogNet feature set is suboptimal. In other words, the FogNet features capture a critical near surface process (near surface sensible heat flux) that influence fog development, some of the microphysical processes associated with fog (e.g NAM visibility algorithm), and general environmental conditions that influence fog development (3D air motions and temperature/moisture profiles). However, critical processes that occur in the surface-1000mb and 1000-975mb layers are not fully accounted for, such as a significant portion of mechanical turbulence that contributes to marine advection fog (Huang et al., 2011, 2015). Further, environmental conditions below 975mb are not fully accounted for, such as a better approximation of the vertical dimension and strength of the lower-level temperature inversion, and of the vertical distribution of moisture, which are critical to the accurate/skillful prediction of both radiation and advection fog (Huang et al., 2016; Price, 2019; Mohan et al., 2020). Finally, the greater resolution of `TKE`, `Q`, and `RH` should improve FogNet's ability to predict stratus-lowering fog (Dupont et al., 2016). Incorporation of these additional features should improve FogNet both performance and trustworthiness by operational meteorologists.

Second, increase the number of radiation fog cases by generating synthetic cases or including additional coastal sites to increase the number of real cases (additional sites should be in close proximity to KRAS to justify using a single domain to train a CNN.) There is evidence from this study (Figures 7 and 8), and from Kamangir et al. (2022), that FogNet - due to the preponderance of fog cases of the advection type - is primarily learning to predict advection fog at the expense of other fog types. By adding more radiation fog cases, we argue that FogNet will better recognize the mechanisms unique to radiation fog, especially when combined with the foregoing first actions.

5 Conclusions & future work

As geoscience continue to adapt complex ML models, there is a need to explain how the models work. XAI has the potential to provide insight into models, both for debugging and to extract novel scientific knowledge from what the model has learned. However, XAI methods are imperfect and may mislead the user. One option is to avoid XAI altogether so as to not be misled, but losing out any potential insights. In this research, we proposed applying XAI when grouping features at multiple levels of granularity and comparing the explanations to guide interpretation.

We performed XAI on channel groups, channels, and CwSPs. We observe some inconsistency among the three partition schemes (Figure 10). Based on channel groups, G3 is a very important channel group. However, CwPS results suggest that it has practically no influence on the model. We argue that each partition schemes asks a different question to the model, and comparing these explanations reveals insights about how the model has learned that would not be possible with a single explanation. G3 is the lower atmospheric moisture and temperature information. This is a 3D vector field that was included in the inputs because of profile characteristics that are correlated with fog. The PFI results demonstrate that the importance of G3 data rapidly decrease as we increase the granularity of the feature groups. This suggests that the model is learning large-scale features in G3, which was the goal. Groups G4 and G5, however, are important even at the most granular level. This allows us to study which regions (e.g. onshore or offshore) are used to make predictions in those groups. Without the multi-scale feature group experiments, we could be misled into thinking that G1 - G3 are not important because their CwSP-based explanations show very little importance or effect.

This research demonstrates a novel methodology for explaining models that use high-dimensional raster predictors. Domain knowledge was applied to interpret the XAI outputs (Section 4.2). Insights regarding the meteorological interpretation of XAI output include: Features with the greatest feature effect/importance captures several mechanisms and environmental conditions associated with coastal fog, based on domain knowledge; the CNN model tends to predict coastal fog based primarily on the strategy used to predict the predominate fog type in the dataset, at the expense of the other types, thus contributing to false alarms given the unique mechanisms/environmental conditions associated with differing fog types; the region influencing CNN output the greatest suggests that the mechanisms responsible for fog at the target were primarily local in nature; feature importance varied with granularity, such that the qualitative difference in feature importance between individual channels and their corresponding group membership could be explained using meteorological reasoning.

5.1 Future Work

We demonstrated that models relying on complex rasters may be highly sensitive to the choice of grouping scheme. However, an issue is the lack of a ground truth attribution. We assume that larger groups will produce more accurate explanations. We also assume that PFI's consistency with GHO and LossSHAP's ranking order indicates accurate relative feature importance from PFI. But this cannot be confirmed without knowing the true attribution. Recently, there has been some research in developing XAI benchmarks: models with ground truth explanations to allow quantitatively ranking XAI methods. Mamalakis et al. (2021) developed a technique for building models where the attribution of each feature toward the output can be directly calculated. This research targets geoscience applications, developing a model whose input is a 2D raster with spatial correlation. They compare several popular XAI methods to determine which is most accurate for explaining that model. This benchmark model could be extended to a multi-channel model where multiple correlation coefficients could be used to make groups of

correlated channels that are combined into a single raster, like FogNet. This multi-channel benchmark could be used to quantitatively compare grouping scheme sensitivity.

We applied several grouping schemes based on geometric partitioning of the raster elements: channels and arbitrary 8×8 superpixels were treated as features. The channel groups were at least defined using forecaster domain knowledge, but not by mathematically analyzing the correlations. The only data-driven definition was performed by CwPS, choosing to recursively divide superpixels when doing so changes the distribution of SHAP values within. Instead, it would be desirable to partition based directly on the characteristics of the data. Spatial statistics could be used to cluster the raster elements into semantically meaningful groups. With complex feature interactions, the ideal groups are not necessarily adjacent elements. Arbitrarily complex volumes could be defined to optimize the accuracy of the XAI results. Given that there are a multitude of methods for clustering (DBSCAN, K-means, Self-Organizing Maps, etc), it remains to determine which data-driven partition best increases explanation accuracy. This is another opportunity to use the XAI benchmarks for a quantitative assessment.

6 Open Research

The FogNet input predictors include NAM NWP model output and the MUR SST analysis product, both of which are in the public domain. FogNet data combines the 12-km NAM output that is archived in grib2 format at `https://www.nco.ncep.noaa.gov/pmb/products/nam` (NOAA, 2006-present) with the Analyzed SST available as a netCDF archive at `https://coastwatch.pfeg.noaa.gov/erddap/griddap/jplMURSST41.html` (NOAA, 2020). The `FogHat` software repository, `https://github.com/conrad-blucher-institute/foghat`, contains utilities to download the NAM and MUR data and generate the FogNet input rasters (Krell et al., 2022a). The predictors and targets used in this study are archived at the FogNet data share server: `https://gridftp.tamucc.edu/fognet/datashare/archive/datasets/` (Krell et al., 2022b).

The FogNet model is available as a software package at `https://github.com/conrad-blucher-institute/FogNet` (Krell et al., 2023). It includes all XAI methods and analysis except for CwPS which is instead implemented as a modification to the `SHAP` Python package by Lundberg & Lee (2017) and has been made available as a fork of the `SHAP` repository: `https://github.com/conrad-blucher-institute/shap` (Krell et al., 2022c). In addition, a companion software repository called `partitionshap-multiband-demo` has been made available with several Jupyter notebooks that demonstrate using CwPS to explain a variety of raster-based models at `https://github.com/conrad-blucher-institute/partitionshap-multiband-demo` (Krell et al., 2022d). `xai-raster-vis-tools`, available at `https://github.com/conrad-blucher-institute/xai-raster-vis-tools`, is another software repository used in this research that contains several scripts for aggregating and visualizing a set of raster XAI outputs (Krell et al., 2022e).

All the software repositories developed as part of the FogNet project (`FogHat`, `FogNet`, `xai-raster-vis-tools`, and `partitionshap-multiband-demo`) as well as the data and scripts archived on the FogNet data share server (Krell et al., 2022b) are released under a Creative Commons 0 1.0 Universal licence. The only exception is the `SHAP` fork containing CwPS since it retains the MIT license used by the original `SHAP` repository.

Acknowledgments
This material is based upon work supported by the National Science Foundation under awards 2019758 and 1828380.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Au, Q., Herbinger, J., Stachl, C., Bischl, B., & Casalicchio, G. (2021). Grouped feature importance and combined features effect plot. *arXiv preprint arXiv:2104.11688*.

Baker, R., Cramer, J., & Peters, J. (2002). Radiaion fog: Ups airlines conceptual models and forecast methods. In *10th conference on aviation, range, and aerospace meteorology.*

Beucher, A., Rasmussen, C. B., Moeslund, T. B., & Greve, M. H. (2022). Interpretation of convolutional neural networks for acid sulfate soil classification. *Frontiers in Environmental Science*, 679.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Cho, Y.-K., Kim, M.-O., & Kim, B.-C. (2000). Sea fog around the korean peninsula. *Journal of Applied Meteorology*, *39*, 2473–2479.

Clare, M., Sonnewald, M., Lguensat, R., Deshayes, J., & V, B. (2022). Explainable artificial intelligence for bayesian neural networks: Toward trustworthy predictions of ocean dynamics. *Journal of Advances in Modeling Earth Systems*, *14*, 1–27.

