ADVANCING EARTH AND SPACE SCIENCES

JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000669

Probabilistic Forecasts of September Arctic Sea Ice Extent at the Interannual Timescale With Data-Driven Statistical Models

L. Hoffman[1], **F. Massonnet**[1], and **A. Sticker**[1]

[1]Earth and Life Institute (ELI), Earth and Climate, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Key Points:

- Statistical models can skillfully and reliably predict Arctic September sea ice extent (SIE) state transitions with a time lag of up to 5 years
- A reliable transfer operator predicts an 86% chance that September SIE will exceed the forced trend for the 2024–2028 period
- A neural network using information about the area of thick sea ice shows skill when initialized before the spring predictability barrier

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. Hoffman,
lauren.hoffman@uclouvain.be

Citation:

Hoffman, L., Massonnet, F., & Sticker, A. (2025). Probabilistic forecasts of September Arctic sea ice extent at the interannual timescale with data-driven statistical models. *Journal of Geophysical Research: Machine Learning and Computation*, 2, e2025JH000669. https://doi.org/10.1029/2025JH000669

Received 3 MAR 2025
Accepted 29 JUL 2025

Author Contributions:

Conceptualization: L. Hoffman, F. Massonnet
Data curation: L. Hoffman, A. Sticker
Formal analysis: L. Hoffman
Funding acquisition: F. Massonnet
Investigation: L. Hoffman
Methodology: L. Hoffman
Project administration: F. Massonnet
Resources: F. Massonnet
Supervision: F. Massonnet
Validation: L. Hoffman, F. Massonnet
Visualization: L. Hoffman

Abstract The widespread impacts of declining Arctic sea ice necessitate accurate and reliable predictions. While much focus has been placed on subseasonal to seasonal forecasts or multidecadal projections, seasonal to interannual predictions—crucial for planning and infrastructure—have received less emphasis. Internal climate variability is a dominant source of uncertainty on these timescales, yet initialized dynamical climate model predictions have limited usefulness due to biases and long-term drift that leads to poor skill beyond seasonal timescales. This study develops statistical models—transfer operators (TO) and neural networks (NN)—to forecast probabilistic state transitions of Arctic September sea ice extent (SIE) internal variability. Trained on 24,420 transitions from the CMIP6 archive, these models make accurate and reliable predictions across multiple initialization months. At interannual timescales, they outperform simple persistence in predicting SIE trends. At seasonal timescales, their skill is comparable to other numerical and statistical models in the Sea Ice Outlook. While TO performance declines for spring initializations, NNs incorporating information about the area of thick ice can overcome the spring predictability barrier for March–May initializations. For the next decade, the TO suggests that September SIE will likely remain above the projected forced trend, while the NN predicts it will likely be lower. However, both models predict that the trend in September SIE will likely be higher (65%–96% chance) than the CMIP6 projected forced trend over the next 3 years, suggesting near-term stability. These results highlight the potential of statistical approaches for improving Arctic sea ice predictions on critical planning timescales.

Plain Language Summary The year-to-year variability in Arctic sea ice cover is challenging to predict, but is of great importance due to its widespread environmental, geopolitical, and logistical impacts. We show that statistical models can make skillful and reliable predictions of Arctic sea ice extent (SIE) by learning from variability in the sea ice state simulated by climate models. We compare the performance of two different statistical models in hindcasting September Arctic SIE at seasonal timescales, for known years of rapid and slower sea ice loss, and for years of extreme sea ice loss. These models are then used for future forecasts, where they predict an anomalously high Arctic SIE over the next 3-year period. However, the different model types disagree in predictions over the next decade: one model predicts Arctic SIE will be anomalously high, and one predicts it will be anomalously low. This work is a novel example of statistical models providing skillful and reliable predictions of Arctic sea ice on timescales further out than one season.

