Evaluating the Trustworthiness of Explainable Artificial Intelligence (XAI) Methods Applied to Regression Predictions of Arctic Sea Ice Motion

LAUREN HOFFMAN,[a] MATTHEW R. MAZLOFF,[a] SARAH T. GILLE,[a] DONATA GIGLIO,[b] AND PATRICK HEIMBACH[c]

[a] *Scripps Institution of Oceanography, University of California San Diego, La Jolla, California*
[b] *Department of Atmospheric and Oceanic Sciences, University of Colorado Boulder, Boulder, Colorado*
[c] *Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, Texas*

(Manuscript received 5 April 2024, in final form 20 November 2024, accepted 23 December 2024)

ABSTRACT: Recent advances in explainable artificial intelligence (XAI) methods show promise for understanding predictions made by machine learning (ML) models. XAI explains how the input features are relevant or important for the model predictions. We train linear regression (LR) and convolutional neural network (CNN) models to make 1-day predictions of sea ice velocity in the Arctic from inputs of present-day wind velocity and previous-day ice velocity and concentration. We apply XAI methods to the CNN and compare explanations to variance explained by LR. We confirm the feasibility of using a novel XAI method [i.e., global layerwise relevance propagation (LRP)] to understand ML model predictions of sea ice motion by comparing it to established techniques. We investigate a suite of linear, perturbation-based, and propagation-based XAI methods in both local and global forms. Outputs from different explainability methods are generally consistent in showing that wind speed is the input feature with the highest contribution to ML predictions of ice motion, and we discuss inconsistencies in the spatial variability of the explanations. Additionally, we show that the CNN relies on both linear and nonlinear relationships between the inputs and uses nonlocal information to make predictions. LRP shows that wind speed over land is highly relevant for predicting ice motion offshore. This provides a framework to show how knowledge of environmental variables (i.e., wind) on land could be useful for predicting other properties (i.e., sea ice velocity) elsewhere.

SIGNIFICANCE STATEMENT: Explainable artificial intelligence (XAI) is useful for understanding predictions made by machine learning models. Our research establishes trustability in a novel implementation of an explainable AI method known as layerwise relevance propagation for Earth science applications. To do this, we provide a comparative evaluation of a suite of explainable AI methods applied to machine learning models that make 1-day predictions of Arctic sea ice velocity. We use explainable AI outputs to understand how the input features are used by the machine learning to predict ice motion. Additionally, we show that a convolutional neural network uses nonlinear and nonlocal information in making its predictions. We take advantage of the nonlocality to investigate the extent to which knowledge of wind on land is useful for predicting sea ice velocity elsewhere.

KEYWORDS: Arctic; Sea ice; Machine learning; Model interpretation and visualization; Neural networks; Regression

1. Introduction

Machine learning (ML) is emerging as a powerful tool for applications in Earth sciences (Eyring et al. 2024; Reichstein et al. 2019; McGovern et al. 2019; Gil et al. 2018; Karpatne et al. 2019; Toms et al. 2020; Camps-Valls et al. 2021; Gordon and Barnes 2022). Deep learning models in the form of neural networks can make highly skilled predictions through the use of a hierarchy of features that allows the models to incorporate complex, nonlinear relationships between the predictors. A major limitation of deep learning models has been their lack of interpretability. Neural networks are often thought of as "black-box" models because their complex architecture makes it difficult to interpret the reasoning behind any given prediction (Karpatne et al. 2019; McGovern et al. 2019). In fact, linear models are often used instead of deep learning for applications that favor interpretability at the expense of a more skillful prediction (Schneider et al. 2021).

However, recent advances in explainable artificial intelligence (XAI) techniques have shown promise in probing the inner workings of the black box and providing a useful understanding of how complex ML models make their predictions (Samek et al. 2021). For applications in geosciences, these XAI methods are used to understand when and where the input features are relevant or important in making predictions and can provide this information in the form of useful relevance heatmaps (McGovern et al. 2019; Toms et al. 2020). There are several examples of recent studies using XAI to both understand ML models and answer scientific questions (Ebert-Uphoff and Hilburn 2020; Gagne et al. 2019; Toms et al. 2020; Hilburn et al. 2021; Labe and Barnes 2021). McGovern et al. (2019) provided a comprehensive review of several different XAI methods applied to problems in the geosciences. Furthermore, recent studies have focused on comparing explainability outputs from different methods

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-24-0027.s1.

Corresponding author: Lauren Hoffman, lahoffma@ucsd.edu

DOI: 10.1175/AIES-D-24-0027.1 e240027

(Mamalakis et al. 2022; Flora et al. 2024; Bommer et al. 2024). An understanding of how ML predictions are made not only improves user trust in the model predictions but also, in the case of a skillful model, can be used to draw information from the data and learn about emergent physical behaviors that have not yet been recognized (Murdoch et al. 2019; Ebert-Uphoff and Hilburn 2020).

In this study, we employ linear regression (LR) and convolutional neural network (CNN) models to make 1-day predictions of ice motion from inputs of present-day wind velocity and previous-day ice velocity and concentration. We build off of the models described in Hoffman et al. (2023), which were shown to have high predictive skill. We apply XAI to our statistical models to confirm historical findings that show that wind velocity explains a large portion of the variability in sea ice motion in the central Arctic on 1-day time scales (Thorndike and Colony 1982) and to gain further understanding of what information in the input features is most relevant or important for the model predictions of sea ice motion. We explore the feasibility of using a novel, global formulation of an XAI method known as layerwise relevance propagation (LRP) for this particular application. In this case, "global" refers to the XAI method providing explanations for all predictions (grid points), compared to a "local" method that would only explain the prediction at one grid location (Molnar 2020). We evaluate feasibility based on the consistency of local and global LRP explanations with other XAI methods and our understanding of the physics of sea ice motion. We put an emphasis on analyzing the localization of XAI explanations because our main goal is to extend the scope of LRP from local to global. Additionally, we analyze the degree of linearity of the CNN predictions to identify possible discrepancies between explanations given by the LR and CNN. In summary, we investigate the following questions:

- What do the outputs from XAI methods show us about the contribution of the various input features for making 1-day predictions of sea ice motion in the Arctic?
- To what extent does the nonlocality of the CNN predictions and XAI explanations provide information about the contribution of the input features on land to the prediction of sea ice motion offshore?
- To what extent does a CNN rely on nonlinear information when making predictions?

Following this introduction, section 2 discusses the datasets and the processing steps applied to the data before they are used to train the models. In section 3, we provide details about the model setup and descriptions of the various XAI methods applied to the models. We show the results from both localized and global XAI methods applied to the LR and CNN models in section 4. We finish with a discussion of these results in section 5 and final remarks in section 6.

2. Data

We train the LR and CNN models based on the frameworks in Hoffman et al. (2023) to make predictions of present-day zonal and meridional sea ice velocity ($u_{i,t}$ and $v_{i,t}$) from inputs of

- present-day zonal and meridional wind velocity ($u_{a,t}$ and $v_{a,t}$),
- previous-day zonal and meridional sea ice velocity ($u_{i,t-1}$ and $v_{i,t-1}$), and
- previous-day sea ice concentration (c_{t-1}).

Daily maps of wind velocity, ice velocity, and ice concentration from 1989 to 2021 are from satellite and reanalysis products (Table S1). Wind velocity is from the Japanese Meteorological Agency 55-year Reanalysis-based surface dataset for driving ocean–sea ice models (JRA55-do; Tsujino et al. 2018); ice velocity is from the Polar Pathfinder Daily Sea Ice Motion Vectors, version 4 (Tschudi et al. 2019); ice concentration is from the Nimbus-7 SMMR and DMSP SSM/I–SSMIS Passive Microwave Data (Cavalieri et al. 1996). We use present-day wind rather than previous-day wind as a predictor due to the spurious nature of atmospheric wind speeds and because the canonical linear relationship between ice and wind speed (Thorndike and Colony 1982) is based on present-day interactions. Additionally, present-day wind is readily obtained by observation, while sea ice properties are provided from satellites at a 1-day lag. We refer to Hoffman et al. (2023) for further information regarding the choice of using these particular datasets.

We apply the following processing to the data before using them to train the models:

- Daily values: The satellite products for ice motion and concentration are provided on a daily basis, but for wind velocity from JRA55-do, we calculate a daily mean from 3-hourly wind velocity vectors.
- Regrid to 25-km Equal-Area Scalable Earth (EASE) Grid: We use the nearest-neighbor approach to regrid wind and ice concentration to the 25-km EASE grid of the Polar Pathfinder Ice Motion product for consistency. The EASE grid provides information at latitude and longitude coordinates in a 361 × 361 grid. Taking the form of the Lambert azimuthal equal-area projections in polar regions, the EASE grid was defined by the NOAA–NASA Polar Pathfinder Program to accurately represent area and support spatial comparisons from the satellite data (Brodzik et al. 2012).
- Imputation to fill in the North Pole hole in sea ice concentration: The Nimbus-7 sea ice concentration dataset is missing data in a circular sector centered over the North Pole as a result of the orbit inclination of the satellite. This dataset is generated from brightness temperature measurements from the SMMR, SSM/Is, and SSMIS sensors, which have pole holes from latitudes of 84.5°, 87.2°, and 89.18°N, respectively (Cavalieri et al. 1996). In the analysis in Hoffman et al. (2023), these grid locations were simply filled in with zeros before implementation in model training. For XAI studies here, we have filled in this polar hole using imputation methods, replacing locations with "Not a Number" (NaN) values with the mean of the 40 nearest neighboring grid points. We opted to apply imputation here because while the polar hole did not significantly impact the overall performance of the model [which was the focus of Hoffman et al. (2023)], we want to ensure that it does not play a role in the XAI explanations as we know that the CNN exploits nonlocal information in making its predictions.

- Remove the seasonal cycle from each parameter: We are interested in the response of ice motion to the inputs on daily time scales without the impact of long-term variability. The seasonal cycle explains a much larger portion of the variance in wind, ice speed, and ice concentration (7.4%, 11%, and 63%, respectively) in comparison to the long-term linear trend, which explains 0.07%, 0.8%, and 2.3% for wind, ice speed, and ice concentration, respectively (Fig. S1 in the online supplemental material). Therefore, we only subtract the daily climatology and not the long-term linear trends for each parameter. For the case that we do not remove the seasonal cycle, the large amount of variance explained by the seasonal cycle in sea ice concentration increases the relevance of sea ice concentration as a predictor of sea ice motion in LRP explanations (not shown).
- Feature scaling: Feature scaling is an important part of preprocessing data as it attempts to give all of the attributes equal weight by expressing them in the same units and within a common range (García et al. 2015). We standardize all data to zero mean and one standard deviation, based on gridwise statistics of the training data. For each input feature, we remove the mean at each grid location and divide by the standard deviation calculated over all grid locations. The mean is removed by grid point, but the standard deviation is calculated across the domain to preserve the variance structure of the input features.
- Replace "NaN" values with zero: Sea ice velocity and concentration are set to zero rather than NaN in regions without ice because the CNN requires inputs to be numerical values.

3. Methods

a. Model setup

The model frameworks are taken from Hoffman et al. (2023) and further described here. Results from a manual k-fold cross-validation show that there is a low coefficient of variation in the performance (correlation and skill) between fifteen different LR and CNN model runs, where the years used in the train (28 years), validation (2 years), and test (2 years) sets are varied (Table S2). Therefore, we move forward in our analyses using only one of these model runs.

1) LR

The LR is expressed as Eq. (1):

$$\mathbf{u}^*_{i,t} = A\mathbf{u}^*_{a,t} + B\mathbf{u}^*_{i,t-1} + C\mathbf{c}^*_{i,t-1} + D. \quad (1)$$

Here, the inputs and coefficients (i.e., $\mathbf{u}^*_{a,t}$ and A, etc.) are the complex numbers, with the real and imaginary parts representing the zonal and meridional components and their respective parameters. The LR is applied gridwise so that each grid location has a unique set of LR coefficients to predict sea ice velocity at a particular grid point. Each grid point has three inputs (wind velocity, sea ice velocity, and sea ice concentration) and one output (sea ice velocity). The prediction (output) occurs at some other location [i.e., the input features at each (X, Y) predict the output at a particular location (A, B)]. We run the LR at a number of different analysis locations (A, B), each of which provides a heatmap covering the entire spatial domain showing a map of the variance explained by each input feature for predicting sea ice velocity at the given (A, B) location. This is in contrast to how the LR was applied previously in Hoffman et al. (2023), where each location was used to predict itself. The method used in this study allows us to take into account nonlocal linear interactions when investigating relevance. We apply a time-variable mask that only uses grid points and times where the ice concentration is greater than zero. We apply ridge regression with a ridge parameter of $\lambda = 10^{-2}$ to prevent unrealistically large LR parameters (Marquardt and Snee 1975).

2) CNN

The CNN architecture is illustrated in Fig. S2. The model is set up with five repeating units of Conv2D-ReLU-MaxPool, where the hyperparameters (i.e., number of filters and filter size) for each layer are shown in Table S1. These are followed by a 20% dropout, a flattening, and a dense layer that applies a regression to predict the output at each grid point. Inputs and outputs are in the form of spatial maps; we refer to each input map from each predictor as an "input feature." We emphasize that this CNN makes a regression (rather than classification) prediction of ice motion at each grid location.