Covert, I., Lundberg, S., & Lee, S.-I. (2020). Feature removal is a unifying principle for model explanation methods. *arXiv preprint arXiv:2011.03623*.

Croft, P., Pfost, R., Medlin, J., & Johnson, G. (1997). Fog forecasting for the southern region: A conceptual model approach. *Weather and Forecasting*, *12*, 545–556.

Dupont, J., Haeffelin, M., Stolaki, S., & Elias, T. (2016). Analysis of dynamical and thermal processes driving fog and quasi-fog life cycles using the 2010-2013 paris fog dataset. *Pure and Applied Geophysics*, *173*, 1337–1358.

Fei, T., Huang, B., Wang, X., Zhu, J., Chen, Y., Wang, H., & Zhang, W. (2022). A hybrid deep learning model for the bias correction of sst numerical forecast products using satellite data. *Remote Sensing*, *14*(6), 1339.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., & Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, *147*(8), 2827–2845.

Glickman, T. (2000). *Glossary of meteorology.* American Meteorological Society.

Gultepe, I., Milbrandt, J., & Zhou, B. (2017). Marine fog: A review on microphysics and visibility prediction. In D. Koračin & C. Dorman (Eds.), *challenges and advancements in observations, modeling, and forecasting* (pp. 345–394). Springer.

Gultepe, I., Tardif, R., Michaelides, S. C., Cermak, J., Bott, A., Bendix, J., ... Cober, S. G. (2007). Fog research: A review of past achievements and future perspectives. *Pure and Applied Geophysics*, *164*, 1121–1159.

Hamilton, M., Lundberg, S., Zhang, L., Fu, S., & Freeman, W. T. (2021). Model-agnostic explainability for visual search. *arXiv e-prints*, arXiv–2103.

Helber, P., Bischke, B., Dengel, A., & Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(7), 2217-2226. doi: 10.1109/JSTARS.2019.2918242

Hilburn, K. A., Ebert-Uphoff, I., & Miller, S. D. (2021). Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using goes-r satellite observations. *Journal of Applied Meteorology and Climatology*, *60*(1), 3–21.

Huang, H., Huang, J., Liu, C., Mao, W., & Bi, X. (2016). Improvement of regional prediction of sea fog on guangdong coastland using the factor of temperature difference in the near-surface layer. *Journal of Tropical Meteorology: English Edition*, *22*, 66–73.

Huang, H., Huang, J., Liu, C., Yuan, J., Mao, W., & Liao, F. (2011). Prediction of sea fog of guangdong coastland using the variable factors output by grapes model. *Journal of Tropical Meteorology*, *17*, 182–190.

Huang, H., Liu, H., Huang, J., Mao, W., & Bi, X. (2015). Atmospheric boundary layer structure and turbulence during sea fog on the southern china coast. *Monthly Weather Review*, *143*, 1907–1923.

Kamangir, H., Collins, W., Tissot, P., King, S. A., Dinh, H. T. H., Durham, N., & Rizzo, J. (2021). Fognet: A multiscale 3d cnn with double-branch dense block and attention mechanism for fog prediction. *Machine Learning with Applications*, *5*, 100038.

Kamangir, H., Krell, E., Collins, W., King, S. A., & Tissot, P. (2022). Importance of 3d convolution and physics on a deep learning coastal fog model. *Environmental Modelling & Software*, 105424.

Koračin, D., Dorman, C., Lewis, J., Hudson, J., Wilcox, E., & Torregrosa, A. (2014). Marine fog: A review. *Atmospheric Research*, *143*, 142–175.

Kreil, D. P., Kopp, M. K., Jonietz, D., Neun, M., Gruca, A., Herruzo, P., ... Hochreiter, S. (2020). The surprising efficiency of framing geo-spatial time series forecasting as a video prediction task–insights from the iarai traffic4cast competition at neurips 2019. In *Neurips 2019 competition and demonstration track* (pp. 232–241).

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2022a). *Foghat* [software]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7117449` doi: 10.5281/zenodo.7117449

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2022b). *Fognet data share server* [collection]. Conrad Blucher Institute. Retrieved from `https://gridftp.tamucc.edu/fognet/`

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2022c). *Modifications made to shap library* [software]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7117410` doi: 10.5281/zenodo.7117410

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2022d). *partitionshap-multiband-demo* [software]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7117459` doi: 10.5281/zenodo.7117459

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2022e). *xai-rastervis-tools* [software]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.7151017` doi: 10.5281/zenodo.7151017

Krell, E., Kamangir, H., Collins, W., King, S. A., & Tissot, P. (2023). *Fognet* [software]. Conrad Blucher Institute. Retrieved from `https://doi.org/10.5281/zenodo.7892917` doi: 10.5281/zenodo.7892917

Lagerquist, R. (2020). *Using deep learning to improve prediction and understanding of high-impact weather* (Unpublished doctoral dissertation). University of Oklahoma.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, *10*(1), 1–8.

Li, P., Wang, G., Fu, G., & Lu, C. (2016). On spatiotemporal characteristics of sea fog occurrence over the northern atlantic from 1909 to 2008. *Journal of Ocean University of China*, *15*, 958–966.

Li, Z.-L., Tang, B.-H., Wu, H., Ren, H., Yan, G., Wan, Z., ... Sobrino, J. (2013). Satellite-derived land surface temperature: Current status and perspectives. *Remote Sensing of Environment*, *131*, 14–37.

Liu, D., Yang, J., Niu, S., & Li, Z. (2011). On the evolution and structure of a radiation fog event in nanjing. *Advances in Atmospheric Sciences*, *28*, 223–237.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information pro-*

cessing systems 30 (pp. 4765–4774). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *arXiv preprint arXiv:2103.10005*.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199.

Meyer, M., & Lala, G. (1990). Climatological aspects of radiation fog occurrence at albany new york. *Journal of Climate*, *3*, 577–586.

Mohan, T., Temimi, M., Ajayamohan, R., Nelli, N., Fonseca, R., Weston, M., & Valappil, V. (2020). On the investigation of the typology of fog events in a arid environment and the link with climate patterns. *Monthly Weather Review*, *148*, 3181–3202.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., ... Bischl, B. (2020). General pitfalls of model-agnostic interpretation methods for machine learning models. *arXiv preprint arXiv:2007.04131*.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, *116*(44), 22071–22080.

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*, 281–293.

NOAA. (2006-present). *North american mesoscale forecast system* [dataset]. Author. Retrieved from https://www.nco.ncep.noaa.gov/pmb/products/nam/

NOAA. (2020). *Multi-scale ultra-high resolution (mur) sst analysis fv04.1* [dataset]. Author. Retrieved from https://data.noaa.gov/dataset/dataset/multi-scale-ultra-high-resolution-mur-sst-analysis-anomaly-fv04-1-global-0-01-2002-present-mont1

Petterssen. (1940). *Weather analysis and forecasting.* McGraw-Hill.

Price, J. (2019). On the formation and development of radiation fog: An observational study. *Boundary-Layer Meteorology*, *172*, 167–197.

Stull, R. (1988). *An introduction to boundary layer meteorology.* Klumer Academic Publishers.

Tardif, R., & Rasmussen, R. (2007). Event-based climatology and typology of fog in the new york city region. *Journal of Applied Meteorology and Climatology*, *46*, 1141–1168.

Taylor, P. (2015). Air sea interactions — momentum, heat, and vapor fluxes. In G. North & J. Pyle (Eds.), *Encyclopedia of atmospheric sciences (second edition)* (pp. 129–135). Academic Press.

Toth, G., Gultepe, I., Milbrandt, J., Hansen, B., Pearson, G., Fogarty, C., & Burrows, W. (2010). *The environment canada handbook on fog and fog forecasting.* Environment Canada. Retrieved from https://publications.gc.ca/site/eng/9.693869/publication.html

Wallace, J., & Hobbs, P. (1977). *Atmospheric science: An introductory survey.* Academic Press.

Xu, G., Xian, D., Fournier-Viger, P., Li, X., Ye, Y., & Hu, X. (2022). Am-convgru: a spatio-temporal model for typhoon path prediction. *Neural Computing and Applications*, 1–17.

Yang, L., Liu, J.-W., Ren, Z.-P., Xie, S.-P., Zhang, S.-P., & Gao, S.-H. (2017). Atmospheric conditions for advection-radiation fog over the western yellow sea. *Jour-*

nal of Geophysical Research: Atmospheres, *123*, 5455–5468.

Yu, F., Hao, H., & Li, Q. (2021). An ensemble 3d convolutional neural network for spatiotemporal soil temperature forecasting. *Sustainability*, *13*(16), 9174.