1. Introduction

The state of Arctic sea ice is one of the key indicators and drivers of climate change (IPCC, 2023). An overall decline in Arctic sea ice extent (SIE) has been observed in the satellite record in all months for over four decades (Onarheim et al., 2018; Serreze et al., 2007; Stroeve & Notz, 2018). September shows the largest reduction in SIE, with a trend of $-13.2\% \pm 1.1\%$ per decade for the 1979–2023 period (Roach & Meier, 2024). An observed downward trend in sea ice area has also been captured by the multimodel ensemble spread of simulations from the sixth phase of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. (2015)), but individual models still face challenges in accurately representing this decline (Lee et al., 2023; Notz et al., 2020).

This long-term sea ice decline is the combined result of increasing radiative forcing associated with rising atmospheric greenhouse gas concentrations (IPCC, 2023; Notz & Stroeve, 2016) and internal (natural) variability (Swart et al., 2015). This internal variability can manifest in the form of (a) extended periods of stable or even increasing SIE, (b) periods of large interannual variability in SIE, or (c) extended periods of sea ice decline, known as rapid ice loss events (RILEs; Holland et al. (2006), Kay et al. (2011), Swart et al. (2015)). Internal

10.1029/2025JH000669

Writing – original draft: L. Hoffman
Writing – review & editing: L. Hoffman, F. Massonnet, A. Sticker

variability remains a dominant source of uncertainty in predicting Arctic SIE on seasonal to interannual timescales (Bonan et al., 2021), as its drivers are still poorly understood and constrained (Dörr et al., 2023). Additionally, sea ice predictability is related to the mean ice state (Massonnet et al., 2018), and is therefore expected to change as Arctic sea ice declines (Holland et al., 2019). Therefore, while significant advancements have been made in predicting Arctic sea ice using dynamical numerical models (Holland & Hunke, 2022), they have not yet reached the full potential for predictive skill (Hawkins et al., 2016). Over the past few years, statistical models have been developed in addition to dynamical models for seasonal prediction. Statistical models are generally as skillful as dynamical models for predicting pan-Arctic September SIE at lead times of 1–3 months (Bushuk et al., 2024). In fact, Bushuk et al. (2024) highlight two statistical models (Chi et al., 2021; Ionita et al., 2019) with skill levels approaching the upper limit of pan-Arctic September SIE predictability for a 1 July initialization (Tietsche et al., 2014).

In this study, we explore the potential of statistical models for predictions of SIE beyond one season. The statistical models are trained to predict the internal variability of September SIE in the Arctic, that is, the residual SIE obtained after removing the estimated forced component of SIE variability. Specifically, we train a number of transfer operator (TO) and neural network (NN) models with SIE variations simulated by climate models from CMIP6 (Eyring et al., 2015) that were shown in Notz et al. (2020) to represent sea ice properties from observations within their ensemble spreads. For each method (TO and NN), we build 1200 models, each trained to predict September SIE from a different starting month (January–December, 12 months), hindcast lag (1–10 years), and averaging time (1–10 years). We show that these models can produce skillful and reliable forecasts of SIE compared to anomaly persistence. Skill refers to the model's ability to capture the signal, while reliability indicates the degree to which the forecast probabilities match the frequency of observed occurrences in the test data set. Future predictions indicate that, over the next 3 years, there is a 65%–96% chance that September SIE will be higher than the projected forced trend. Additionally, we provide insight into the feasibility of using TO and NN models to predict state transitions, and discuss the behavior of these models for different input conditions.