The CNN is implemented in Python using the TensorFlow/Keras library (Abadi et al. 2015). The model is trained to optimize the loss function (normalized root-mean-square error) using an Adam optimizer and L2 regularization (Table S1). Data are split into train, validation, and test sets with an 88%–6%–6% split (28, 2, and 2 years of daily data). We run the model over 50 epochs with a batch size of 365; the number of epochs refers to the number of times the model will work through the entire training dataset, while the batch size represents the number of samples the model works through before updating the internal model parameters (i.e., weights and biases).

b. Explainability methods

In this study, we explore the level to which explanations provided by the local LRP explainability method can be aggregated to produce global explanations. We obtain global explanations by taking the average of local LRP explanations over the domain of interest (the Arctic) (Murdoch et al. 2019; Molnar 2020; Wilming et al. 2022). The global LRP is useful for our study because our ML model is built to make predictions of sea ice velocity at every grid location throughout the Arctic, and local LRP can only provide explanations for the model prediction at individual grid points. Global explanations will enhance our exploration of how ML models use each of the input features (i.e., maps of present-day wind velocity u_a; previous-day ice velocity u_i; and previous-day ice concentration c_i) to predict the output (i.e., maps of present-day ice velocity u_i).

To our knowledge, this study is the first application of a global implementation of LRP to a regression problem in

TABLE 1. Details about the various explainability methods applied to the ML models in this study based on the classifications discussed in section 3b and the appendix. The asterisk (*) refers to XAI methods not applied in this study; asterisked rows in the CPU time column refer to estimates made for the methods not used in this study: for the PERT, global case, we multiplied the computational cost for the local PERT by the number of points included in the global LRP analysis (i.e., 219 points); the PFI, local is assumed to have the same computational cost as a local PERT, as they involve the same steps.

Method	Model applied to	Methodology (linear, PERT, or propagation)	Type of explanation (sensitivity or salience)	Importance or relevance	Model awareness (specific or agnostic)	CPU time, OM (s)
(i) Variance explained by inputs in LR:	LR	Linear	Salience	Relevance	Model specific	—
LR, gridwise, local						2
LR, gridwise, global						5
(ii) PERT:	CNN	PERT based	Sensitivity	Relevance	Model agnostic	—
PERT, local						4
PERT, global*						6*
(iii) PFI:	CNN	PERT based	Sensitivity	Importance	Model agnostic	—
PFI, local*						4*
PFI, global						4
(iv) LRP:	CNN	Propagation based	Salience	Relevance	Model specific	—
LRP, local						5
LRP, global						7

geosciences. Therefore, we compare it to other XAI methods to provide context. Because our focus is to evaluate the trustworthiness and utility of a particular method (global LRP) to represent the physical drivers of ice motion, we choose a suite of XAI methods that provide us with similar comparisons for moving from local to global explanations. For each scope (local and global), we investigate explanations from a linear, perturbation (PERT), and propagation method. We stick to one of each of these categories due to the intensive nature of applying global explanations on this scale. However, we note that several XAI methods exist that are not applied in this study. Mamalakis et al. (2022), Flora et al. (2024), and Bommer et al. (2024) discuss an array of those that have been successfully implemented for geoscience problems, including but not limited to deep Shapley additive explanations (Lundberg and Lee 2017) and integrated gradients (Sundararajan et al. 2017). We choose to leave these methods out of our analysis because Han et al. (2022) show that Shapley methods pose challenges for continuous data, and both are vulnerable to producing noisy explanations (Mamalakis et al. 2022).

For this study, localized methods include analyzing the variance explained by each parameter in LR, perturbation analyses with the CNN, and gridwise-LRP applied to the CNN. Global methods include the analysis of the variance explained by each parameter in LR, permutation feature importance (PFI) applied to the CNN, and a global implementation of the LRP applied to the CNN. Each of these methods is summarized in Table 1 and Fig. 1 and further described below. We refer the reader to the appendix for further clarification about the classification of XAI methods. Locations used in local (red points) and global analyses (yellow points) are indicated in Fig. 2. The performance of the CNN is shown to be high throughout most regions of the Arctic (Fig. 2; correlation between prediction and observations is above 0.4 for all points and above 0.7 for most points), which justifies our use of explainability methods to understand skillful model predictions.

We compare outputs from different XAI methods applied to a CNN and an LR case for a robust analysis and to bolster confidence in the understanding of our models. Results from Hoffman et al. (2023) show that LR is comparable in skill to the CNN (correlation of 0.78 ± 0.02 for LR and 0.81 ± 0.02 for the CNN); therefore, we use LR as a baseline for comparison with outputs from various XAI methods applied to the CNN, as done by Toms et al. (2020). We run XAI analyses on the training data based on suggestions from Lakshmanan et al. (2015) and Flora et al. (2024), who argue that using out-of-sample data (i.e., the testing data) could expose extrapolation errors that might not be useful for understanding the model behavior. We note that we found only slight differences in the explanations when applying XAI to the testing rather than training data, where explanations were more localized for the training case (not shown).

1) VARIANCE EXPLAINED BY INPUTS IN LINEAR REGRESSION

In contrast to the CNN, LR is an interpretable model and can be understood through the analysis of the model parameters without the application of post hoc XAI methods. The linear regression parameters identify locations where each of the input features is relevant for predicting sea ice motion. However, the inputs are not orthogonal; wind and ice velocity in particular are highly correlated (not shown). Therefore, we use the fraction of variance explained by each of the inputs in LR as a metric for relevance, rather than the LR parameter itself. We calculate variance explained for each input from Eq. (2):

$$R^2 = 1 - (\mathbf{x} - \mathbf{E}\mathbf{m}_i)^{\mathrm{T}}(\mathbf{x} - \mathbf{E}\mathbf{m}_i)(\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}, \tag{2}$$

where $\mathbf{x}$ is the data, $\mathbf{E}$ is the LR kernel, $\mathbf{m}$ is the LR model, and $\mathbf{E}\mathbf{m}_i$ is the LR model prediction with the two parameters

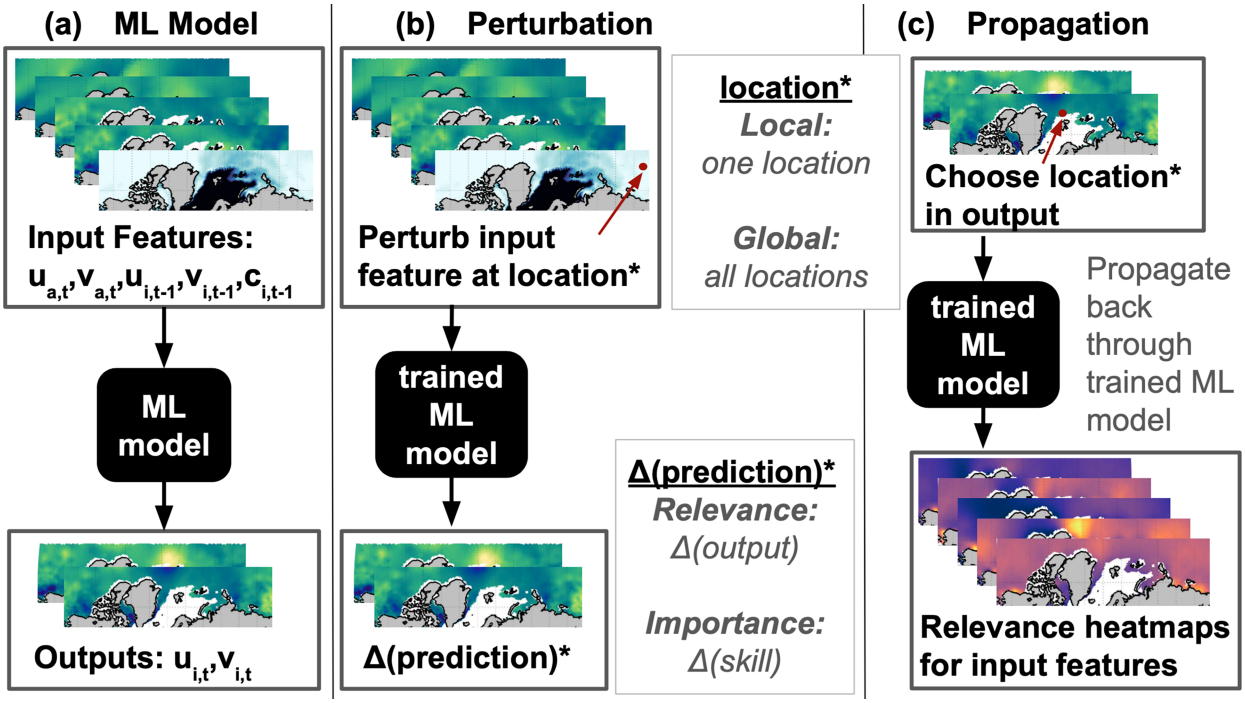

FIG. 1. Schematic of the (a) CNN applied in this study for predicting present-day sea ice velocity components (outputs) from present-day wind velocity, previous-day sea ice velocity, and previous-day sea ice concentration (inputs). (b),(c) XAI methods applied to understand model predictions. (b) PERT methods measure sensitivity and analyze how the model prediction [either the output itself (relevance), or the skill of the model (importance)] changes in response to PERTs of an input feature at specific grid locations (a single grid point for local; all grid points for global). (c) Propagation methods measure salience and analyze the relevance of every grid location in each input feature for making a prediction at a specific output grid location. Local studies show relevance heatmaps for the input features in predicting a particular grid point in the output, while global studies show relevance heatmaps for the input features in predicting the entire spatial domain. Maps shown here cover the spatial region north of 60°N latitude.

that are not being analyzed set to zero. In comparison to attribution studies using the CNN, LR only shows information about linear relationships. We run the LR analysis for both the local and global cases. For the local case, there are 17 different sets of gridwise LR models to predict sea ice velocity at each of the 17 analysis locations from the inputs at every other grid point in the Arctic. For each analysis location (A, B) (red points in Fig. 2), there is a separate LR model at each grid point, (X, Y), where the inputs are the aforementioned features at that grid point and the output is the sea ice velocity at the analysis location (A, B). The local case allows us to investigate how the sea ice velocity at each of the 17 analysis locations is dependent on the input features at every other grid location throughout the Arctic. For the global case, we apply the local method at 219 analysis locations throughout the Arctic (yellow points in Fig. 2). Each of the 219 analysis locations produces an individual variance explained heatmap for each input feature, showing the relevance of the entire spatial domain for making a prediction at that particular location. We average the heatmaps produced by each of these locations to obtain the global variance explained heatmaps. The global case allows us to investigate how the sea ice velocity throughout the Arctic is dependent on the input features at every other grid location throughout the Arctic.

We note that because LR is an interpretable model, there are nuances in classifying this method within the scope of local or global explainability. Because there is a different LR model at each grid location, each model itself is fully described (global) by the LR parameters at that grid point. However, we refer to local and global explainability for the two LR cases as described above in this study. The analysis of the variance explained by each input feature in LR allows comparison with the explainability studies applied to the CNN. We use this method as a baseline for comparison with explainability methods applied to a more complex model (i.e., a neural network).

2) PERTURBATION ANALYSIS

Perturbation analyses provide information about the sensitivity of an ML model prediction to a perturbed element (McGovern et al. 2019; Ivanovs et al. 2021). The relevance of the perturbed element is quantified from the magnitude of the change to the prediction. We are interested in understanding the relative relevance of each of the input predictors at various geographical locations. Thus, we apply perturbation analyses for each input feature at 17 different locations throughout the Arctic. We follow a procedure similar to that of Sinha and Abernathey (2021). After the CNN has been

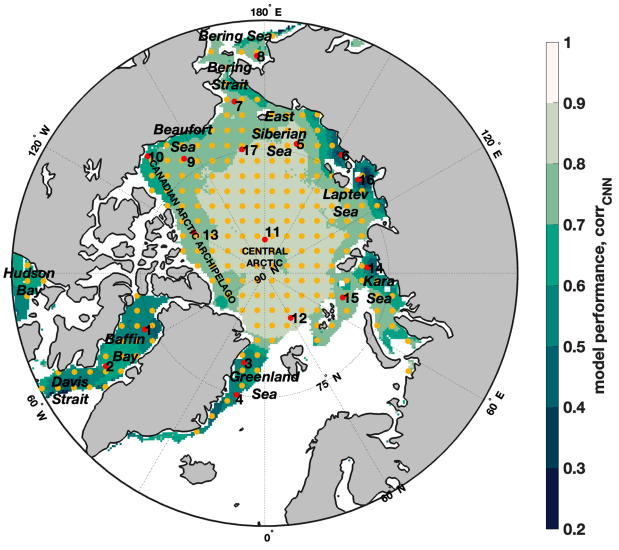

FIG. 2. Locations used in analysis for local (red points; labeled 1–17) and global LRP implementations of XAI (yellow points). Regions of the Arctic are labeled for reference during discussion. Map colors show the performance of the CNN in terms of the correlation [Eq. (4)] between the model prediction and observations.

trained, we apply the model to the training dataset (1979–2017) to make a prediction (the control). Next, a new dataset is made for each of the five input features and each of the 17 hand-picked locations, for a total of $17 \times 5 = 85$ perturbation runs. Here, the input feature is perturbed at the chosen location while keeping the rest of the variables and locations fixed. For each perturbation, we add a fraction of the standard deviation ($+0.5\sigma$) of the input feature. We also tested using $+1\sigma$ perturbations to the inputs. This impacted the magnitude of the change in the output, but not the comparative relevance of each of the input features, which is our main interest here. The model is run on each perturbed test dataset to make a prediction. For each perturbation, the root-mean-square difference (RMSD) is calculated over the temporal domain from Eq. (3):

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum_{i}^{n}(\hat{y}_i - y_i)^2}, \tag{3}$$

where n is the number of time steps, $\hat{y}_i$ is the prediction made using the input with a perturbation applied at a specific grid point, and y_i is the control model prediction using the nonperturbed input. The output is a map of the spatial extent of the effect of the perturbation for each input feature and grid location analyzed.