Zakhvatkina, N., Smirnov, V., & Bychkova, I. (2019). Satellite sar data-based sea ice classification: An overview. *Geosciences*, *9*(4), 152.

Figure 1.

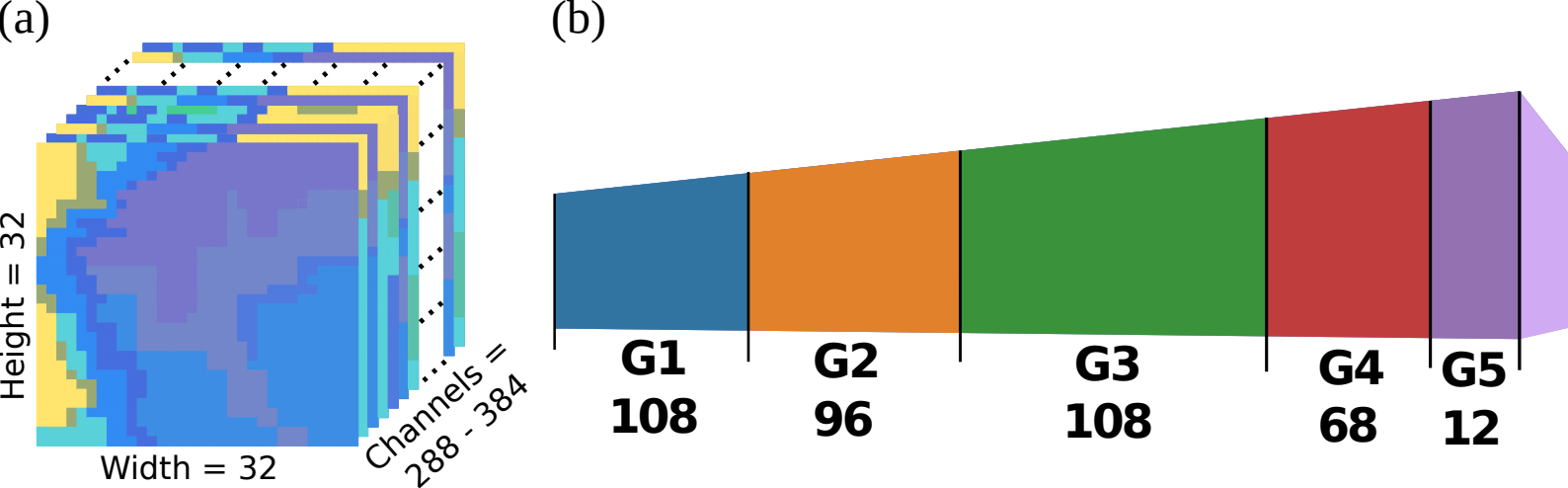

(a)
Height = 32
Width = 32
Channels = 288 - 384
(b)
G1
108
G2
96
G3
108
G4
68
G5
12

Figure 2.

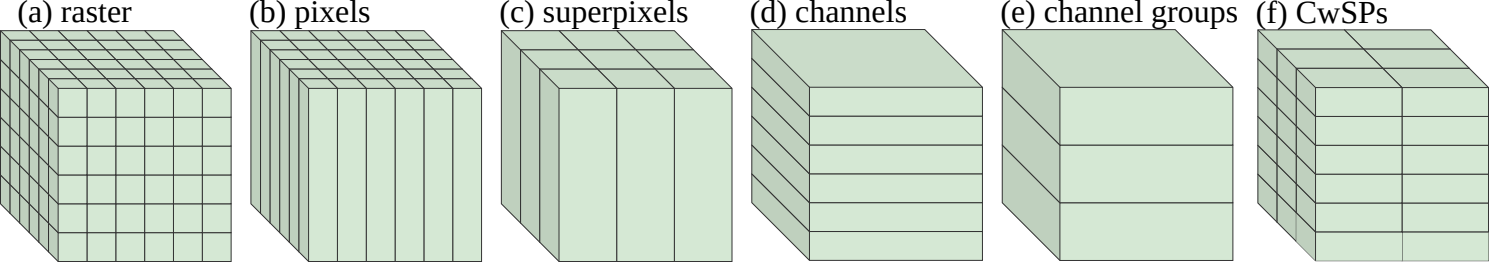
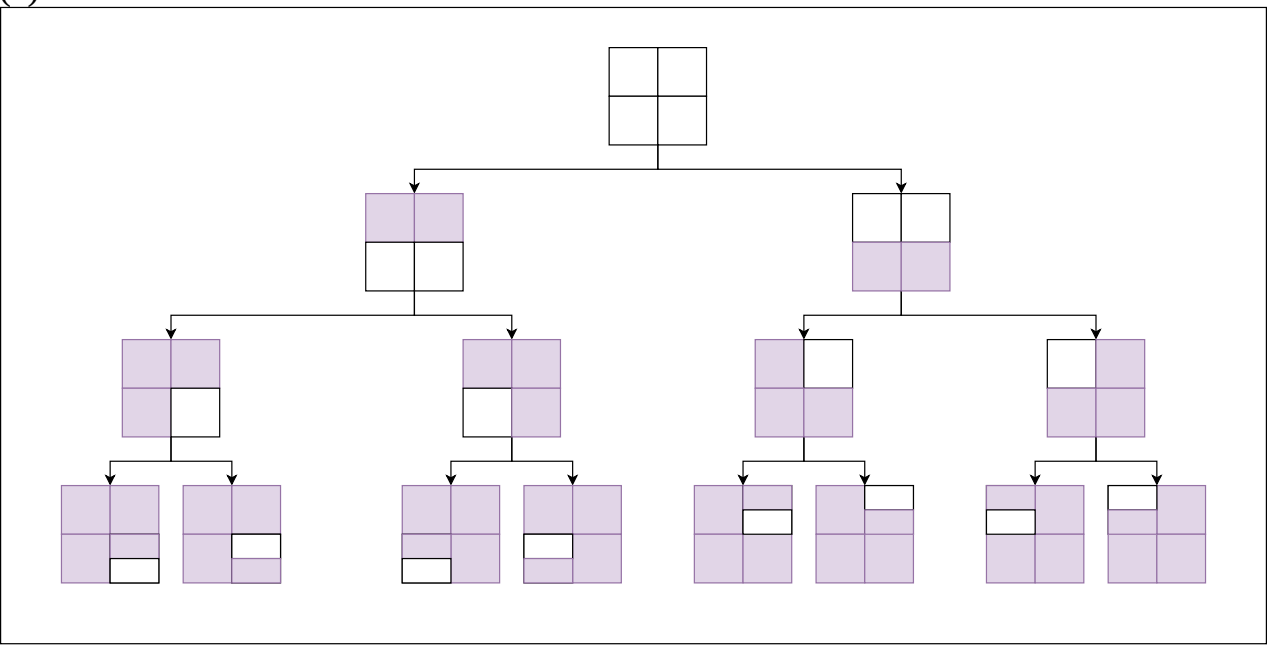

(a) raster
(b) pixels
(c) superpixels
(d) channels
(e) channel groups
(f) CwSPs

Figure 3.