2. Data

We train our statistical models using sea ice conditions from historical simulations of climate models from the Coupled Model Intercomparison Project Phase 6 (CMIP6). The sea ice conditions from the future high-emitting greenhouse gas scenario (SSP5-8.5) of CMIP6 are used during our testing phase, as described below in Section 3.1. We use the parameters *siextentn* (precomputed SIE in the Northern Hemisphere), *siconc* (sea ice concentration), and *sithick* (floe ice thickness) from six different models in CMIP6 (model name [number of members in brackets]): ACCESS-CM2 [10], ACCESS-ESM1-5 [40], CanESM5 [25], IPSL-CM6A [33], MIROC6 [50], and MRI-ESM2-0 [10] (Table S1 in Supporting Information S1). These CMIP6 models are chosen based on the selection criterion from Notz et al. (2020), where the ensemble spread captures the observations of both the 2005–2014 September mean sea ice area and the observed sensitivity of sea ice area to cumulative CO_2 emissions over the 1979–2014 period (listed in bold font in Table S4 in Supporting Information S1 of Notz et al. (2020)). Additionally, Sticker et al. (2024) analyzed the occurrence of RILEs, defined as a period lasting at least 4 years during which the trend in the 5-year running mean minimum SIE is lower than -0.3 Mkm^2 per year, across 26 models from CMIP6 (the first ensemble member of each model, as well as five large ensembles). All of the models used in our study were included in their analysis, which found a 92% probability of at least one September RILE occurring over the 1970–2099 period. They also showed that all the models used in our study exhibited at least one RILE during July–September in 1970–2014 (see Figures 2 and 3 of Sticker et al. (2024)). Lastly, our selection of models from CMIP6 meets the criteria that outputs exist for the full temporal range (1850–2014 for historical simulations and 2015–2100 for SSP5-8.5 simulations), and that there are 5 or more members in both historical and SSP5-8.5 simulations.

Our model predictions are based on two predictors: the total SIE and the area of sea ice thicker than a given threshold (SIAt). SIE is provided directly by CMIP6, while SIAt is calculated as the sum of the area of each grid cell where $sithick * siconc$ is greater than $t = 1.25$ m. The choice of SIAt as a predictor is motivated not only by the notion that the proportion of thick ice in winter and spring is a known predictor of summer SIE (Chevallier & Salas-Mélia, 2012; Goosse et al., 2009) but also that the fraction of thin ice controls the longer-term sea ice decline at decadal timescales (Massonnet et al., 2012). Additionally, Chevallier and Salas-Mélia (2012) found that an anomaly of sea ice area in September is potentially predictable up to 6 months in advance using the area covered by ice thicker than 0.9–1.5 m, which motivates our thickness threshold of 1.25 m.

We use observations for the 1979–2024 time period to make predictions and evaluate the model performances. The National Snow and Ice Data Center (NSIDC) Sea Ice Index (Fetterer et al., 2017) provides monthly mean SIE, while the Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS; Schweiger et al. (2011), Zhang and Rothrock (2003)) provides monthly mean SIT (from which we calculate SIAt as the sum of each grid cell area where SIT > 1.25 m). Although the PIOMAS data set exhibits some of the same spatial biases found in CMIP6 models (e.g., an overly smooth thickness gradient toward the Canadian Arctic Archipelago and insufficient thick ice near the coast (Petty et al., 2023)), a study by Labe et al. (2018) shows that PIOMAS provides a realistic representation of sea ice thickness when compared with satellite and submarine observations. Additionally, Petty et al. (2023) find that differences from satellite-derived estimates are typically on the order of tens of centimeters. This evidence supports our decision to use PIOMAS as a testing data set for our statistical models.

3. Methods

3.1. Model Inputs

We build an array of statistical models in the form of transfer operators (TO) and neural networks (NN) to predict the internal variability of September Arctic SIE. The TO models use only SIE as an input, while the NN models use both SIE and the total sea ice area where the sea ice thickness is greater than a threshold value of 1.25 m (SIAt). Each model makes a direct prediction of September SIE from sea ice conditions in a specific earlier month, using lead times ranging from 1 month to 10 years. For each input month (e.g., August, July, etc.), we train separate models using different lead times ($\tau = 1$–10 years) and right-aligned moving average windows ($T = 1$–10 years), where both the input and the target variables are averaged over the same number of years. These multiyear averages are computed across the same calendar month over previous years. Additionally, because we train a separate model for each initialization month, a nominal lead time of 1 year includes models initialized anywhere from 1 to 12 months prior to September. For instance, a model initialized in September of the previous year has a 12-month lead time, while one initialized in October has an 11-month lead time, and so on—yet all are considered $\tau = 1$ year under our notation. In a case with $\tau = 1$ year, $T = 3$ years, and a July initialization, the model uses the average of July SIE (or SIE and SIAt) over years t, $t - 1$, and $t - 2$ to predict the average of September SIE over the same 3 years. A separate model is trained for each combination of input month, averaging window, and lead time, resulting in 1200 model instances (10 lead times × 10 averaging windows × 12 input months) for each model architecture (TO and NN).