3) PFI

Another way to evaluate the sensitivity of the ML model to the various input features is through PFI methods (Breiman 2001; Radivojac et al. 2004). Similar to perturbation, in PFI, the relative importance of the input features is determined by the extent to which the ML model predictions are impacted by changing elements of the input features. To apply PFI, we randomly shuffle the values of each input feature between examples (i.e., time steps) for the training dataset (i.e., 1979–2017). We then use the shuffled data to make a prediction with a trained CNN. The performance (correlation) of the CNN is calculated and compared to that of a control case without randomization at each grid location:

$$\mathrm{corr}_{x,y} = \frac{\sum_{i}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i}^{n}(y_i - \overline{y})^2}}. \tag{4}$$

Here, n refers to the number of time steps over which the metric is evaluated, the overbar represents the time average, x represents the observations, and y represents the model output. We look at the difference in the performance between each permuted case and a control case. The most important feature is the one for which permutation results in the largest loss in performance in comparison to the control run. From this analysis, we can determine the input feature with the overall highest importance, in addition to spatial locations that are identified as important by the CNN. We also perform PFI by shuffling the spatial values of each input (not shown) and find there is little difference in the result compared to the case where the input features are shuffled between examples. This PFI method has been applied to ML predictions made in geosciences in the form of classification problems involving sea ice (Shen et al. 2017) and weather event detection (Molina et al. 2021; Lakshmanan et al. 2015), as well as for regression problems for predicting motions of wave gliders (Amador et al. 2021).

4) LRP

We use the LRP method to trace the explanation of CNN predictions (Bach et al. 2015; Montavon et al. 2015, 2019, 2018). LRP provides information on the relevance of individual input features by progressively redistributing the activation score of the output layer during a backward propagation through the neural network to the first layer after the model weights and biases have been frozen. The output from LRP is a heatmap showing the relevance of each input feature at each mapped location (latitude and longitude). Each grid location in the output produces an individual LRP heatmap for each input feature showing the relevance of the entire spatial domain for making a prediction at that particular grid location. For more details, Toms et al. (2020) and Ebert-Uphoff and Hilburn (2020) provide excellent overviews of LRP for other geoscience applications.

We proceed with the analysis of LRP in two different ways: local and global. For the local case, we pick 17 different latitude–longitude locations in the Arctic (red points in Fig. 2) and analyze the relevance heatmaps produced from each of these locations in comparison to other explainability methods that are confined to one grid point (i.e., maps of variance explained by LR and perturbation). For global analyses, we

create integrated heatmaps by running LRP for 219 selected points throughout the Arctic (yellow points in Fig. 2) and averaging the relevance maps produced by each of these locations at each time step. This allows LRP to be compared to more spatially comprehensive explainability methods, such as the variance explained by LR parameters or PFI outputs. We note that a fully global implementation would integrate the LRP heatmaps produced by every output grid point in the Arctic, not just the selected 219 points. However, we stick to this subset of points for the sake of computational efficiency.

As with the other XAI methods, we run the LRP analysis on the training data. We average the LRP relevance heatmaps over the temporal domain. We use the iNNvestigate package (Alber et al. 2019) with the "sequential preset A" configuration of LRP that applies different rules at different layers of the model, where LRP-$\alpha_1\beta_0$ is applied for convolutional layers and LRP-ϵ is applied for dropout, flatten, and dense layers. This configuration is similar to the LRPcomp that was demonstrated to be a good explainability method for geoscience applications in Mamalakis et al. (2022). The LRPcomp of Mamalakis et al. (2022) uses LRP-z where we apply the LRP-ϵ rule. In comparison to LRP-z, LRP-ϵ includes a parameter that leads to less noisy explanations (Montavon et al. 2019). We refer to Montavon et al. (2019) for more information about the different LRP rules.

The LRP relevance scores can be positive or negative. For the case of a classification problem, the sign of the relevance shows which points contributed positively or negatively to a correct prediction. However, here, we apply LRP to a regression problem and the sign of the relevance indicates whether the input at a particular location contributed positively or negatively to the prediction itself, i.e., which regions in each input made the prediction of sea ice velocity different (larger magnitude relevance for larger contributions to velocity changes, regardless of sign) than zero. Therefore, we take the absolute value as we are looking for the magnitude of the contribution, rather than the directionality. We tested whether a large negative relevance score was indicative of a highly relevant point by integrating relevance maps calculated for points within given regions and found that both large negative and large positive relevance scores were indicative of highly relevant points (not shown). Additionally, Mamalakis et al. (2022) have identified that LRP automatically assigns zero relevance to zero values in the input features due to input multiplication at the final step of LRP. This is less important for the input features for sea ice velocity and concentration in our study, where a zero value represents regions where there is no sea ice and it would not make sense to distribute relevance to these points. However, there could be places where the standardized wind velocity input is zero and subject to this ignorant-to-zero-input issue.

While LRP methods were developed for classification rather than regression (Bach et al. 2015; Montavon et al. 2019), there have been examples of LRP being applied for regression in other fields (Dobrescu et al. 2019; Rahman et al. 2021; Schnake et al. 2020) and within geosciences (Ebert-Uphoff and Hilburn 2020; Hilburn et al. 2021). Additionally, Letzgus et al. (2021) and Mamalakis et al. (2023) discuss methods to extend LRP to regression problems. These methods involve retraining the CNN with respect to a carefully chosen reference value (i.e., subtracting the reference value from the outputs before training). Here, the question that is being asked by XAI is "where are the input features relevant in predicting the output to be different from the baseline reference value?" The choice of the baseline reference value changes the scientific question being asked by XAI (Letzgus et al. 2021; Mamalakis et al. 2023).

In this study, we use XAI to evaluate a CNN trained to make regressive predictions of sea ice motion at each grid location throughout the Arctic. We reiterate that the model is trained on data from which the seasonal cycle has been removed and which is standardized to have zero mean and one standard deviation. We move forward using the default baseline reference value of zero for the method's analysis of this study. This is synonymous with choosing a reference value as the mean. We will apply a carefully chosen reference value in future work when the aim is to use XAI to answer a scientific question about the underlying physics of ice motion.

c. *Localization*

One of the benefits of using a CNN (in contrast to a nonconvolutional neural network) is its ability to incorporate information from neighboring pixels into predictions by performing convolutions over multiple grid points. For a traditional CNN, the spatial extent of the influence of the input features on a prediction at a particular grid location is known as the receptive field, which can be extended by adjusting the network architecture (i.e., adding more convolutional layers, changing filter sizes, etc.) (Araujo et al. 2019). In addition to convolutional layers, our CNN includes a fully connected layer (Table S1), which extends the long-range dependencies to give the model a global capacity for feature interactions (Ding et al. 2021). The nonlocality of the CNN will be incorporated into explainability. We analyze the radius of influence of the explanations from different local XAI methods to assess the degree to which nonlocal points are incorporated into the model predictions.

d. *Decomposition into linear and nonlinear responses using perturbations*

Perturbation studies are also used to decompose the model into its linear or nonlinear responses based on techniques described by Verdy et al. (2014) and Swierczek et al. (2021). In this case, we run perturbation experiments where we either add or subtract one standard deviation from a particular grid point in each of the input features (i.e., $\pm 1\sigma$). The differences between the model response to positive h_+ and negative h_- anomalies and the control run h_0 are given as $\delta h_1 = (h_+ - h_0)$ and $\delta h_2 = (h_- - h_0)$, respectively. We decompose the perturbed runs as Taylor expansions around the control run and separate the odd-order and even-order terms into $\delta H_1 = (1/2)(\delta h_1 - \delta h_2)$, and $\delta H_2 = (1/2)(\delta h_1 + \delta h_2)$, which, assuming third-order and higher terms are negligible, are representative of the linear and nonlinear responses, respectively. Under the condition that δH_2 is roughly an order of magnitude smaller than δH_1, we can assume

the model is primarily linear and that δH_1 is representative of this linear response. When the magnitude of δH_2 approaches δH_1 (i.e., positive and negative perturbation responses no longer cancel out), the model response is assumed to be nonlinear, and the analysis of the higher-order terms in the Taylor series becomes more important.

We calculate these terms for each point in time and space at the 17 perturbation locations. We calculate the root-mean-square (RMS) of δH_1 and δH_2 and the fraction of the response that is linear, $\delta H_1/\sqrt{(1/2)(\delta H_1^2 + \delta H_2^2)}$, for (i) spatial and (ii) temporal domains from Eq. (5),

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i}^{n} h^2}, \tag{5}$$

where h is the term of interest (either δH_1, δH_2, or the fraction of the response that is linear) and n is the number of data points. Spatial analyses use the RMS taken over the temporal domain (n is the number of time steps) and provide maps that show which parts of the Arctic have a linear or nonlinear response to a perturbation at a given location. For temporal analyses, we take the RMS over space (n is the number of spatial coordinates) and follow the RMS calculation by computing the monthly mean of each of the three metrics and comparing them to the monthly mean sea ice concentration at each perturbation location to understand how the model response is linked to the sea ice state.

4. Results

a. *Local attribution studies*

We use local XAI methods to understand the relevance or importance of each of the input features for making a prediction at a specific grid location. Our focus is on determining the spatial structure of the feature attribution in order to determine how much information is extracted from nonlocal inputs by each method: variance explained by local LR, perturbation analysis, and localized LRP. These local XAI methods produce a separate relevance heatmap for each specific grid location and time step to which they are applied, indicating either (i) the degree to which the input at a specific location impacts the prediction of the output throughout the Arctic (perturbation) or (ii) the degree to which the entire map of each input was relevant in predicting the output at a specific location (variance explained by LR and LRP). We run each of these analyses for 17 different locations (red points in Fig. 2) and average over all time steps. We identify the spatial extent of the relevance of each of the input features (e.g., wind velocity) in predicting the output (i.e., sea ice velocity) for location 11 (Fig. 3). The spatial mean and standard deviation are indicated in the legend. We show the mean to compare the overall relevance attributed to each input feature for each method; therefore, these statistics are calculated over the entire spatial domain shown (i.e., north of 60°N). An extensive look at this set of maps for each location can be found in Figs. S3–S19.

The outputs from XAI perturbation and propagation-based studies are normalized by dividing the temporal mean by the value at the 99.5th percentile for each particular method to create similar magnitudes for visual comparison among the different methods. We normalize to the 99.5th percentile rather than the overall maximum because of the risk of the absolute maximum being an outlier. Because the normalization is based on overall magnitude instead of a particular spatial location, the location of the normalization value is not necessarily the same for each method. Thus, when we refer to relevance, we are referring to the XAI outputs, each normalized to the 0.5% maximum value for each of the perturbation and LRP methods, respectively. The relevance scores are not used to compare between the methods but instead are used to compare how each method assigns relevance to the various input features.

1) XAI EXPLANATIONS

The variance explained by each LR parameter decreases with increasing distance from the analysis point for wind velocity and ice velocity (Figs. 3a,b). The spatial mean of the variance explained by the various LR parameters is the highest for wind velocity, followed by ice velocity, and is much lower for sea ice concentration (legends in Figs. 3a–c). The variance explained by wind velocity also extends over land, indicating wind on land can be linked to sea ice velocity offshore. We show results for location 11. Figures in the supplemental information show that the same patterns of a decreasing variance explained with an increasing distance from the analysis point tend to hold for all 17 analysis points (Figs. S3–S19).

Perturbation analyses on the CNN show decreasing relevance with increasing distance from the analysis point for all predictors (Figs. 3d–f). Wind velocity has the highest relevance and largest radius of influence, followed by ice velocity and then ice concentration. The perturbation relevance score for wind velocity remains above 0.8 even for locations far from the analysis point (e.g., off the northeast coast of Greenland), while those for ice velocity and ice concentration drop below 0.2 at the same locations.

The LRP applied at the analysis location also shows high localized relevance that decreases with increasing distance from the analysis point for wind and sea ice velocity (Figs. 3g–i). Sea ice concentration shows a substantially smaller relevance (the mean in the legend is nonzero) but is not entirely irrelevant as in the LR case. Here, ice velocity shows the largest localized relevance in predicting next-day ice velocity, followed by wind velocity and ice concentration. Conversely, wind velocity is found to have the highest spatial mean relevance. While these regions may not represent regions with the highest relevance, the relevance of wind over land is consistent with the variance explained by LR in many cases. Generally, areas of high relevance tend to extend coherently within a certain radius of the analysis point for all methods.