(a) Partition Tree



(b) Owen value calculation

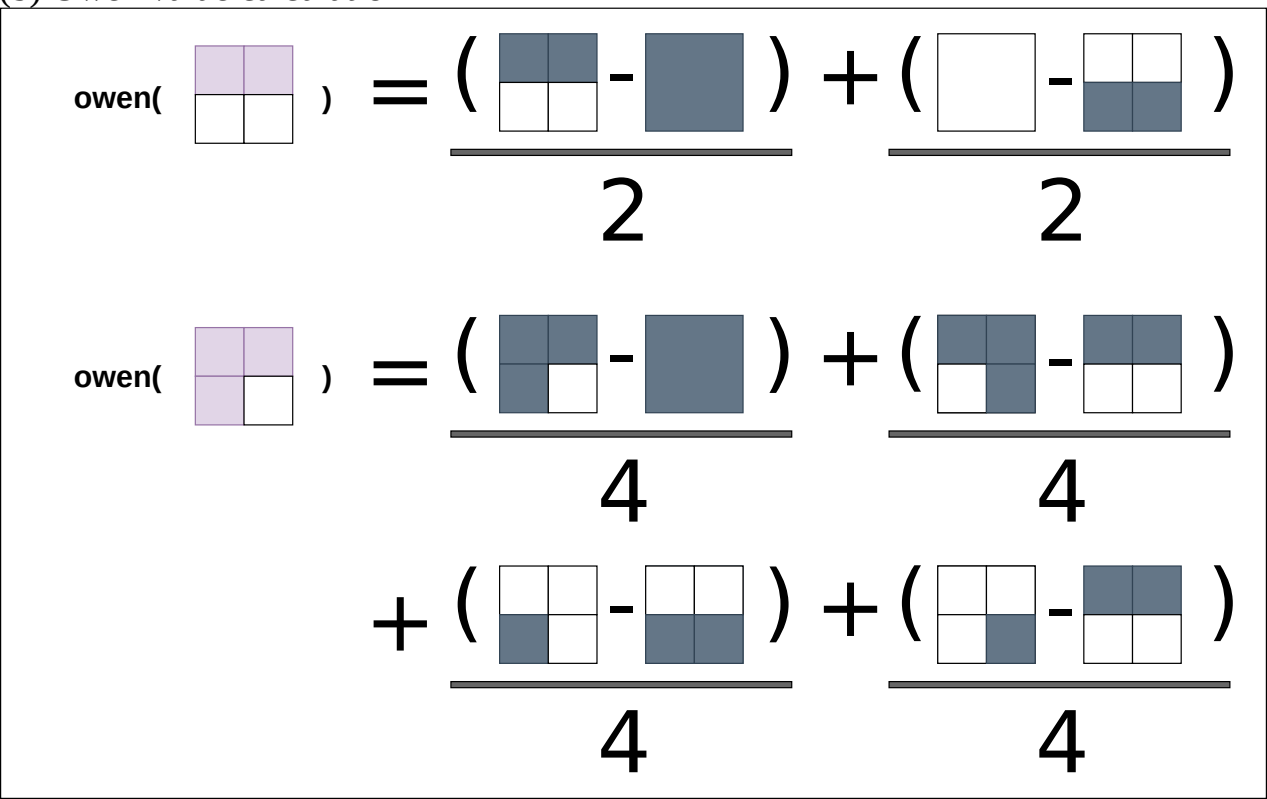
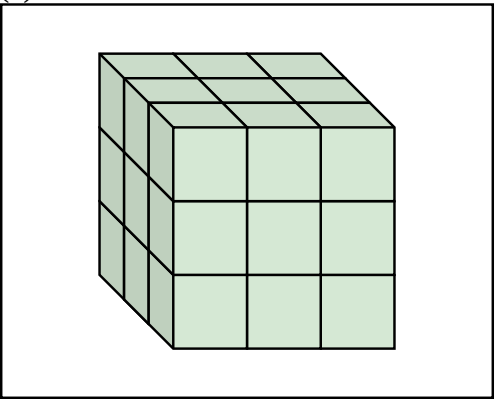
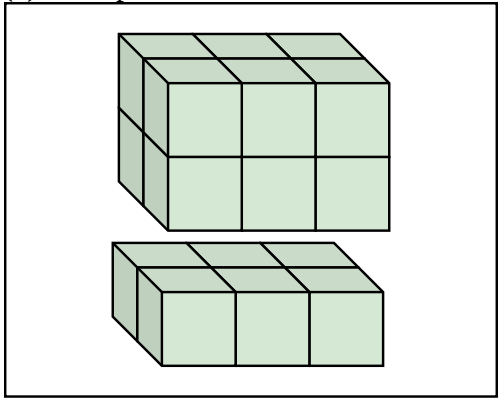


Figure 4.

(a) initial



(a) row split



(a) column split

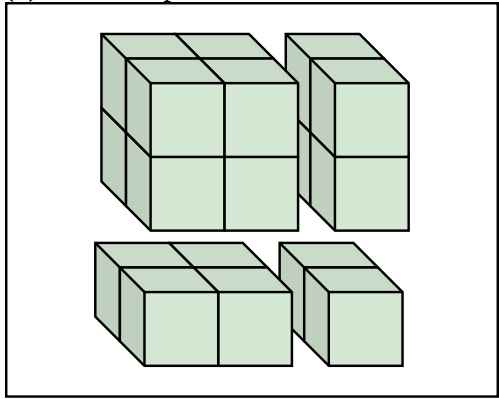
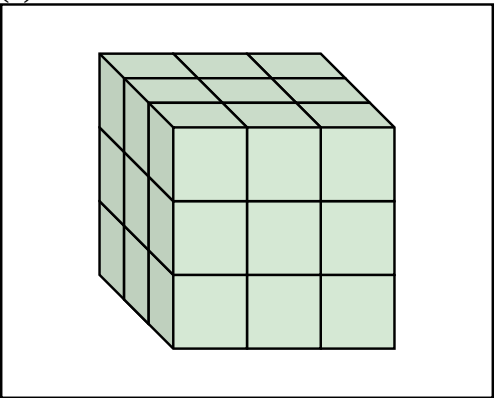
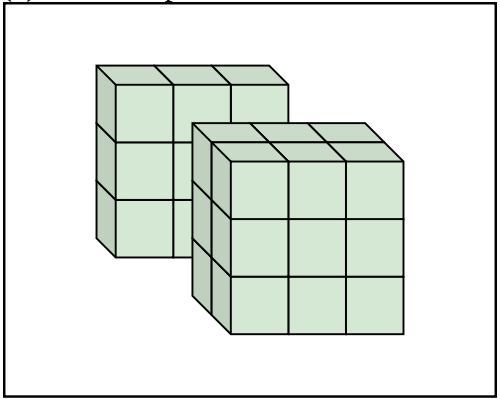


Figure 5.

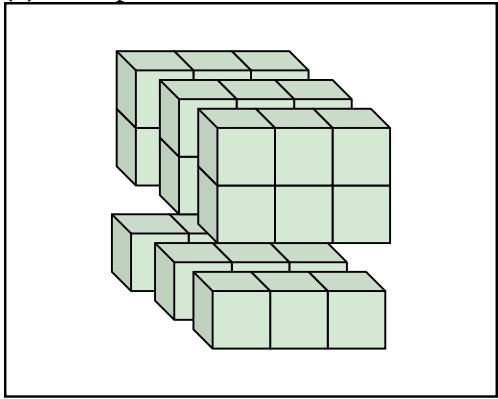
(a) initial



(a) channel split



(a) row split



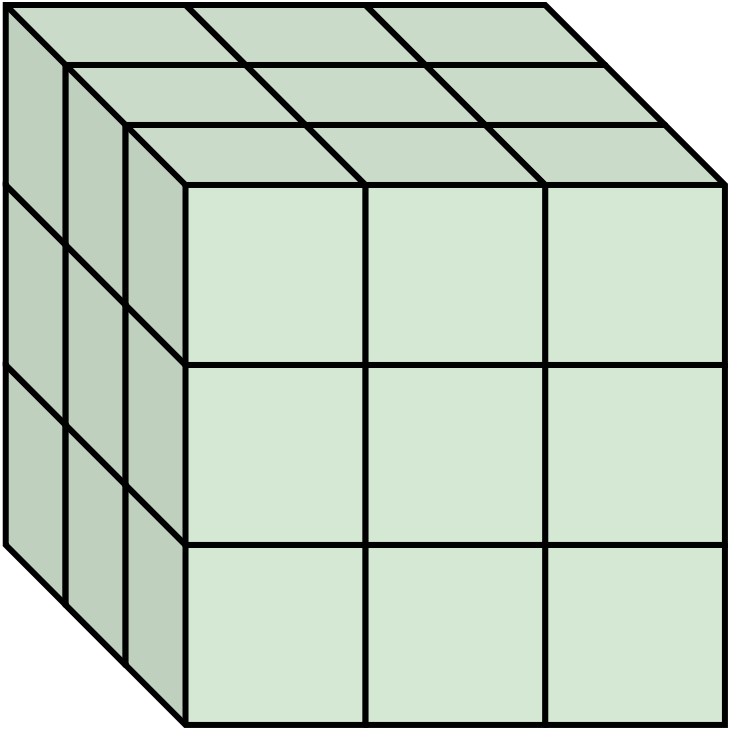


Figure 6.