We separate the CMIP6 data into training and test sets so that we can verify the performance of the statistical models on data on which they have not been trained. We choose five ensemble members from each CMIP6 model that has more than 10 members total (i.e., ACCESS-ESM1-5, CanESM5, IPSL-CM6A, and MIROC6) as the test set, and use the remaining data for training and validation. This approach allows us to evaluate the models while ensuring the test data are independent of the training data, but is still consistent with the dynamics of the CMIP6 models. This prevents inflated performance metrics that result from using the training data in the model evaluation. We run the statistical models using different subsets of the CMIP6 members for testing and training and find little variation in performance. We also find that using data from multiple CMIP6 models during testing generally produces more skillful and reliable predictions than using any one model (Table S2 in Supporting Information S1).

Because we aim to predict the internal variability in September SIE, we decompose the total variations in SIE (Figure 1a) and SIAt (Figure S1a in Supporting Information S1) into a forced contribution (Figure 1b for SIE and Figure S1b in Supporting Information S1 for SIAt) and a residual contribution (internal variability, Figure 1c for SIE and Figure S1c in Supporting Information S1 for SIAt). For the training and testing data from CMIP6, we calculate the residual as the difference between each ensemble member and the respective model ensemble mean (estimated to represent the forced contribution to the total variability). As a caveat, we note that the CMIP6 models may not accurately represent the forced contribution and often do not correctly simulate the sensitivity of sea ice area to changes in global mean temperature (Notz et al., 2020).

We also test using observations as inputs to the trained statistical models. We use distinct forced contributions for the 1979–2014 and 2015–2024 periods for analyses done with observations because the historical CMIP6 runs only provide output through 2014. We use the historical runs for the 1979–2014 period, and the SSP5-8.5 runs for 2015–2024. The forced component is taken to be the weighted mean of the CMIP6 ensemble means, where the weights are based on the number of ensemble members used from each CMIP6 model. We find that the model performance does not vary substantially when using ensemble means of individual CMIP6 models, versus the weighted