2) LOCALIZATION

We also analyze the extent to which the relevance values vary with distance from the analysis location for all locations and each of the attribution methods and input features (Fig. 4). Here, we only show data within 2000 km of the

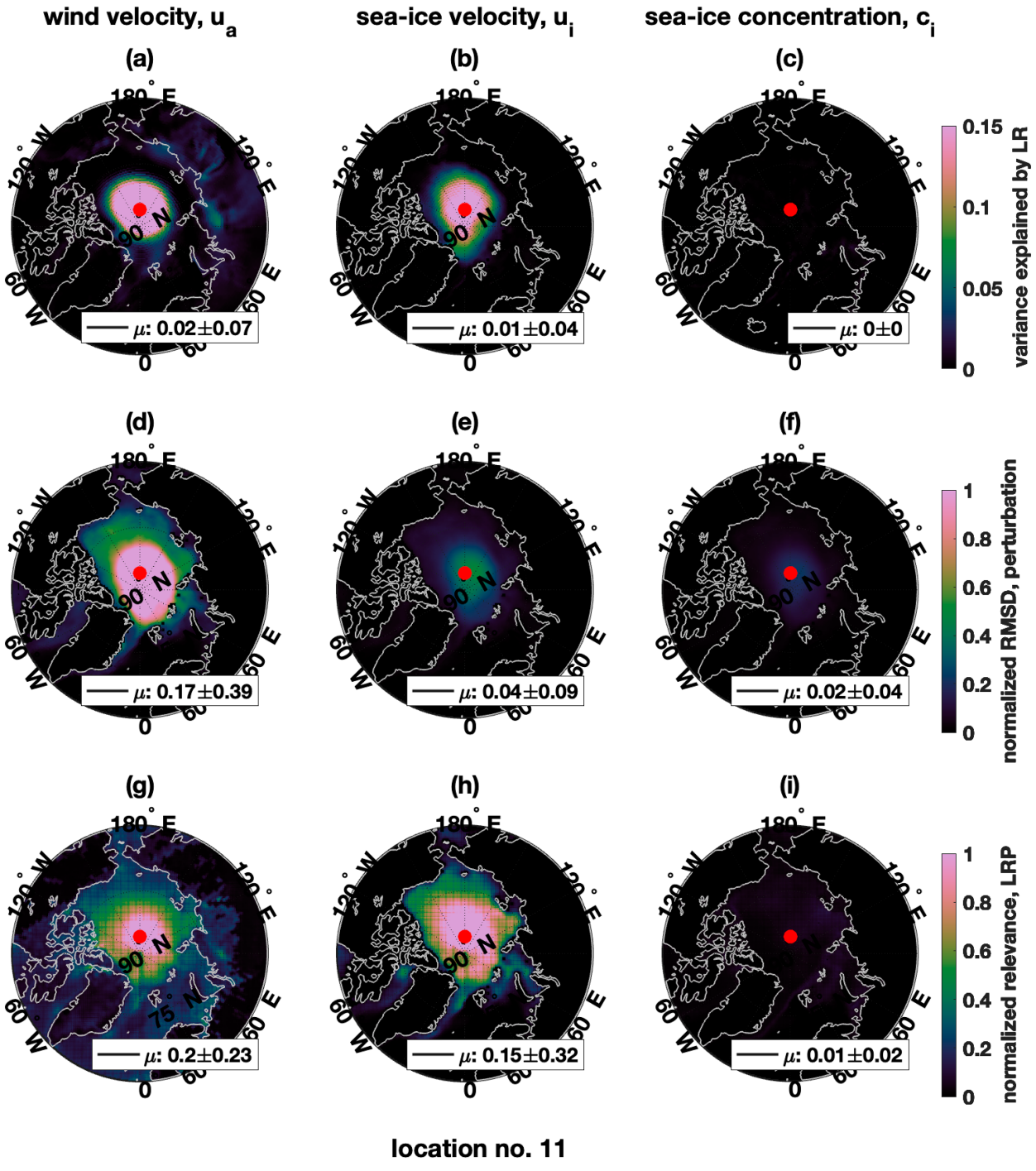

FIG. 3. Results from localized XAI studies for each of the input features at location 11, indicated by the red dot. These relevance heatmaps show the spatial extent to which each input feature is relevant in predicting the outputs at the location indicated by the red dot. The columns represent each of the different input features: (a),(d),(g) wind velocity u_a; (b),(e),(h) ice velocity u_i; and (c),(f),(i) sea ice concentration c_i. The rows represent the different sensitivity methods: (a)–(c) variance explained by LR, (d)–(f) normalized RMSD from PERT analysis, and (g)–(i) normalized relevance score from LRP. The bottom two rows are normalized by dividing by the top 0.5% relevance value for each method. The spatial mean and standard deviation are indicated in the legend; these statistics are calculated over the entire spatial domain shown.

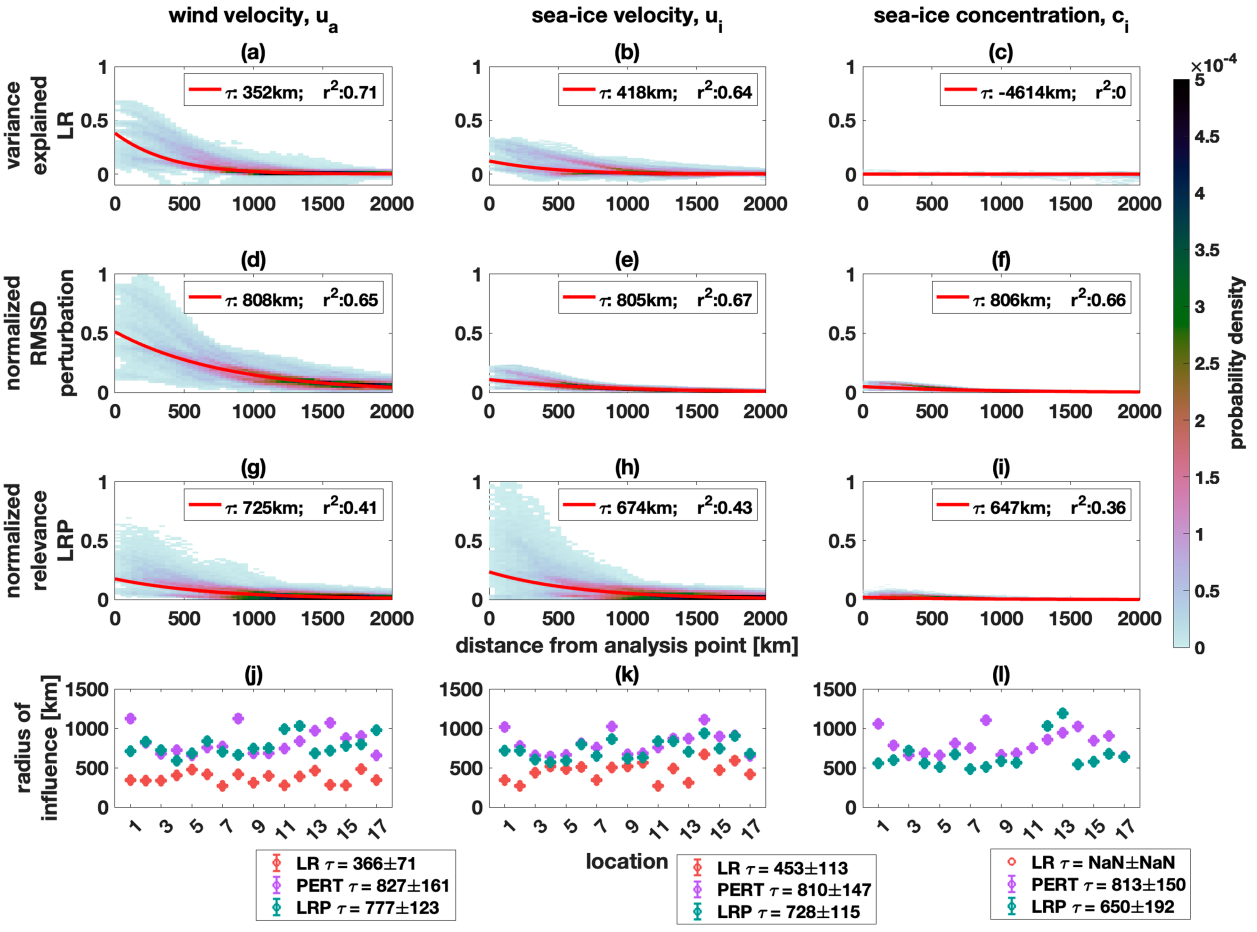

FIG. 4. (a)–(i) Probability density of the relevance of each localized sensitivity study as a function of the distance from the point of analysis for each of the input features and for all locations. The columns represent each of the different input features: (a),(d),(g) wind velocity u_a; (b),(e),(h) ice velocity u_i; and (c),(f),(i) sea ice concentration c_i. The rows represent the different sensitivity methods: (a)–(c) variance explained by LR, (d)–(f) normalized RMSD from PERT analysis, and (g)–(i) normalized relevance score from LRP. The middle two rows are normalized by dividing by the maximum relevance value for each method. The red lines represent exponential fits to the data. The legend gives the *e*-folding distance for that fit which gives a measure of the radius of influence for each of the relevance methods. The legend also shows the r^2 values for each fit; an $r^2 > 0.12$ is statistically significant with 95% confidence based on the degrees of freedom for each fit. (j)–(l) Mean and standard deviation radius of influence (i.e., *e*-folding distance) for each location (points), input feature (columns), and relevance method [colors: red for variance explained by LR, purple for PERT, and green for LRP]. Statistics are calculated using Monte Carlo methods. The mean and standard deviation over all locations for each input feature and method are shown in the legend below each panel. The outlier has been omitted for LRP for c_i at location 11 (33 704 km). The low r^2 value for the fit of the LR relevance of c_i in (c) indicates the inability of the fit to describe the *e*-folding distance for this particular case; thus, it is omitted from (l).

analysis point because our interest concerns the regions of high relevance. The red lines represent exponential fits, and the legend shows the r^2 value and the *e*-folding distance, which is a measure of the radius of influence of each method. Relevance decreases exponentially with increasing distance from the analysis point. This is true for almost all attribution methods and input features and makes sense because the spatial correlation of ice motion in the Arctic decreases exponentially with increasing distance from the point of interest (not shown). The exception is the LR variance explained by sea ice concentration (Fig. 4c), which exhibits a low r^2 for the exponential fit, indicating its statistical insignificance. We do not necessarily expect the relevance of each input feature to be modeled well by a Gaussian, but use this as a simplified representation of the radius of influence of each of the input features in making predictions of the output at various locations throughout the Arctic. While the r^2 values calculated over the distribution of the relevance for the combined 17 locations show fits that are statistically significant (legends in Figs. 4a–i), the fits calculated at each location separately (represented in Figs. 4j–l) do not all exhibit significant r^2 values, particularly for the case of ice concentration.

For the LR variance explained, the relevance score for wind velocity and ice velocity has similar radii of influence (Figs. 4a,b). Outputs from perturbation show that the predictors have largely different relevance scores but similar radii of influence

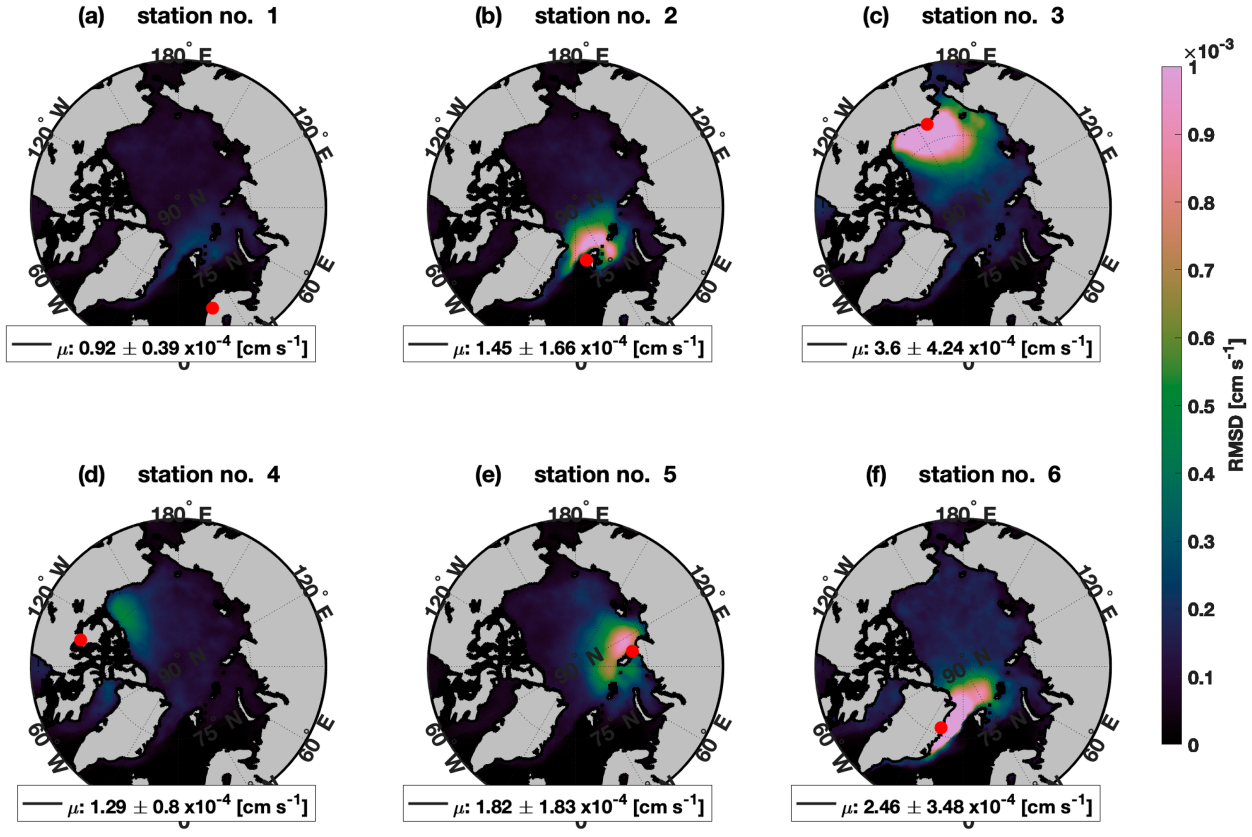

FIG. 5. Maps showing the RMSD of the response of the CNN prediction of sea ice velocity to PERTs of the input features at the locations of six different Arctic research stations. The PERT outputs have units of sea ice velocity (cm s^{-1}). The legend shows the mean and standard deviation of the RMSD response for locations where there is ice (i.e., not including land).