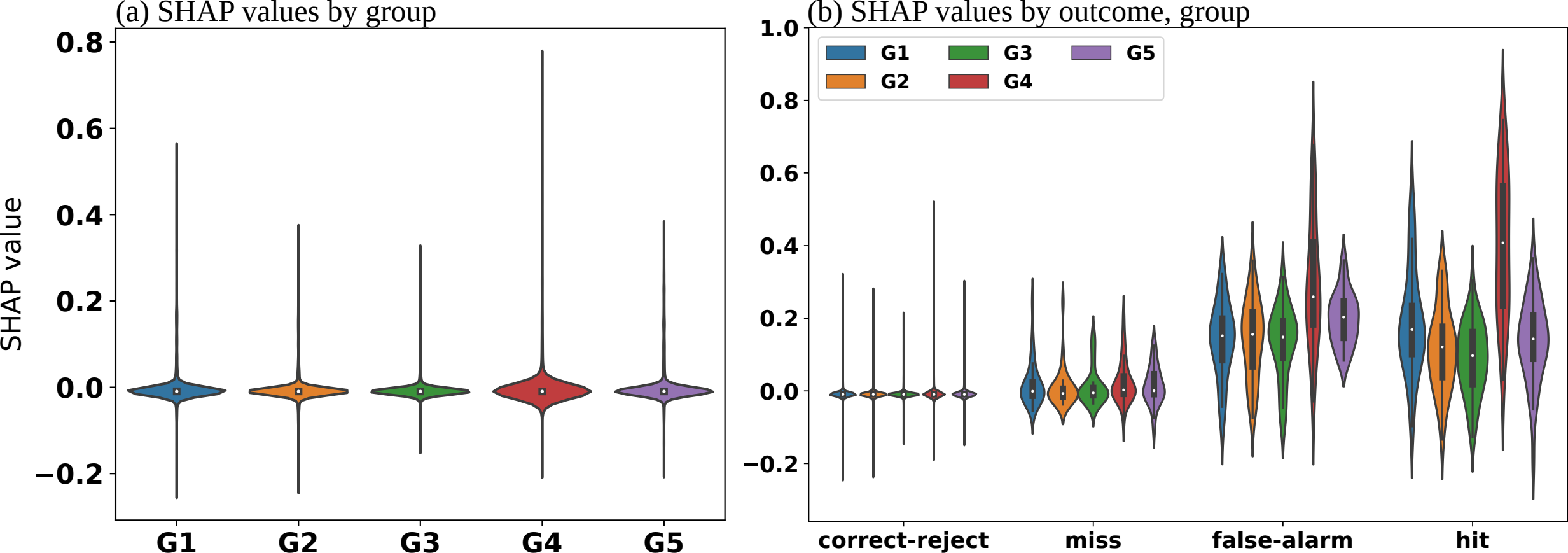
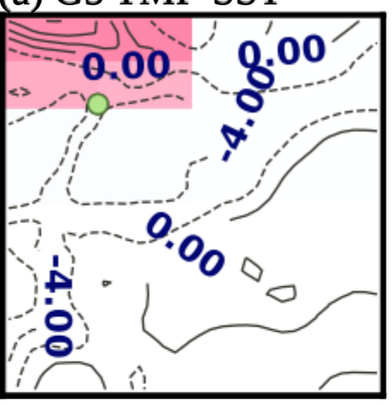
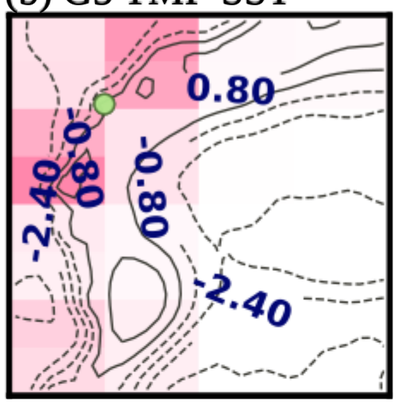
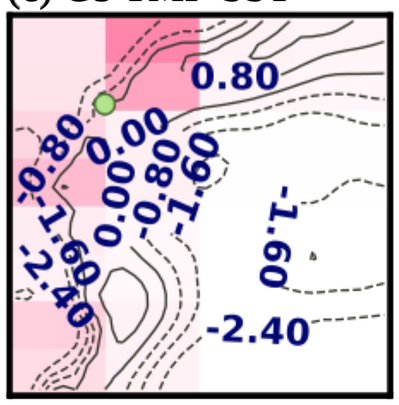

(a) SHAP values by group
SHAP value
0.8
0.6
0.4
0.2
0.0
−0.2
G1
G2
G3
G4
G5
(b) SHAP values by outcome, group
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
G1
G2
G3
G4
G5
correct-reject
miss
false-alarm
hit

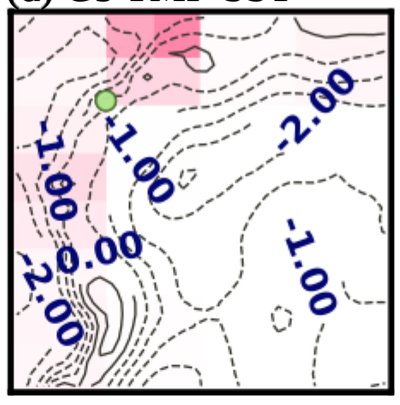
Figure 7.

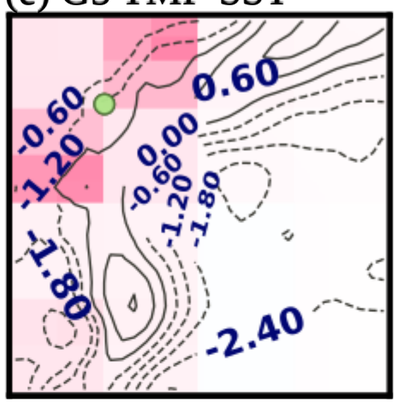
Hit (R/A-R) | **Hit (A)** | **False Alarm** | **Miss (R/A-R)** | **Miss (A)** | **Correct Reject**