10.1029/2025JH000669

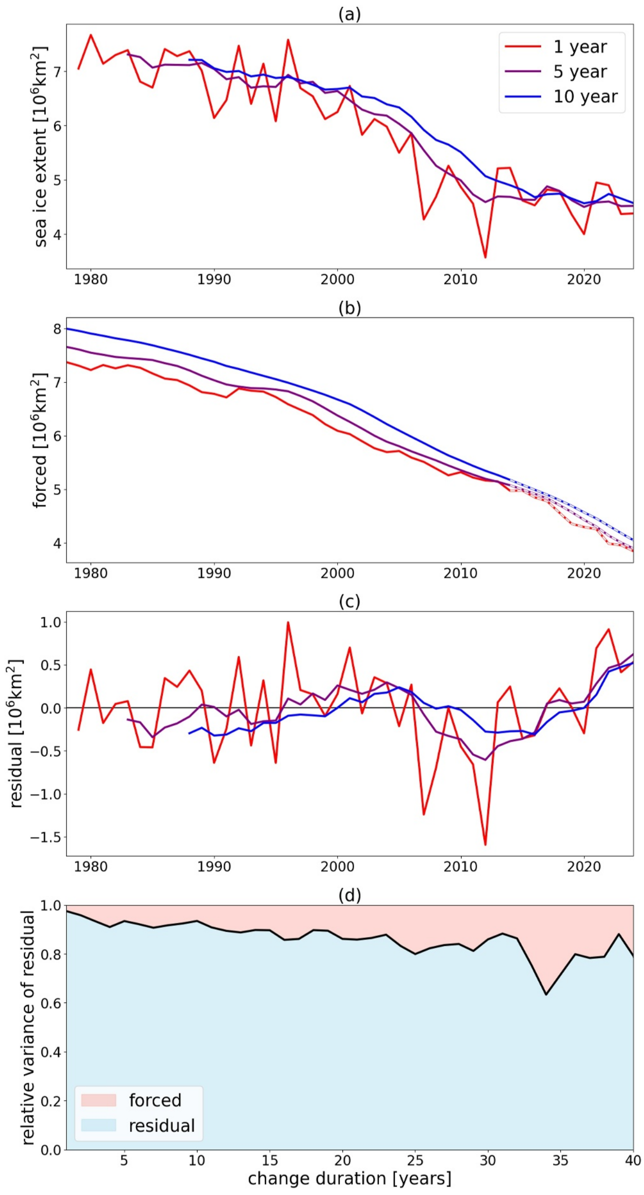

Figure 1. Decomposition of September sea ice extent (SIE) into forced and residual components. For 1 year (red), 5 year (purple), and 10 year (blue) moving means, (a) September SIE from NSIDC is decomposed into a (b) forced and (c) a residual term. The forcing is taken to be the weighted mean of the ensemble mean from six different CMIP6 models from historical (1979–2014; solid lines) and SSP5-8.5 (2015–2024; dashed lines) simulations. (d) The fraction of variance is attributed to the forcing (red) versus the residual (blue) as a function of the number of years over which the variance is calculated (*x*-axis).

multimodel ensemble mean (Table S3 in Supporting Information S1). We only use the ensemble members that are available in both the historical and the SSP5-8.5 runs of the CMIP6 models (the number of ensemble members is in parenthesis in Table S1 in Supporting Information S1). Lastly, we correct the CMIP6 forced response for the trend and mean bias between the CMIP6 data set and observations by rescaling these two attributes to the observed value.

Other methods exist to define the forced term for the observations. For example, the forced term could be taken as either the linear trend in observed SIE or a least squares fit between SIE and global mean temperature. We run tests using these other two types of forcing and find that the model predictions and performances are not highly sensitive to how the forced term is defined (not shown). For these tests, we use global mean temperature from GISTEMP Team (2025) and Lenssen et al. (2024). We opt for using the reconstructed form of the forced term (i.e., weighted CMIP6 ensemble mean of the historical (1979–2014) and SSP5-8.5 (2015–2024) runs) to maintain consistency between how the forcing is defined in the testing and training phases of the model evaluation.

The model inputs are in the form of residuals obtained by removing the forced component from the raw time series and further standardizing to zero mean and one unit standard deviation. For preprocessing both SIE and SIAt, we thus (a) calculate the forced component, (b) calculate the residual as the total signal minus the forced component, and (c) normalize based on the statistics of the training data by subtracting the mean and dividing the residual by the standard deviation. Hence, the probabilistic forecasts are to be understood as normalized departures from the estimated forced component, that is, how the SIE departs from a reference state corresponding to the SIE evolution that would be caused by forcing agents alone. We find that while observations exhibit a higher frequency of low September values for both SIE and SIAt relative to the CMIP6 models, the distributions of the standardized residuals are in good agreement (Figures S2 and S3 in Supporting Information S1), lending additional support to our use of CMIP6 output for training the statistical models. Throughout this study, we show future forecasts in terms of the standardized residual. Hindcasts are presented in terms of the full September SIE signal with the forced term added back into the predicted residual.