(Figs. 4d–f). Similar to results in Fig. 3, wind velocity has the highest relevance for perturbation, followed by ice velocity and ice concentration. Consistent with results in Fig. 3, sea ice velocity has the highest localized (i.e., close to the analysis point) relevance for the LRP method. The XAI methods applied to the CNN (perturbation and LRP) show larger radii of influence that are 1.6–2.3 times that of the variance explained by LR parameters (legends in Figs. 4a–i), which suggests that the far-reaching spatial information incorporated into the CNN predictions includes nonlinear interactions between locations and or input features.

Results in Figs. 4a–i show the distribution for all 17 locations. We also calculate the radius of influence for each location individually (Figs. 4j–l) and show the mean and standard deviation over all locations for each input feature and method (legend in Figs. 4j–l). The radius of influence is omitted for the LR variance explained by sea ice concentration (Fig. 4l) due to the insignificance of the fit (i.e., low r^2 in Fig. 4c). The mean radius of influence for the LRP and perturbation methods is similar and falls within one standard deviation of each other for each of the input features. The radius of influence for LRP and perturbation is also similar for each location (green and purple points in Figs. 4j–l). The LRP shows an outlier in the radius of influence for c_i at location 11. This value is omitted from Fig. 4l and from the calculation of the mean value in the legend and is likely a result of the low spatial variability in the relevance of ice concentration.

3) PERTURBATION OF WIND AT RESEARCH STATIONS ON LAND

The CNN incorporates nonlocal relationships to make predictions. Results from localized LR and LRP studies interestingly show that wind on land is relevant for predicting sea ice velocity at a location in the central Arctic (Fig. 3g). We run perturbation analyses at the location of six land-based research stations in the Arctic to understand the spatial extent to which they are relevant for making predictions of sea ice motion offshore. We show results for these perturbation studies with wind because it is the only nonzero input predictor over land (Fig. 5).

We find that information about the wind at land-based research stations is useful for making predictions of sea ice velocity at offshore locations. The spatial extent to which information about onshore wind is relevant for predicting offshore sea ice velocity depends on the location of the research station. The relevance does not extend as far offshore for stations that are further from the ice (stations 1 and 4 in Figs. 5a,d) in comparison to stations that are closer to the ice (stations 2, 3, 5, and 6 in Figs. 5b,c,e,f).

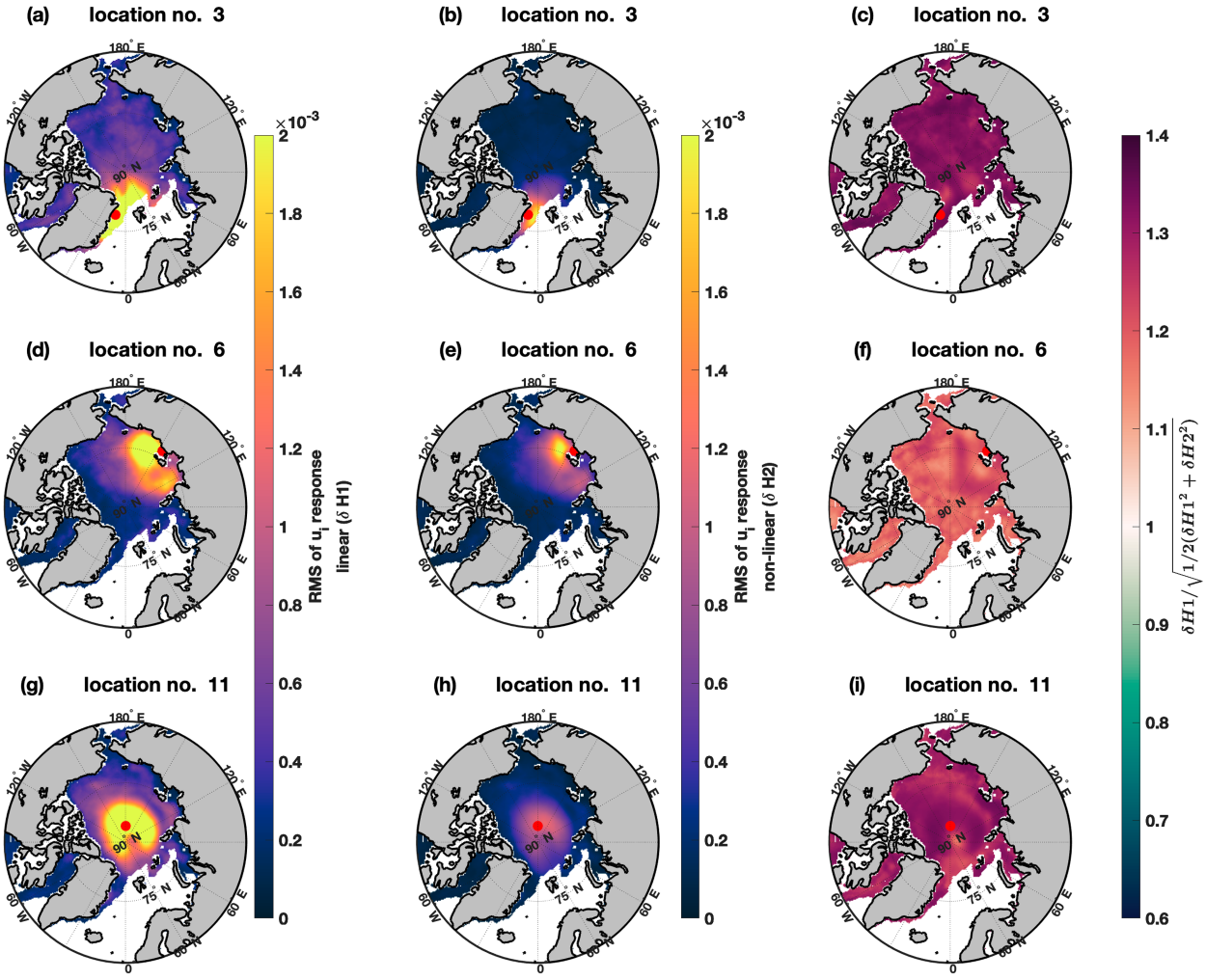

FIG. 6. Maps showing the RMS response of the CNN prediction of sea ice velocity to PERTs of the input predictors at three different locations [(a)–(c) location 3; (d)–(f) location 6; and (g)–(i) location 17]. The response is separated into (a),(d),(g) linear and (b),(e),(h) nonlinear components. (c),(f),(i) The ratio of the linear to nonlinear response. A ratio greater than one indicates stronger dominance of the model by a linear response. The red dots represent the location that was perturbed for each case.

4) DECOMPOSITION INTO LINEAR AND NONLINEAR RESPONSE COMPONENTS USING PERTURBATIONS

The response of the CNN prediction of sea ice velocity to localized perturbations in each of the input features is separated into linear δH_1 and nonlinear δH_2 components through the process discussed in section 3b(2). We note that these terms represent the odd and even-ordered terms in the Taylor series expansion and that the ability to refer to them as the linear and nonlinear components is based on the assumption that δH_1 is roughly an order of magnitude (OM) greater than δH_2 (i.e., the second-order and higher-order terms in the Taylor series are small). Maps of the ratio $\delta H_1/\delta H_2$ show little spatial variability and remain around 2–4 even for locations far away from the perturbation point (not shown). While this is not quite one OM difference, we move forward with the assumption and mention alternative approaches in the discussion. We analyze the fraction of the response that is linear $[\delta H_1/\sqrt{(1/2)(\delta H_1^2 + \delta H_2^2)}]$ to determine when and where the model is dominated by linear or nonlinear terms. When this ratio is greater than one, we can say the model is dominated by the linear term. When the ratio is equal to one, the linear and nonlinear terms are equal. Nonlinear terms play an important role when the ratio is less than one and as it approaches zero. We analyze the spatial and temporal RMS responses of the CNN to perturbations at 17 different locations throughout the Arctic (red points in Fig. 2).

Maps of the temporal RMS responses to perturbations at various locations (Fig. 6) indicate that linear terms dominate the CNN predictions throughout most of the Arctic. We can see that for all three locations shown (locations 3, 6, and 11), the linear portion of the RMS response is roughly 2–4 times larger than the nonlinear portion (Figs. 6a,d,g vs Figs. 6b,e,h). We define this as a "weakly linear" response because the linear terms are not a full OM greater than the nonlinear terms.

We note that the magnitude of the perturbation response is quite small; however, we are not so much concerned with the overall magnitude of the change as we are with the comparison between δH_1 and δH_2. We applied different amounts of noise to assess the magnitude of the nonlinear signal (i.e., also applied perturbations of $\pm 0.5\sigma$; not shown) and find that larger perturbations show that nonlinear terms are more important.

Both the linear and nonlinear responses are localized around the perturbation location (red dot in maps of Fig. 6). The fraction of the response that is dominated by the linear term depends on the analysis location (Figs. 6c,f,i). Analysis locations in the central Arctic typically exhibit a more linear response, while nonlinearity becomes more important for analysis locations near coastal regions and in the peripheral seas (i.e., location 6 vs 11 in Figs. 6f,i).

We also analyze the monthly mean of the spatial RMS response (i.e., the RMS is taken over space) to perturbations at 17 different locations throughout the Arctic and compare the linear and nonlinear components to the monthly mean sea ice concentration at each of these locations (Fig. 7). We show the fraction of the response that is linear compared to ice concentration (Fig. 7b). In this case, the model is described by the linear terms for $0.2 < c_i < 0.8$, i.e., the ratio is greater than one in Fig. 7b. The model tends to have a stronger dependence on the nonlinear term for extremes in sea ice concentration (i.e., $c_i < 0.2$ and $c_i > 0.8$ in Fig. 7b).

b. *Global attribution studies*

In the previous section, we analyzed the spatial extent of the relevance of each input feature for the models in making predictions at a particular location. In this section, we aim to understand the relevance of each input feature for making predictions over the entire spatial domain (in this case, the Arctic). We achieve this by integrating the attribution of the model's predictions to each input over all spatial locations. We compare explainability methods that analyze global attribution of each input in predicting ice motion: variance explained by LR parameters, PFI, and global LRP (Fig. 8). The composited explainability outputs are normalized by dividing by the top 0.5% maximum value to create similar scales for comparison among the three methods. However, we emphasize that we are not comparing attribution scores between the methods, but analyzing how each of the methods distributes attribution to the various inputs. As with the localized studies, normalization is based on magnitude and not a particular spatial location, and therefore, a different location is used to normalize across methods (not shown).

Maps in Fig. 8 show the spatial extent of the attribution of each of the input features (columns) to the ML models trained to predict sea ice motion for each XAI method (rows). The spatial mean and standard deviation of the relevance or importance are indicated in the legends. Similar to the local case, the statistics are calculated over the areas north of 60°N. We note that XAI highlights regions over land for the LRP and variance explained by LR parameters methods, but not for the PFI method. The PFI is unable to attribute importance to areas where there is no ice: there is technical nuance in calculating the correlation over land in that both x and y are zero in Eq. (4) where there is no sea ice, which leads to a "divide by zero" error, and thus a NaN value for the correlation in these areas. The LR and LRP are also only applied to locations in the Arctic where there is ice (see Fig. 2), but the nature of these methods allows relevance to be distributed to nonlocal points.

For each method, we find that wind velocity has the highest spatial mean relevance or importance, followed by ice velocity and then ice concentration (legends in Fig. 8). These results are largely consistent with a visual inspection of the heatmaps in Fig. 8. The spatial analysis of the variance explained by the LR parameters indicates that wind velocity has the highest relevance in predicting ice motion for the LR model. This is particularly true in the central Arctic, but relevance also remains high in coastal regions and over land. Previous-day ice velocity has the second largest LR variance explained in the central Arctic, but exhibits a low variance explained in most coastal regions (Figs. 8a–c).