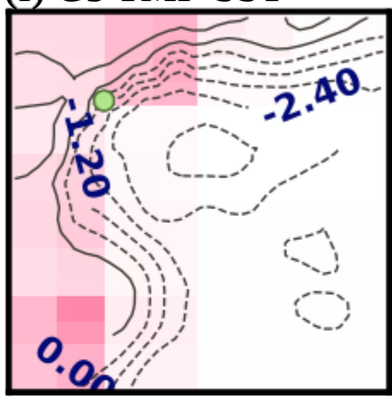
(a) G5 TMP-SST 

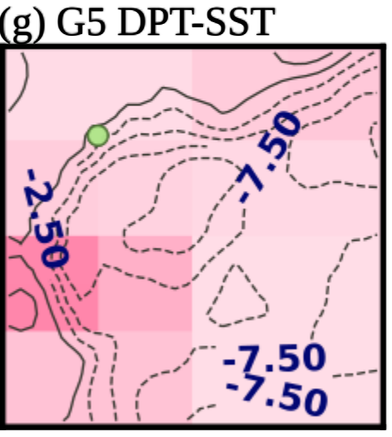
(b) G5 TMP-SST 

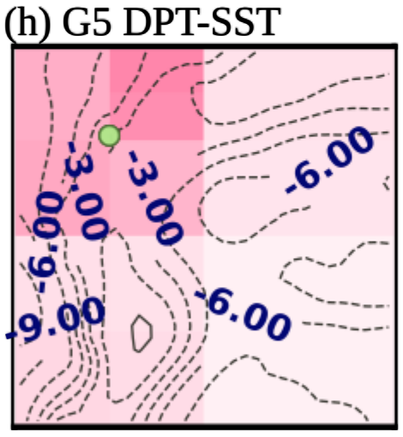
(c) G5 TMP-SST 

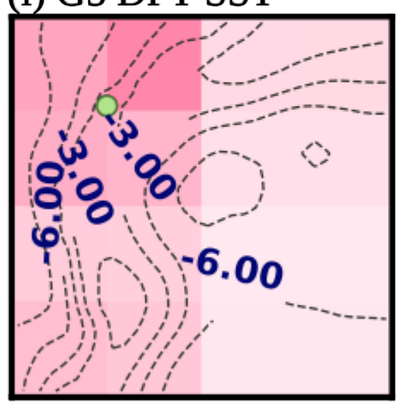
(d) G5 TMP-SST 

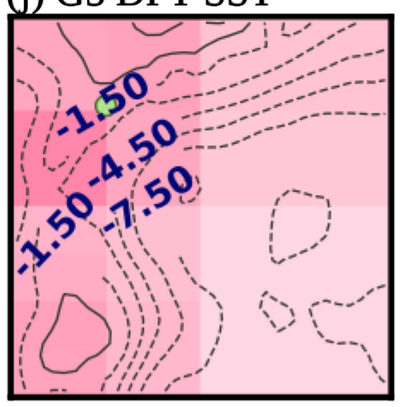
(e) G5 TMP-SST 

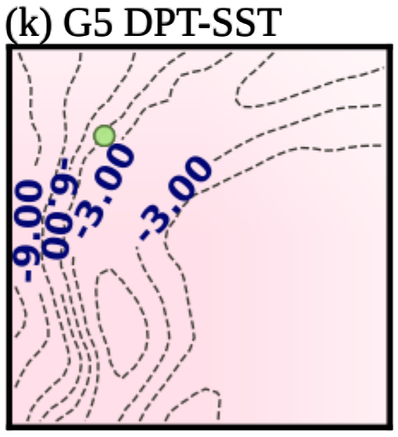
(f) G5 TMP-SST 

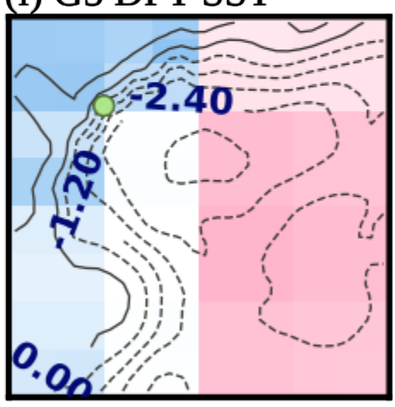
(g) G5 DPT-SST 

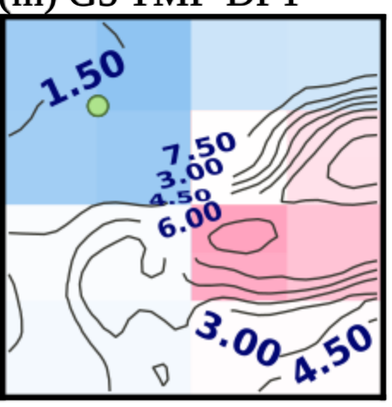
(h) G5 DPT-SST 

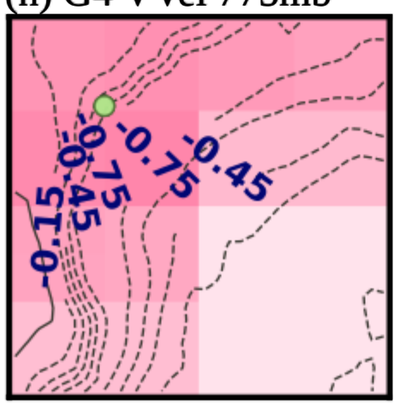
(i) G5 DPT-SST 

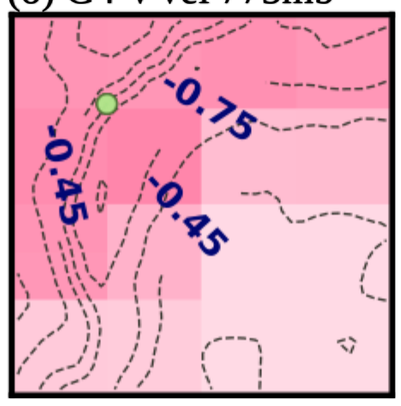
(j) G5 DPT-SST 

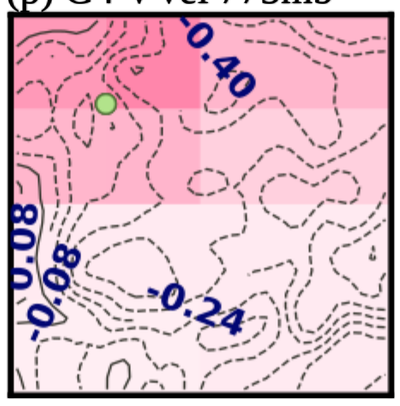
(k) G5 DPT-SST 

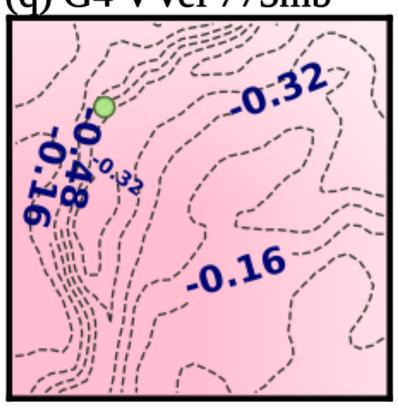
(l) G5 DPT-SST 

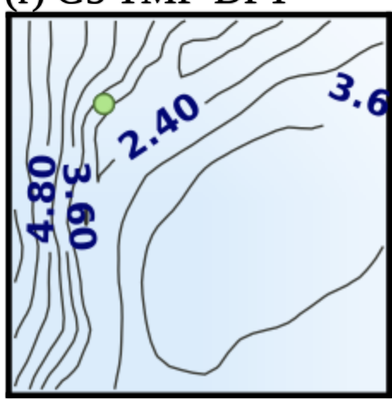
(m) G5 TMP-DPT 

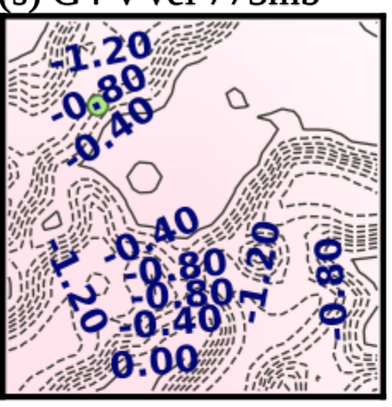
(n) G4 VVel 775mb 

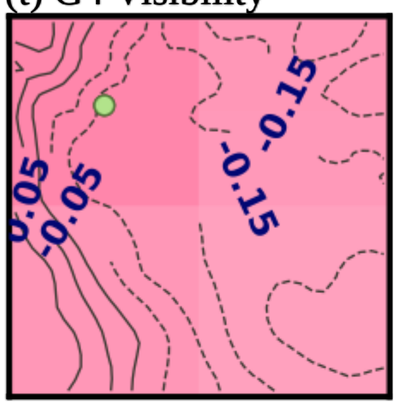
(o) G4 VVel 775mb 

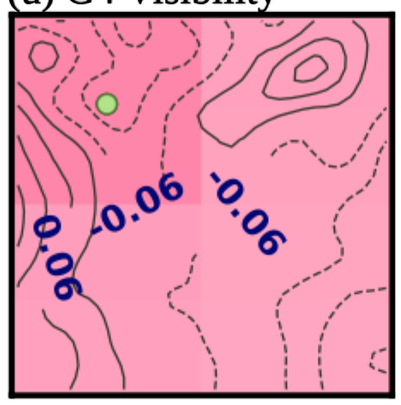
(p) G4 VVel 775mb 

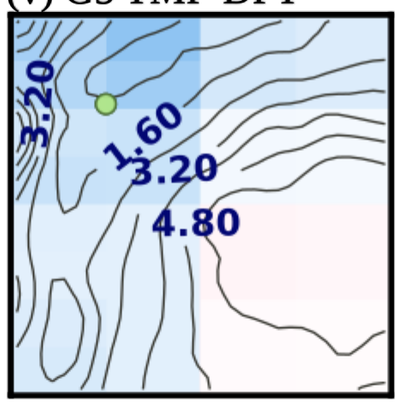
(q) G4 VVel 775mb 

(r) G5 TMP-DPT 