3.2. Model Setup

3.2.1. Transfer Operator

A transfer operator is a statistical model that provides the probability of a given output state based on the value of the input state (Froyland, 2001). We apply a transfer operator in the form of a stationary first-order Markov transition matrix (Hamilton, 1994). This type of transfer operator has been applied previously to predict climate states; for example, Sévellec and Drijfhout (2018) build PROCAST to predict state transitions of global mean temperature and sea surface temperature. In this work, we train a transfer operator to predict the probability distribution for September SIE based on the input SIE at a previous time step. The TO is trained using the CMIP6 model data described in Section 2.

The models are set up so that the output is a probability distribution based on bins defined during model training. The bins are in terms of a standardized residual SIE and are centered at zero mean. Based on Sévellec and Drijfhout (2018), we use a uniform bin size of $6\sigma/\eta = 0.25$ that is dependent on the number of bins, $\eta = 24$, and the standard deviation, $\sigma = 1$. We extend the most extreme bins to infinity to account for extremes in SIE.

10.1029/2025JH000669

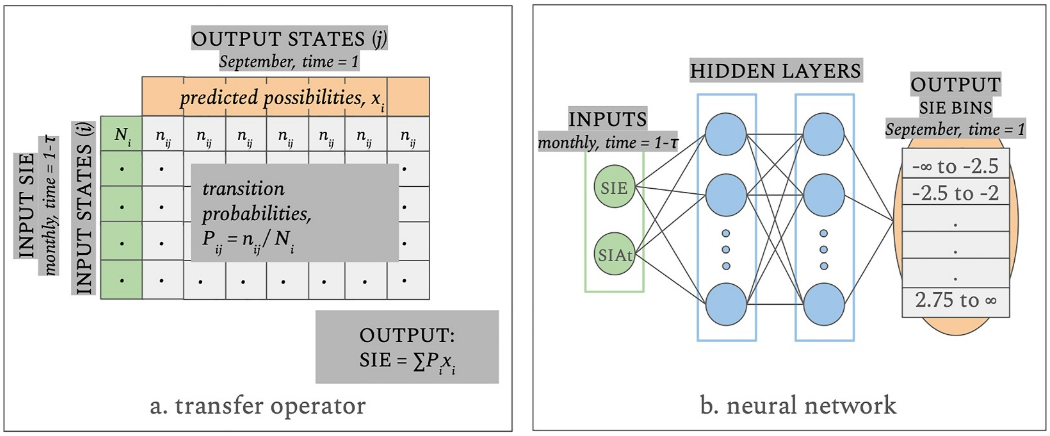

Figure 2. Schematic of the (a) transfer operator (TO) and (b) neural network (NN) models trained to predict the internal variability in September Arctic sea ice extent (SIE). A TO calculates the probability of transitioning to a particular output state based on the value of the input, where the output is September SIE and the input is SIE in a previous month and year. Similarly, the NN is trained to predict the probability that September SIE (output) falls within a given range based on the value of the inputs (SIE and the SIAt in a previous month and year). We train a separate model of each type for each initialization month, hindcast lag (τ = 1–10 years), and averaging time (T = 1–10 years).

We apply the following steps to build the transfer operator: (a) bin the input and output variables (monthly and September SIE) into a predefined and fixed number of states; (b) count the number of inputs in each state, N_i; (c) for each input state, count the number of trajectories to each output state, $n_{i,j}$; and (d) calculate the probability of transitioning from the ith state to the jth state as $P_{i,j} = n_{i,j}/N_i$. $P_{i,j}$ can thus be understood as the probability that the SIE ends up in state j given that it was in state i at the previous step. Figure 2a shows a summary schematic for this transfer operator.