The PFI method also shows that wind velocity is the most important predictor throughout the Arctic. However, the spatial variability in importance for the PFI is not consistent with the LR case. In contrast, for PFI, the importance of wind velocity is higher in coastal regions and peripheral seas (particularly the Laptev Sea, Kara Sea, north of Greenland, and Hudson Bay) and comparatively exhibits a lower importance in the central Arctic, which is the opposite of what is seen in LR. Spatial patterns in the importance of the ice velocity for the PFI analysis are also in contrast to those of LR: for PFI, the importance of ice velocity is higher for coastal regions than for the central Arctic (i.e., the Beaufort Sea, Bering Sea, and East Siberian Sea). Overall, wind has a higher importance relative to sea ice velocity for the PFI method than for LR.

In contrast to the other two XAI methods, wind is not consistently the most relevant predictor throughout the Arctic for LRP. The global LRP shows similarities to PFI for the spatial distribution of the relevance of wind velocity: Wind velocity is relevant throughout the Arctic, but some coastal regions (i.e., surrounding Greenland, Laptev Sea, Bering Strait, Beaufort Sea) have a slightly higher relevance than other regions. In contrast to PFI, the LRP relevance for previous-day sea ice velocity is high in the central Arctic and low in coastal regions and peripheral seas. In this regard, the relevance map of LRP is similar to that of LR. Also, similar to LR, LRP shows that wind is relevant throughout most of the Arctic, including over land. Interestingly, sea ice concentration shows a small relevance for the LRP at coastal and peripheral seas, where no relevance was shown for this input feature in other methods. We discuss potential mechanisms for the variability in the spatial structure of the attribution between these methods in the next section.

5. Discussion

This study aims to confirm the feasibility of using a novel XAI method (a global implementation of LRP) to understand predictions made by machine learning models built to predict

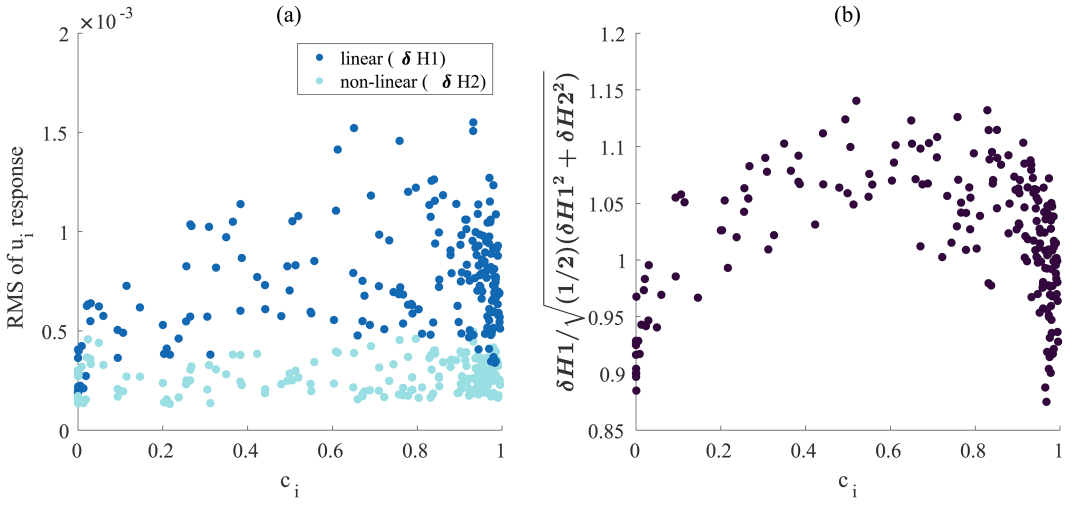

FIG. 7. (a) Monthly mean of the linear and nonlinear RMS response of the CNN prediction of sea ice velocity to a small PERT of the input features at 17 different locations vs monthly mean sea ice concentration at each location. (b) The ratio of the linear to the nonlinear RMS response vs monthly mean sea ice concentration at each location. A higher ratio indicates stronger dominance of the model by a linear response.

sea ice motion in the Arctic on 1-day time scales. To do this, we compare outputs from various XAI methods and analyze the degree of nonlocality and nonlinearity inherent in predictions from a CNN. Results from this study are used to answer the following questions.

a. What do the outputs from XAI methods show us about the contribution of the various input features for making one-day predictions of sea ice motion in the Arctic?

Generally, we find that wind velocity is the input feature that contributes the most to predictions of sea ice motion. This is confirmed in Fig. 9, which shows that the spatial mean value of the contribution of wind velocity (red points) is higher than that of the other input features (purple and blue points) for all local (Figs. 9a–c) and global (Figs. 9d–f) XAI methods. We note that the spatial mean is calculated over the entire domain shown in Fig. 8. Additionally, the spatial mean values are not to be used to compare between the methods, just to compare how the different methods distribute relevance or importance to the various input features. The fact that we generally find wind to be the most relevant predictor is consistent with historical results from Thorndike and Colony (1982), who found that local wind explained up to 70% of the variability in ice motion on short time scales.

Localization analyses also confirm consistencies between the local XAI methods. We show that relevance decreases exponentially with increasing distance from the analysis point for all methods and input features (Fig. 4). The radii of influence of the relevance for LRP and perturbation fall within one standard deviation of each other for each of the input features. Both the LR and CNN incorporate nonlocal information, while the CNN alone incorporates nonlinearities. Thus, the larger radius of influence for XAI methods applied to the CNN in comparison to LR suggests that this nonlocal relevance also includes information about nonlinear interactions.

The overlap of relevant regions between LR and XAI applied to the CNN suggests that XAI methods typically show high localization for regions where there is a known statistical relationship between the inputs and outputs. Regions of high relevance that exist outside of the region where there is high variance explained by LR could be the result of spurious relationships in the training data; they could also provide useful insight into relationships between the inputs and outputs that are not fully understood by linear statistics. Here, it is important to discern whether this nonlocal relevance is consistent with physics. For example, McNutt and Overland (2003) discuss that while wind forcing is important for sea ice dynamics at all scales, it is of particular importance at the coherent scale (75–300 km), and that coherent-scale motion provides nonlocal forcing to motion at the aggregate scale (10–75 km). We use sea ice motion on a 25-km grid, so it is feasible that both of these scales are represented in our study and that some of the nonlocal effects we are seeing in the XAI are indeed physical.

We move forward with the global implementation of LRP because local analyses confirm consistencies between the XAI methods and with what is known about the physics of sea ice motion (i.e., it is largely driven by wind). These consistencies are also confirmed in the overall mean of the global explanations, as discussed above.

Analyses of the spatial extent of the contributions of the input features for each of the XAI methods show varying consistency with historical results that sea ice motion can be largely explained by wind velocity in the central Arctic, but the relationship weakens in coastal regions (Thorndike and Colony 1982; Kimura and Wakatsuchi 2000; Kwok et al. 2013; Maeda et al. 2020). This is true for LR, but the PFI shows increased importance of wind velocity in coastal regions and LRP shows that the relevance of wind velocity is relatively uniform throughout the Arctic and is lower than that for sea ice velocity at most locations. Historically, decreases in the

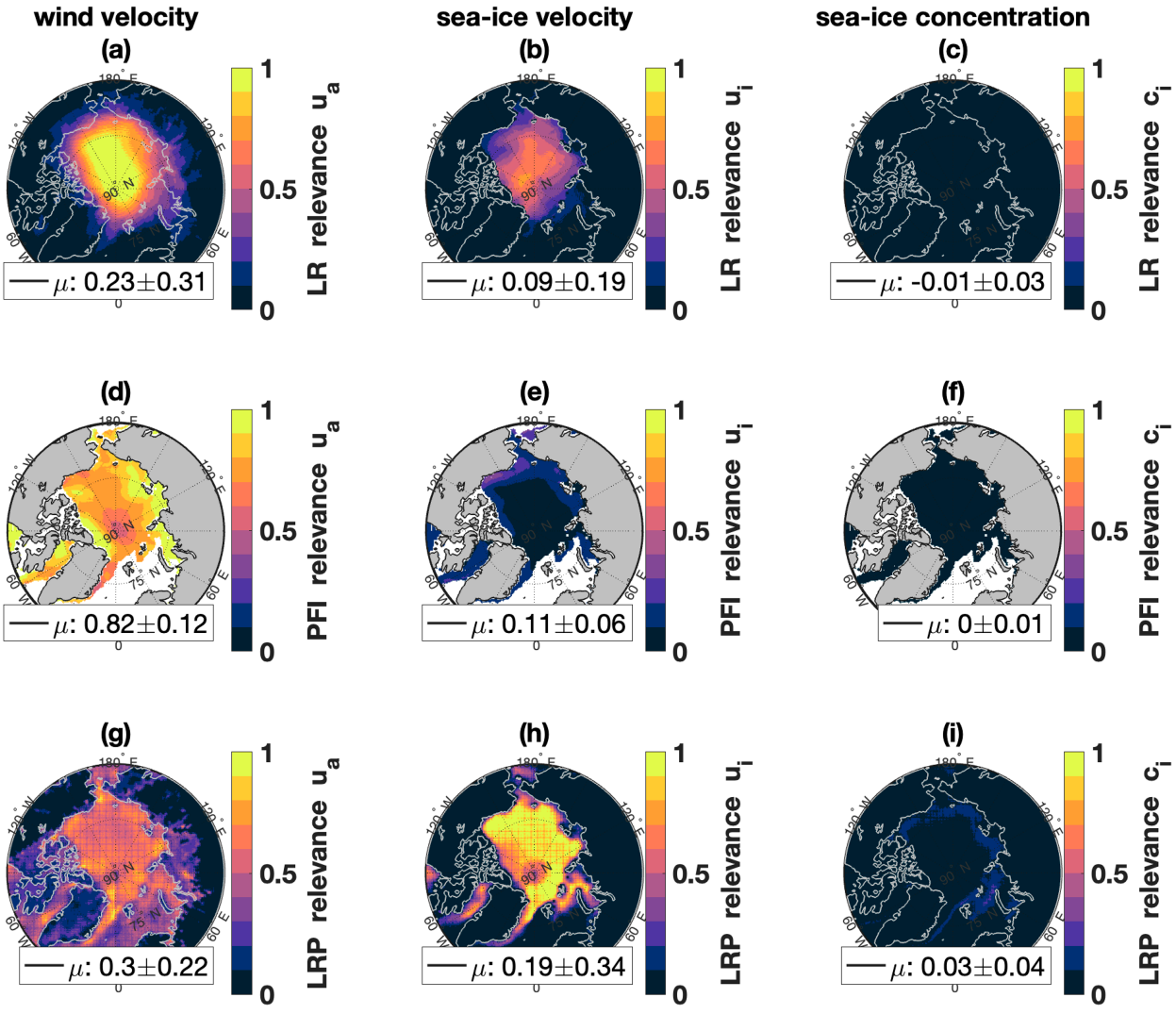

FIG. 8. Results from global sensitivity studies for each of the input features and methods. The columns represent each of the different input features: (a),(d),(g) wind velocity u_a; (b),(e),(h) ice velocity u_i; and (c),(f),(i) sea ice concentration c_i. The rows represent the different sensitivity methods: (a)–(c) variance explained by LR parameters; (d)–(f) PFI; and (g)–(i) LRP. We normalize by dividing by the top 0.5% maximum importance or relevance value of the spatial mean for each method. We note that PFI is only applied at locations where c_i is nonzero and therefore do not show relevance or importance over land. The spatial mean and standard deviation are indicated in the legend. The statistics are calculated over all areas north of 60°N.

linear relationship between ice motion and wind velocity near the coast have been attributed to increased ice stresses in these regions (Hibler 1979; Thorndike and Colony 1982; Kimura and Wakatsuchi 2000). Often, larger internal ice stresses are associated with higher ice concentration (Hibler 1979). Interestingly, we show that LRP shows increased relevance for ice concentration near the coast in comparison to other regions, which could be linked to this mechanism. Last, we note that the attribution of previous-day sea ice velocity in coastal regions compared to the central Arctic is higher for PFI, but lower for LR and LRP.

Some of the discrepancies in attribution between the global XAI methods could stem from the nature of the different types of models and XAI method applied. While LR incorporates nonlocal information, it is inherently nonlinear. The other two XAI methods are applied to the CNN, which is inherently nonlinear and nonlocal. Predictions of ice motion at a given coastal location could be using nonlinear information about wind from other locations. Additionally, the PFI is a sensitivity method measuring importance. The higher importance for wind in coastal regions for PFI means that when wind is randomized, the model becomes less skillful at predicting sea ice motion in coastal regions. This decrease in performance could be related to a loss of information from either local or nonlocal winds. The LRP is a salience method that measures relevance. The fact that wind has coastal relevance in LRP suggests that wind velocity from coastal locations is relevant for predicting ice velocity throughout the Arctic.