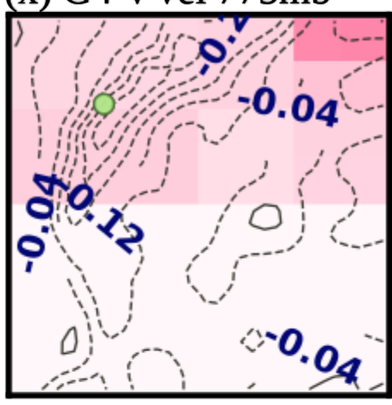
(s) G4 VVel 775mb 

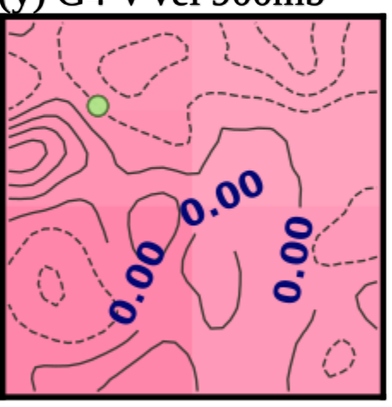
(t) G4 Visibility 

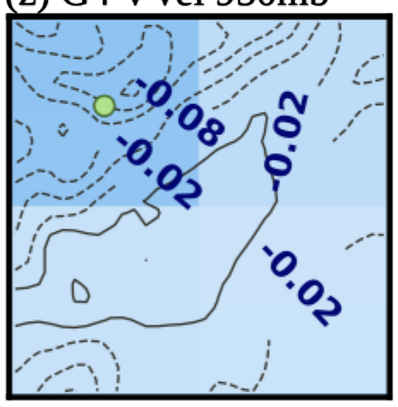
(u) G4 Visibility 

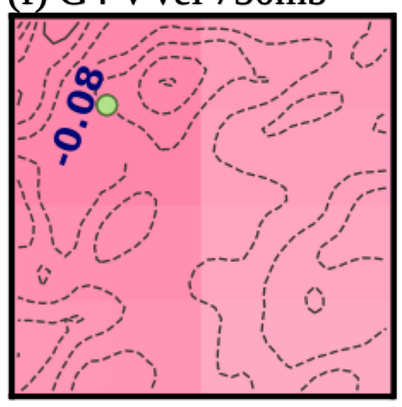
(v) G5 TMP-DPT 

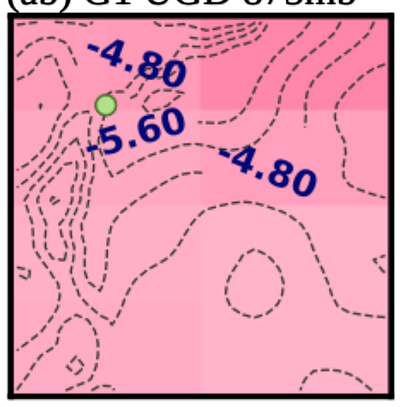
(w) G2 TKE 925mb 

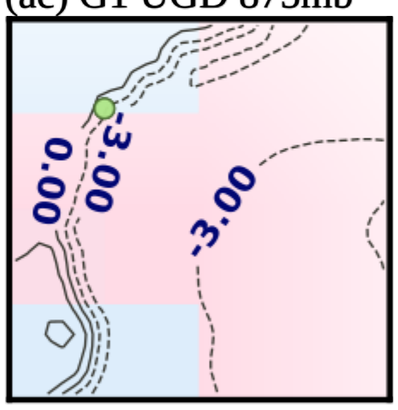
(x) G4 VVel 775mb 

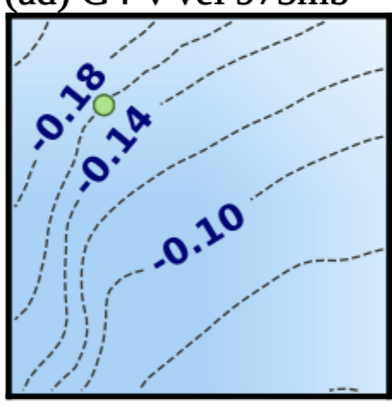
(y) G4 VVel 900mb 

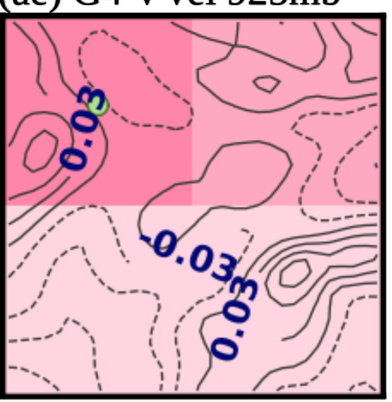
(z) G4 VVel 950mb 

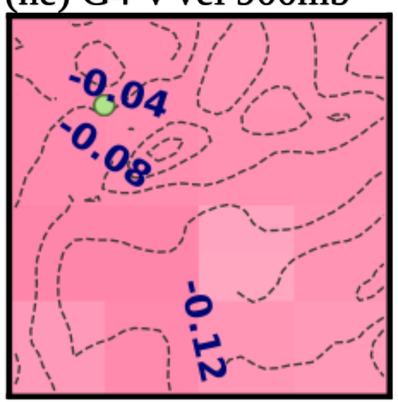
(f) G4 VVel 750mb 

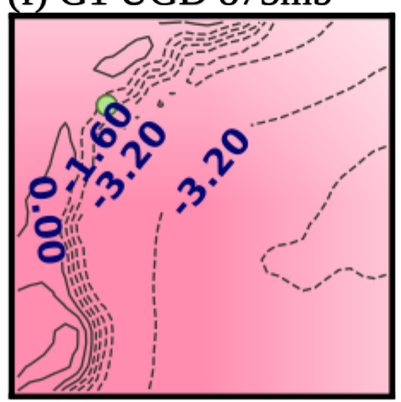
(ab) G1 UGD 875mb 

(ac) G1 UGD 875mb 

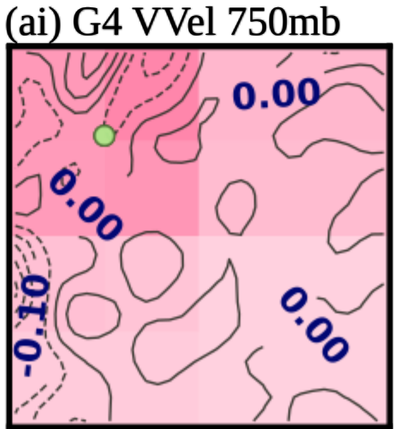
(ad) G4 VVel 975mb 

(ae) G4 VVel 925mb 

(ne) G4 VVel 900mb 

(f) G1 UGD 875mb 

(ah) G4 VVel 750mb 

(ai) G4 VVel 750mb 

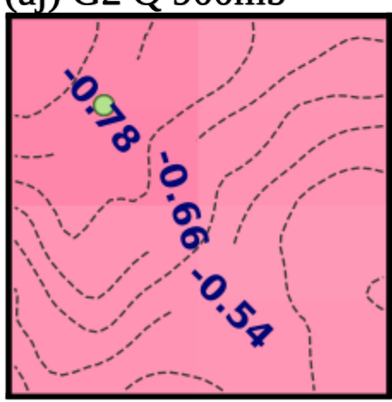
(aj) G2 Q 900mb 

Figure 8.

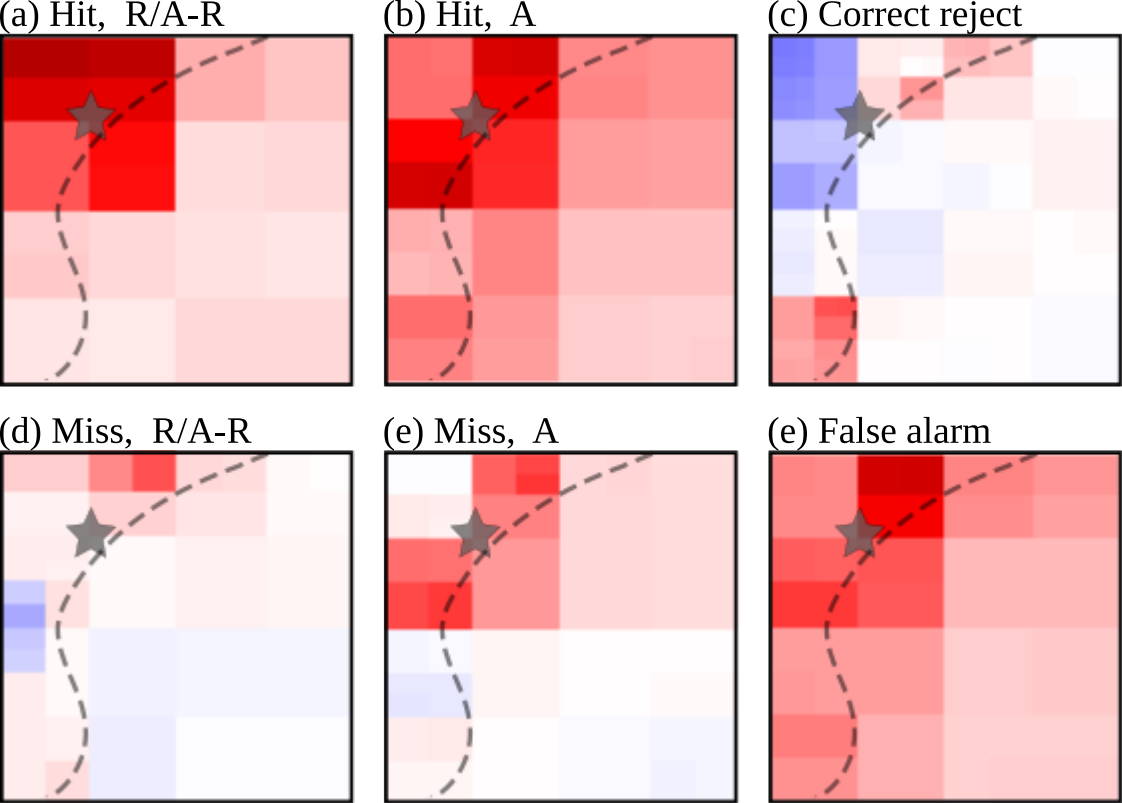

(a) Hit, R/A-R
(b) Hit, A
(c) Correct reject
(d) Miss, R/A-R
(e) Miss, A
(e) False alarm

Figure 9.