The overall mean prediction is the expected value, $y = \mu = \Sigma x_i \times P(x_i)$, where x_i is the bin value and $P(x_i)$, the bin probability. The bin value, x_i, represents the midpoint of the bin; for extreme bins, this value is set to the bin edge plus (or minus for the negative bins) half of the bin width. The standard deviation of the prediction is calculated as $\sigma = (x_i - \mu)^2 \times P(x_i)$. This is still in the form of a standardized residual; therefore, we apply postprocessing to obtain the full predicted SIE signal as follows: (a) unstandardize based on the training data statistics (multiply by the standard deviation and add back the mean of the training data) and (b) add the forced component back in. The forced component is different in the pre- and postprocessing phases because the inputs are from the monthly mean of the given input month and the output is for September. For preprocessing, we use the forcing value based on the starting month. For hindcasts, we use the September forcing for the output. We show future predictions in terms of probability distributions of the standardized residual. Therefore, our forecasts represent the probability that a future SIE will be lower or higher than the forced contribution.

A natural thought is that including more information in the input would improve the predictive skill of the model. However, we find that a TO that uses only one input (SIE) outperforms a TO with two inputs (SIE and SIAt, not shown). This likely results from the exponential increase in the number of input states with each additional input, which makes it more difficult to capture the statistical behavior of transitions because there are not enough data in each state to skillfully represent the transitions.

3.2.2. Neural Network

We train an NN on a task similar to that of the transfer operator. Our NN is set up for a classification task and predicts the probability distribution for September SIE based on input conditions at a previous time step. For the NN, we have two inputs: monthly SIE and SIAt. The output is a probability distribution based on the number of bins described in Section 3.1. The probability distribution is obtained by fuzzy classification encoding using a Gaussian kernel (Amo et al., 2004; Zadeh, 1965), whereby a particular value of SIE is transformed into a probability distribution where bins are assigned probabilities based on their proximity to the actual SIE value. We apply a softmax function to the output layer of our NN, which ensures that the probabilities of the output classes add up to one. We then apply fuzzy classification decoding to get the output in terms of a continuous value for

SIE, rather than a probability distribution. Fuzzy classification decoding determines the SIE by computing a weighted sum of the SIE class probabilities.

The NN is set up with two hidden layers, each containing 10 nodes with ReLU activation. The model is trained to optimize the Kullback-Leibler (KL) divergence loss function using an Adam optimizer and L2 regularization. The KL divergence is useful for softmax classification problems, as it measures the difference between two probability distributions (Kullback & Leibler, 1951). We run the model for 10 epochs with a batch size of 200. The NN is implemented in Python using the TensorFlow/Keras library (Abadi et al., 2015). The CMIP6 model data are split into training, validation, and testing subsets with a 73%–15%–12% split before input into the NN. Testing data include the last five members from each CMIP6 model used in this study that has more than 10 members (i.e., ACCESS-ESM1-5, CanESM5, IPSL-CM6A-LR, and MIROC6); the remainder of the data is used for training and validation. This split is chosen so that testing can be used represent the "perfect model" case for the NN, in that the NN is being tested on data from models in CMIP6 whose statistics are represented in the trained NN. Further details about the data and model parameters can be found in Table A1.

In contrast to the TO, we find that an NN set up with two input predictors tends to outperform an NN that uses just SIE as an input (Figures S4 and S5 in Supporting Information S1). Where the TO relies on binned values of the inputs, the inputs to the NN are continuous, and therefore, the model does not suffer from the same issue as the TO where additional inputs decrease the ability of the model to describe statistical transitions because there are too many bins.

3.3. Model Evaluation

We use four different metrics for a robust evaluation of the model performance: the coefficient of determination $\left(R^2\right)$, reliability, anomaly correlation coefficient (ACC), and root mean squared error ($RMSE$). Deterministic skill metrics (R^2, ACC, and $RMSE$) are applied to understand the skill of the ensemble mean, while probabilistic metrics, here reliability, assess the ability of the model forecast probabilities to represent the observed frequencies (Hawkins et al., 2016). The R^2 metric is used here as a skill score that reflects how well the model reduces prediction error relative to the variance in the observations. It is commonly used in climate prediction as a benchmark against the climatological mean. The ACC is more sensitive to the phase variability of the anomalies, and is applied in this study to enable direct comparison to results from Bushuk et al. (2024). The ACC focuses on whether the model captures the correct timing of the peaks and troughs in the observed data. The $RMSE$ measures the absolute error, and penalizes both magnitude and phase errors in the predictions.