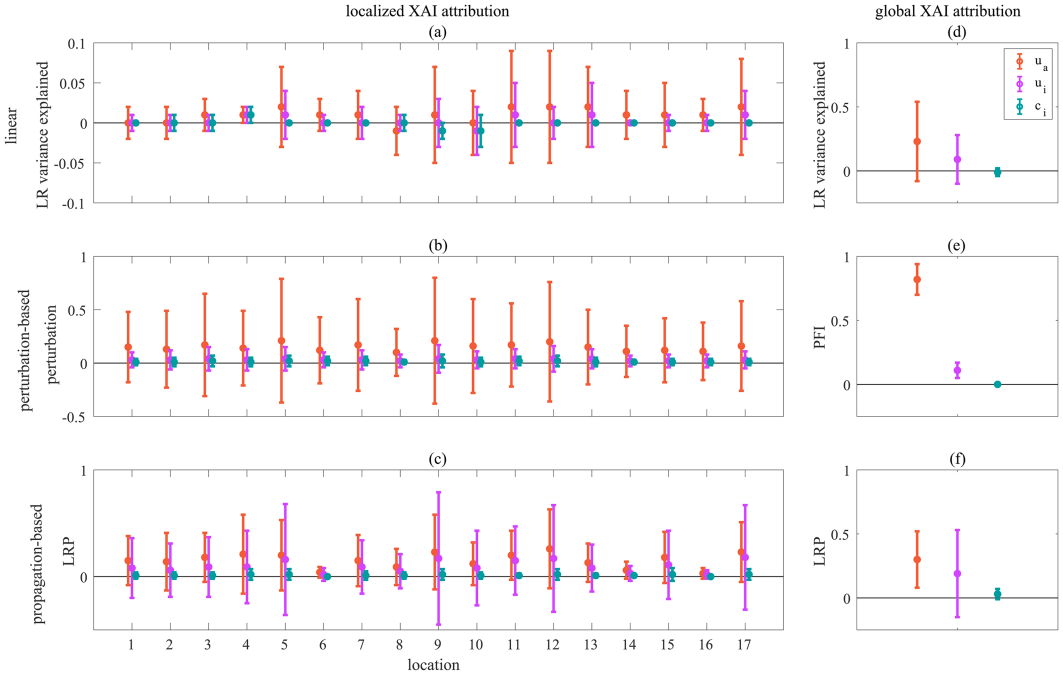

FIG. 9. Spatial mean relevance and importance values for (a)–(d) local and (e)–(f) global XAI methods. Local results are shown for all 17 locations. The rows represent the different types of XAI methods [(a) linear, (b) PERT based, or (c),(d) propagation based], and the colors represent each of the different input features: red, wind velocity u_a; purple, ice velocity u_i; and blue, sea ice concentration c_i. The information in this figure summarizes the spatial mean values provided in the legends of Figs. 3, 8, and S2–S18.

Additionally, the fact that sea ice velocity has higher relevance for LRP at most locations could be a result of correlations between the input features that are difficult for the LRP to disentangle (Flora et al. 2024). Last, while ice concentration may not be that important by itself, the degree of nonlinearity is tied to extreme values of ice concentration, which suggests that ice concentration may play an important role in nonlinear feature interactions.

b. To what extent does the nonlocality of the CNN predictions and XAI explanations provide information about the relevance of the inputs on land for predicting sea ice motion offshore?

Local XAI methods (variance explained by LR and LRP) show that nonlocal wind over land is relevant for predicting sea ice velocity offshore, suggesting that knowledge of wind over land could be useful in understanding ice motion offshore. While LRP is helpful if you want to know the extent to which the output at one grid point is sensitive to the input features throughout the domain, perturbation is useful if you want to know the spatial extent to which the inputs at one point can inform the output. Therefore, we investigate the effect of perturbing wind velocity on land at the locations of several Arctic research stations. Results show that measurements of wind velocity made at these land-based stations are valuable for making predictions of ice motion offshore. This is likely related to the large-scale nature of atmospheric variability, as there are large spatial correlations in the wind patterns (not shown). Studies such as these could be further utilized for field campaign designs to understand where it would be important to place research stations to obtain the most useful data for predicting sea ice motion.

c. To what extent does the CNN incorporate nonlinear information when making predictions?

The differences between the explanations of the LR and CNN models highlight the nonlinear interactions inherent in the CNN. Furthermore, decomposition of the CNN into linear and nonlinear components shows that the model is weakly described by linear terms and tends toward nonlinearity for perturbations made in peripheral seas and coastal regions. We compare the degree of linearity to sea ice concentration and find that the model tends toward nonlinearity for both high ($c_i > 0.8$) and low ($c_i < 0.2$) ice concentration. In this regime, it is possible that third-order terms are also important. These terms would show up in δH_1, inflating the importance of what we have assumed to be the linear terms in this study. Thus, for these cases, the nonlinear component may be even larger than the linear component of the model. For high ice concentration, the linear relationship between wind speed and ice motion is known to diminish, as ice is not in a state of free drift where it is highly responsive to wind forcing (Lepparänta 2011). Therefore, it makes sense that nonlinearities become more important for increasing ice concentration, as the known linear relationship falls apart and the model must rely on other inputs or input combinations to make predictions. For

low ice concentration, nonlinearities could result from the treatment of sea ice velocity when there is no ice. While in reality the sea ice velocity would be undefined for the case of no ice, due to the inability for the CNN to take NaN values as inputs, we have set u_i to zero when c_i is zero. This could lead to discontinuities in the model predictions, as ice with an extreme low value for concentration could potentially have a high velocity. This is important, and model results may be improved by setting the ice velocity to be some fraction of the wind velocity rather than zero.

We note that our results that the model is weakly linear could be influenced by the fact that the analysis does not quite meet the condition for the assumption that the first term represented the linear part of the model response (i.e., δH_1 is not always greater than δH_2 by a full OM). Future work would address this by evaluating the contributions of higher-order terms of the Taylor expansion, which could contribute to δH_1 being larger than δH_2 in this study. Additionally, other physical mechanisms (e.g., the influence of coastal topography on the winds) could also play a role in the degree of nonlinearity, and we suggest this as an avenue for further study. Future work could also apply perturbations to different combinations of inputs in efforts to tease out the nonlinear interactions that are occurring between the different input features. The short integration period (i.e., 1-day prediction time scale) could also contribute toward the model being less dependent on nonlinear terms, as intrinsic nonlinearity scales with prediction time (Verdy et al. 2014). It will be interesting to see if a CNN built for multiday predictions exhibits more nonlinear behavior, but we leave this for future work.

In summary, while the linear terms are larger than the nonlinear terms, they are not an order of magnitude larger suggesting that nonlinear relationships are an important part of the solution and that these components may be why a CNN outperforms LR for many instances in space and time in Hoffman et al. (2023). Additionally, the tendency for the model to become more nonlinear at lower sea ice concentration highlights the importance of applying models that capture this nonlinear nature, such as a CNN, as sea ice in the Arctic diminishes.

6. Conclusions

In the Arctic with diminishing sea ice cover, predictive models that incorporate nonlinear and nonlocal information (particularly neural networks) will become increasingly more important as the ice enters a state of free drift where motion is driven more by wind forcing than by local internal ice stresses (Spreen et al. 2011; Zhang et al. 2012; Tandon et al. 2018; Maeda et al. 2020). The black-box nature of neural networks makes them difficult to interpret without additional tools. Applying XAI is crucial for gaining trust in model predictions and can also be useful for providing new insights into physical processes when applied to skillful models.

In this study, we apply a novel implementation of a global LRP method, which integrates the local explanations provided for predictions made at each grid point. We note that this global LRP is computationally costly in comparison to other global XAI methods (Table 1). Additionally, the sheer amount of information it provides can be challenging to work with and requires patience. There are also discrepancies in the spatial variability of the explainability for global LRP in comparison to what is known from physics and other XAI methods. However, we have shown that the spatially averaged output from global LRP is consistent with other global XAI methods and from what is expected based on what we know about the physics of sea ice motion. On the basis of this analysis, we recommend global LRP as a powerful tool for understanding regression predictions made by the CNN models applied to problems in the Earth sciences and highlight the benefit that this method provides in showing the nonlinear and nonlocal effects of the model predictors. In future work, we aim to use global LRP to understand the spatiotemporal variability of the relevance of wind for predicting sea ice motion. To avoid complexities that arise from feature correlations, we will simplify this analysis and run using wind velocity as the only input feature.

Overall, we find that XAI methods generally agree that wind velocity is the input feature with the largest contribution to predictions of sea ice motion. We discuss nuances in the spatial variability of the relevance or importance of each predictor produced by each XAI method. We also confirm the ability of the CNN to incorporate nonlocal and nonlinear information into its predictions, which is a highly useful feature.

Acknowledgments. L. H. was supported by ONR (Grant N00014-20-1-2772). M. R. M. was supported by ONR (Grant N00014-20-1-2772) and by NSF (Award OPP-1936222). S. T. G. was supported by NSF (Award OPP-1936222) and by U.S. Department of Energy (DOE) (Award DE-SC002007). D. G. was supported by NSF Award 1928305. P. H. was supported by ONR (Grant N00014-20-1-2772). Figures in this report were prepared using MATLAB, Matplotlib: A 2D Graphics Environment Hunter (2007). Colormaps were obtained using the cmocean package (Thyng et al. 2016) and the CubeHelix Colormap (Green 2011). We thank our reviewers for their helpful feedback.

Data availability statement. We acknowledge all sources of publicly available data that were used in this study. The JRA55-do dataset can be accessed at https://climate.mri-jma.go.jp/pub/ocean/JRA55-do/. Polar Pathfinder Daily 25-km EASE-Grid Sea Ice Motion Vectors, version 4, are made available by the National Snow and Ice Data Center (NSIDC) and can be accessed at https://nsidc.org/data/nsidc-0116/versions/4. Sea Ice Concentrations from Nimbus-7 Passive Microwave Data, version 1, are made available by the NSIDC and can be accessed at https://doi.org/10.5067/8GQ8LZQVL0VL. All of the data and files used for processing for this paper can be accessed through https://doi.org/10.6075/J0S182Q6.

APPENDIX

A Taxonomy of XAI Methods

Several XAI methods have been developed to explain predictions made by ML models (Haar et al. 2023; Linardatos

et al. 2021; McGovern et al. 2019; Samek and Müller 2019; Mayer and Barnes 2021). We briefly discuss the taxonomy of these methods to clarify terminology and ensure consistency throughout the literature. Thorough descriptions of how to classify different explainability methods are provided by Bommer et al. (2024), Flora et al. (2024), Das and Rad (2020), and Mamalakis et al. (2022). In summary, classification is based on the following:

- Usage (posthoc vs antehoc): Antehoc methods modify the model architecture to improve interpretability, whereas posthoc methods are applied to any neural network model after it has been trained.
- Scope (global vs local): Local XAI methods provide explanations for a specific prediction (i.e., a pixel or grid point), whereas global methods show attributions of the input features across all samples (Molnar 2020). By assuming linearity, local explanations can be aggregated to create global explanations (Murdoch et al. 2019; Molnar 2020). We distinguish that here the term "global" is used to describe the scope of the XAI methods rather than the spatial extent of Earth.
- Methodology (perturbation vs propagation): Perturbation-based analyses are iterative and test the model's response to perturbations. A perturbation is applied to the input features at selected grid locations after the model is trained, and the degree to which the model prediction changes in response to the perturbation is an indicator of the relevance of the input feature at the perturbed point for the model in making a prediction (Samek and Müller 2019; Linardatos et al. 2021). On the other hand, propagation-based methods integrate the internal structure of the model into the explanation process and focus on the influence that each input value (i.e., each predictor at each grid point) has on activating part of a neural network (Samek and Müller 2019). Propagation-based methods only require one forward and backward pass through the model to generate a relevance visualization. These methods propagate the model prediction through the neural network and analyze the weights and activations at each layer of the model based on certain propagation rules (i.e., analyzing gradients or applying conservation rules to the relevance values propagated to each node in the model). Outputs from propagation-based methods are heatmaps that have the same dimensions as the input features, allowing a comprehensive evaluation of relevance.
- Model awareness (model specific vs model agnostic): Model-specific methods use components of the model (i.e., model weights) for the explanation and therefore rely on the specific model architecture to produce an explainability output. Model-agnostic methods concern only the model outputs and do not incorporate the model architecture into explainability. Typically, perturbation-based methods are model-agnostic, and propagation-based methods are model-specific.
- Type of explanation output (sensitivity vs salience): Sensitivity is a measure of how much the output will change based on changes to a particular input feature and is measured by the units of output per unit of input. Salience is a measure of the relative contribution of the input feature for making a prediction and indicates how many units of the output are explained by the given input feature. (Mamalakis et al. 2022).
- Feature importance versus relevance: Importance is a measure of the feature contribution to the performance of the model. Relevance methods measure the contribution of the features to the model output (Flora et al. 2024).

REFERENCES

Abadi, M., and Coauthors, 2015: TensorFlow: Large-scale machine learning on heterogeneous distributed systems Version 2.8.0. TensorFlow, https://www.tensorflow.org/.

Alber, M., and Coauthors, 2019: INNvestigate neural networks! *J. Mach. Learn. Res.*, **20**, 1–8.

Amador, A., S. T. Merrifield, R. A. McCarthy, R. Young, and E. J. Terrill, 2021: Wave glider speed model for real-time motion planning. *OCEANS 2021: San Diego—Porto*, San Diego, CA, Institute of Electrical and Electronics Engineers, 1–9, https://doi.org/10.23919/OCEANS44145.2021.9705782.

Araujo, A., W. Norris, and J. Sim, 2019: Computing receptive fields of convolutional neural networks. *Distill*, **4**, e21, https://doi.org/10.23915/distill.00021.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, 2015: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, **10**, e0130140, https://doi.org/10.1371/journal.pone.0130140.

Bommer, P., M. Kretschmer, A. Hedström, D. Bareeva, and M. M.-C. Höhne, 2024: Finding the right XAI method—A guide for the evaluation and ranking of explainable AI methods in climate science. arXiv, 2303.00652v2, https://doi.org/10.48550/arXiv.2303.00652.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, and M. H. Savoie, 2012: EASE-Grid 2.0: Incremental but significant improvements for Earth-gridded data sets. *ISPRS Int. J. Geo-Inf.*, **1**, 32–45, https://doi.org/10.3390/ijgi1010032.