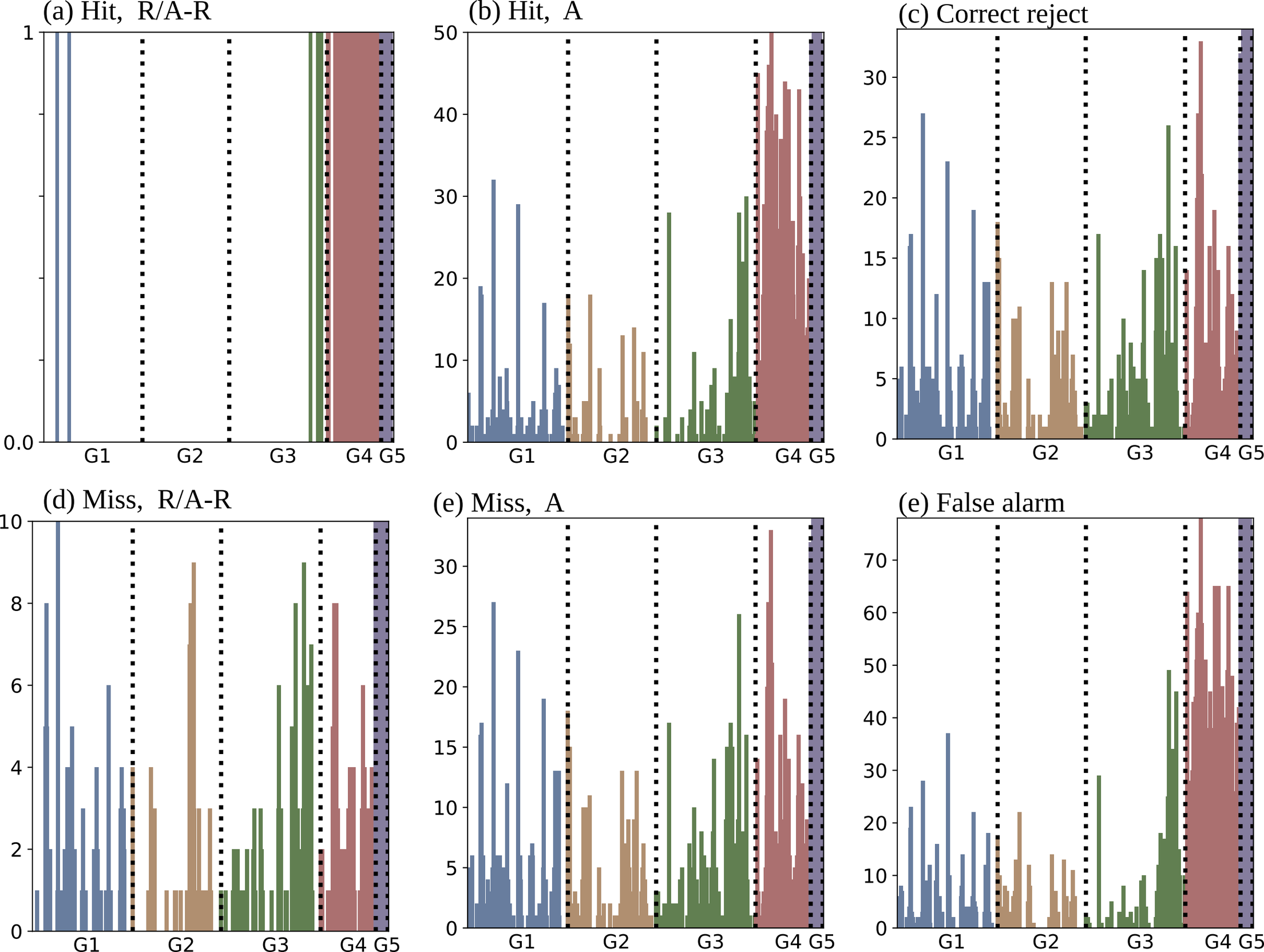
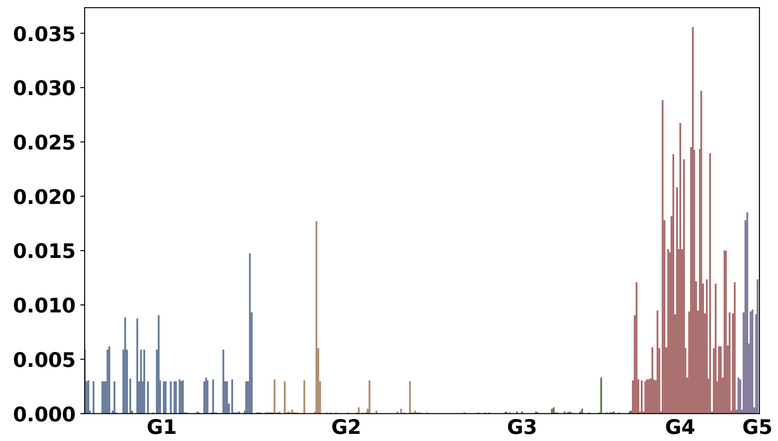

(a) Hit, R/A-R
1
0.0
G1
G2
G3
G4
G5
(b) Hit, A
50
40
30
20
10
0
G1
G2
G3
G4
G5
(c) Correct reject
30
25
20
15
10
5
0
G1
G2
G3
G4
G5
(d) Miss, R/A-R
10
8
6
4
2
0
G1
G2
G3
G4
G5
(e) Miss, A
30
25
20
15
10
5
0
G1
G2
G3
G4
G5
(e) False alarm
70
60
50
40
30
20
10
0
G1
G2
G3
G4
G5

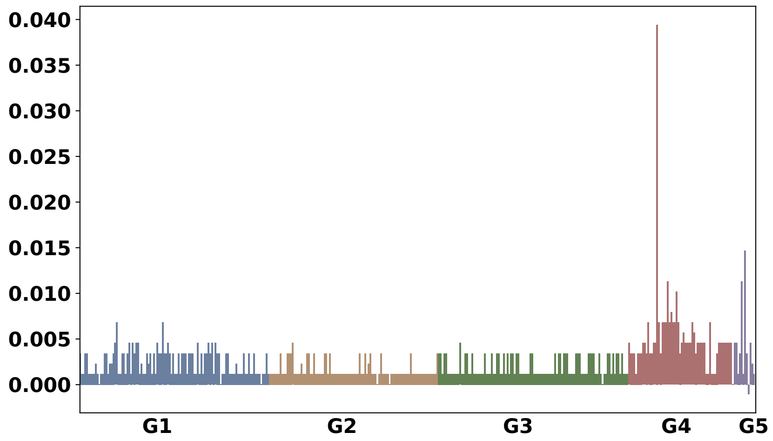
Figure 10.

(a) CwSP, top 15 channels

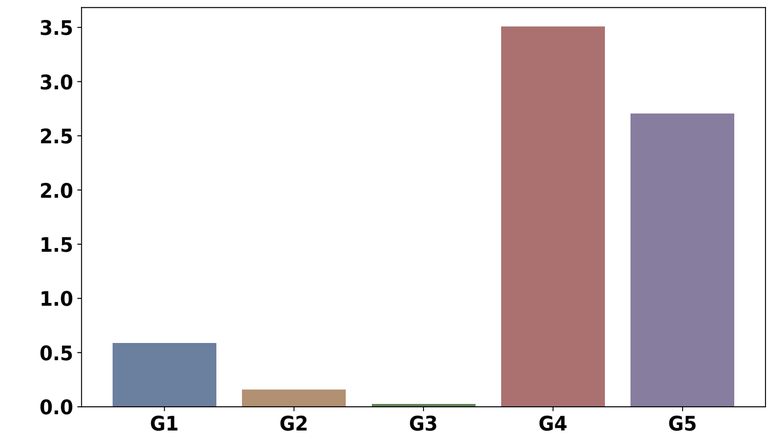
(b) CwSP, channel sums



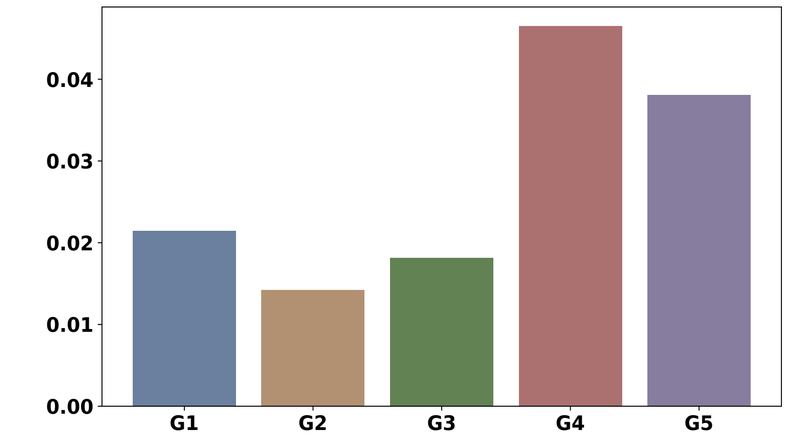
(c) Channel-wise



(d) CwSP, group sums



(e) Channel-wise, group sums



(f) Channel groups