Camps-Valls, G., X. Xiang Zhu, D. Tuia, and M. Reichstein, 2021: Introduction. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, G. Camps-Valls et al., Eds., John Wiley and Sons, Ltd., 1–11.

Cavalieri, D. J., C. L. Parkinson, P. Gloersen, and H. J. Zwally, 1996: Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS passive microwave data, version 1. NASA National Snow and Ice Data Center Distributed Active Archive Center, accessed 14 February 2023, https://nsidc.org/data/NSIDC-0051/versions/1.

Das, A., and P. Rad, 2020: Opportunities and challenges in Explainable Artificial Intelligence (XAI): A survey. arXiv, 2006.11371v2, https://doi.org/10.48550/arXiv.2006.11371.

Ding, X., C. Xia, X. Zhang, X. Chu, J. Han, and G. Ding, 2021: RepMLP: Re-parameterizing convolutions into fully-connected layers for image recognition. arXiv, 2105.01883v3, https://doi.org/10.48550/arXiv.2105.01883.

Dobrescu, A., M. V. Giuffrida, and S. A. Tsaftaris, 2019: Understanding deep neural networks for regression in leaf counting. *2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, Institute of

Electrical and Electronics Engineers, 2600–2608, https://doi.org/10.1109/CVPRW.2019.00316.

Ebert-Uphoff, I., and K. Hilburn, 2020: Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. *Bull. Amer. Meteor. Soc.*, **101**, E2149–E2170, https://doi.org/10.1175/BAMS-D-20-0097.1.

Eyring, V., and Coauthors, 2024: Pushing the frontiers in climate modelling and analysis with machine learning. *Nat. Climate Change*, **14**, 916–928, https://doi.org/10.1038/s41558-024-02095-y.

Flora, M. L., C. K. Potvin, A. McGovern, and S. Handler, 2024: A machine learning explainability tutorial for atmospheric sciences. *Artif. Intell. Earth Syst.*, **3**, e230018, https://doi.org/10.1175/AIES-D-23-0018.1.

Gagne, D. J., II, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

García, S., J. Luengo, and F. Herrera, 2015: Introduction. *Data Preprocessing in Data Mining*, Springer, 1–17, https://doi.org/10.1007/978-3-319-10247-4_1.

Gil, Y., and Coauthors, 2018: Intelligent systems for geosciences: An essential research agenda. *Commun. ACM*, **62**, 76–84, https://doi.org/10.1145/3192335.

Gordon, E. M., and E. A. Barnes, 2022: Incorporating uncertainty into a regression neural network enables identification of decadal state-dependent predictability in CESM2. *Geophys. Res. Lett.*, **49**, e2022GL098635, https://doi.org/10.1029/2022GL098635.

Green, D. A., 2011: A colour scheme for the display of astronomical intensity images. *Bull. Astron. Soc. India*, **39**, 289–295.

Haar, L. V., T. Elvira, and O. Ochoa, 2023: An analysis of explainability methods for convolutional neural networks. *Eng. Appl. Artif. Intell.*, **117**, 105606, https://doi.org/10.1016/j.engappai.2022.105606.

Han, T., S. Srinivas, and H. Lakkaraju, 2022: Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. arXiv, 2206.01254v3, https://doi.org/10.48550/arXiv.2206.01254.

Hibler, W. D., III, 1979: A dynamic thermodynamic sea ice model. *J. Phys. Oceanogr.*, **9**, 815–846, https://doi.org/10.1175/1520-0485(1979)009<0815:ADTSIM>2.0.CO;2.

Hilburn, K. A., I. Ebert-Uphoff, and S. D. Miller, 2021: Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using GOES-R satellite observations. *J. Appl. Meteor. Climatol.*, **60**, 3–21, https://doi.org/10.1175/JAMC-D-20-0084.1.

Hoffman, L., M. R. Mazloff, S. T. Gille, D. Giglio, C. M. Bitz, P. Heimbach, and K. Matsuyoshi, 2023: Machine learning for daily forecasts of Arctic sea-ice motion: An attribution assessment of model predictive skill. *Artif. Intell. Earth Syst.*, **2**, 230004, https://doi.org/10.1175/AIES-D-23-0004.1.

Hunter, J. D., 2007: Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55.

Ivanovs, M., R. Kadikis, and K. Ozols, 2021: Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognit. Lett.*, **150**, 228–234, https://doi.org/10.1016/j.patrec.2021.06.030.

Karpatne, A., I. Ebert-Uphoff, S. Ravela, H. A. Babaie, and V. Kumar, 2019: Machine learning for the geosciences: Challenges and opportunities. *IEEE Trans. Knowl. Data Eng.*, **31**, 1544–1554, https://doi.org/10.1109/TKDE.2018.2861006.

Kimura, N., and M. Wakatsuchi, 2000: Relationship between sea-ice motion and geostrophic wind in the Northern Hemisphere. *Geophys. Res. Lett.*, **27**, 3735–3738, https://doi.org/10.1029/2000GL011495.

Kwok, R., G. Spreen, and S. Pang, 2013: Arctic sea ice circulation and drift speed: Decadal trends and ocean currents. *J. Geophys. Res. Oceans*, **118**, 2408–2425, https://doi.org/10.1002/jgrc.20191.

Labe, Z. M., and E. A. Barnes, 2021: Detecting climate signals using explainable AI with single-forcing large ensembles. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002464, https://doi.org/10.1029/2021MS002464.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol.*, **32**, 1209–1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Lepparänta, M., 2011: *The Drift of Sea Ice*. Springer-Praxis Books, Springer, 350 pp.

Letzgus, S., P. Wagner, J. Lederer, W. Samek, K.-R. Müller, and G. Montavon, 2021: Toward explainable AI for regression models. arXiv, 2112.11407v2, https://doi.org/10.48550/arXiv.2112.11407.

Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis, 2021: Explainable AI: A review of machine learning interpretability methods. *Entropy*, **23**, 18, https://doi.org/10.3390/e23010018.

Lundberg, S., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. arXiv, 1705.07874v2, https://doi.org/10.48550/arXiv.1705.07874.

Maeda, K., N. Kimura, and H. Yamaguchi, 2020: Temporal and spatial change in the relationship between sea-ice motion and wind in the Arctic. *Polar Res.*, **39**, https://doi.org/10.33265/polar.v39.3370.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artif. Intell. Earth Syst.*, **1**, e220012, https://doi.org/10.1175/AIES-D-22-0012.1.

——, ——, and ——, 2023: Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artif. Intell. Earth Syst.*, **2**, e220058, https://doi.org/10.1175/AIES-D-22-0058.1.

Marquardt, D. W., and R. D. Snee, 1975: Ridge regression in practice. *Amer. Stat.*, **29**, 3–20, https://doi.org/10.1080/00031305.1975.10479105.

Mayer, K. J., and E. A. Barnes, 2021: Subseasonal forecasts of opportunity identified by an explainable neural network. *Geophys. Res. Lett.*, **48**, e2020GL092092, https://doi.org/10.1029/2020GL092092.

McGovern, A., R. Lagerquist, D. J. Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

McNutt, S. L., and J. E. Overland, 2003: Spatial hierarchy in Arctic sea ice dynamics. *Tellus*, **55A**, 181–191, https://doi.org/10.3402/tellusa.v55i2.12088.

Molina, M. J., D. J. Gagne, and A. F. Prein, 2021: A benchmark to test generalization capabilities of deep learning methods to classify severe convective storms in a changing climate. *Earth Space Sci.*, **8**, e2020EA001490, https://doi.org/10.1029/2020EA001490.

Molnar, C., 2020: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 320 pp.

Montavon, G., S. Bach, A. Binder, W. Samek, and K.-R. Müller, 2015: Explaining nonlinear classification decisions with deep Taylor decomposition. arXiv, 1512.02479v1, https://doi.org/10.48550/arXiv.1512.02479.

——, W. Samek, and K.-R. Müller, 2018: Methods for interpreting and understanding deep neural networks. *Digital Signal Process.*, **73**, 1–15, https://doi.org/10.1016/j.dsp.2017.10.011.

——, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Muller, 2019: Layer-wise relevance propagation: An overview. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek et al., Eds., Lecture Notes in Computer Science, Vol. 11700, Springer, 193–209.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, 2019: Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA*, **116**, 22 071–22 080, https://doi.org/10.1073/pnas.1900654116.

Radivojac, P., Z. Obradovic, A. K. Dunker, and S. Vucetic, 2004: Feature selection filters based on the permutation test. *Machine Learning: ECML 2004*, J.-F. Boulicaut et al., Eds., Springer, 334–346.

Rahman, M. M., K. Matsuo, S. Matsuzaki, and S. Purushotham, 2021: DeepPseudo: Pseudo value based deep learning models for competing risk analysis. *35th AAAI Conf. Artificial Intelligence*, Online, AAAI Press, 479–487, https://ojs.aaai.org/index.php/AAAI/article/view/16125.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Caravalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Samek, W., and K.-R. Müller, 2019: Towards explainable artificial intelligence. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 1st ed. W. Samek et al., Eds., Springer, 5–22.

——, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, 2021: Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE*, **109**, 247–278, https://doi.org/10.1109/JPROC.2021.3060483.

Schnake, T., O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, 2020: Higher-order explanations of graph neural networks via relevant walks. arXiv, 2006.03589v3, https://doi.org/10.48550/arXiv.2006.03589.

Schneider, T., N. Jeevanjee, and R. Socolow, 2021: Accelerating progress in climate science. *Phys. Today*, **74**, 44–51, https://doi.org/10.1063/PT.3.4772.

Shen, X., J. Zhang, X. Zhang, J. Meng, and C. Ke, 2017: Sea ice classification using Cryosat-2 altimeter data by optimal classifier–feature assembly. *IEEE Geosci. Remote Sens. Lett.*, **14**, 1948–1952, https://doi.org/10.1109/LGRS.2017.2743339.

Sinha, A., and R. Abernathey, 2021: Estimating ocean surface currents with machine learning. *Front. Mar. Sci.*, **8**, 672477, https://doi.org/10.3389/fmars.2021.672477.

Spreen, G., R. Kwok, and D. Menemenlis, 2011: Trends in Arctic sea ice drift and role of wind forcing: 1992–2009. *Geophys. Res. Lett.*, **38**, L19501, https://doi.org/10.1029/2011GL048970.

Sundararajan, M., A. Taly, and Q. Yan, 2017: Axiomatic attribution for deep networks. arXiv, 1703.01365v2, https://doi.org/10.48550/arXiv.1703.01365.

Swierczek, S., M. R. Mazloff, and J. L. Russell, 2021: Investigating predictability of DIC and SST in the Argentine basin through wind stress perturbation experiments. *Geophys. Res. Lett.*, **48**, e2021GL095504, https://doi.org/10.1029/2021GL095504.

Tandon, N. F., P. J. Kushner, D. Docquier, J. J. Wettstein, and C. Li, 2018: Reassessing sea ice drift and its relationship to long-term Arctic sea ice loss in coupled climate models. *J. Geophys. Res. Oceans*, **123**, 4338–4359, https://doi.org/10.1029/2017JC013697.

Thorndike, A. S., and R. Colony, 1982: Sea ice motion in response to geostrophic winds. *J. Geophys. Res.*, **87**, 5845–5852, https://doi.org/10.1029/JC087iC08p05845.

Thyng, K. M., C. A. Greene, R. D. Hetland, H. M. Zimmerle, and S. F. DiMarco, 2016: True colors of oceanography: Guidelines for effective and accurate colormap selection. *Oceanography*, **29** (3), 9–13, https://doi.org/10.5670/oceanog.2016.66.

Toms, B. A., E. A. Barnes, and I. Ebert-Uphott, 2020: Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002002, https://doi.org/10.1029/2019MS002002.

Tschudi, M., W. N. Meier, J. S. Stewart, C. Fowler, and J. Maslanik, 2019: Polar pathfinder daily 25 km EASE-Grid sea ice motion vectors, version 4. NASA National Snow and Ice Data Center Distributed Active Archive Center, accessed 14 February 2023, https://nsidc.org/data/NSIDC-0116/versions/4.

Tsujino, H., and Coauthors, 2018: JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do). *Ocean Modell.*, **130**, 79–139, https://doi.org/10.1016/j.ocemod.2018.07.002.

Verdy, A., M. R. Mazloff, B. D. Cornuelle, and S. Y. Kim, 2014: Wind-driven sea level variability on the California coast: An adjoint sensitivity analysis. *J. Phys. Oceanogr.*, **44**, 297–318, https://doi.org/10.1175/JPO-D-13-018.1.

Wilming, R., C. Budding, K.-R. Müller, and S. Haufe, 2022: Scrutinizing XAI using linear ground-truth data with suppressor variables. *Mach. Learn.*, **111**, 1903–1923, https://doi.org/10.1007/s10994-022-06167-y.

Zhang, J., R. Lindsay, A. Schweiger, and I. Rigor, 2012: Recent changes in the dynamic properties of declining Arctic sea ice: A model study. *Geophys. Res. Lett.*, **39**, L20503, https://doi.org/10.1029/2012GL053545.