The coefficient of determination $\left(R^2\right)$ and reliability are defined as

$$R^2 = 1 - \frac{\overline{\left(\overline{x_i P_i(t)}^i - o(t)\right)^2}^t}{\overline{o(t)^2}^t}, \tag{1}$$

and

$$Reliability = \sqrt{\overline{\left(\frac{\left(\overline{x_i P_i(t)}^i - o(t)\right)^2}{\overline{\left[x_i - \overline{x_i P_i(t)}^i\right]^2 P_i(t)}^i}\right)}^t}, \tag{2}$$

where t is time, i is the state index of each prediction bin, $o(t)$ is the verifying observation at time t, x_i are the predicted possibilities (i.e., the middle value of each state or bin), and P_i are the probabilities of each state. The bar represents a temporal mean or a sum over the different states, depending on the superscript.

The coefficient of determination (i.e., R^2) is used here as a skill score that quantifies how well the model reduces prediction error relative to the variance in the observations. It ranges from $-\infty$ to 1, where $R^2 = 1$ indicates perfect predictions, and $R^2 = 0$ corresponds to no skill beyond predicting the climatological mean. Negative values indicate that the model performs worse than simply predicting the climatological mean. This metric differs from the classical R^2, which is defined as the square of the Pearson correlation coefficient and is bounded between 0 and 1. While the Pearson-based R^2 measures the strength of the linear association between predictions and observations, the skill-based R^2 used here assesses how well the predictions reduce the mean squared error relative to a baseline prediction of the mean. Although this formulation is mathematically equivalent to the Nash-Sutcliffe efficiency commonly used in statistical modeling, we adopt the term coefficient of determination $\left(R^2\right)$ following its frequent usage in climate prediction studies (e.g., Sévellec and Drijfhout (2018)).

The reliability score quantifies the consistency between forecast uncertainty and actual error, and indicates how well the forecast probabilities match the observed relative frequencies. This metric measures the ratio of error to uncertainty. A reliability score of 1 indicates that the forecast probabilities perfectly match the observed frequencies. A reliability less than one indicates that the model is underconfident (i.e., it predicts more uncertainty than the errors justify), while a reliability greater than one indicates overconfidence (i.e., the errors exceed the predicted uncertainty). The reliability score is bounded from 0 to ∞, though values ranging from 0 to 2 are typical.

These two metrics are useful not only for evaluating residual terms and probability distributions (i.e., the direct model outputs) but also to assess typical forecasting system biases like over- or underdispersion (Weisheimer & Palmer, 2014). We use these two metrics to evaluate the model performance. We apply a 95% confidence interval to each of the performance metrics to determine which models can be considered either skillful or reliable. The confidence intervals are computed using a bootstrapping method, where random subsets of the inputs are used to calculate the performance metrics over several iterations. The confidence interval is then calculated as the mean plus or minus the standard deviation multiplied by the z-score for a 95% significance level (i.e., 1.96). A model is considered skillful if the coefficient of determination is greater than that of an anomaly persistence forecast (described below) and has a lower-level confidence interval that is greater than zero. A model is considered reliable if a reliability value of 1 is within the lower and upper confidence intervals.

For a comprehensive understanding of model performance, we evaluate the potential skill and the actual skill. The potential skill refers to the potential for a model to make skillful predictions, while actual skill describes the ability of an imperfect model to make forecasts of the real world (DelSole & Tippett, 2018; Hawkins et al., 2016). To assess potential skill, we analyze the perfect model case, where we calculate the model performance when predictions are made using the test data set, which is from the same data set as the training data (i.e., the CMIP6 models). As described in Section 3.1, the testing data are composed of five ensemble members from each of the CMIP6 models that the statistical models are trained on that have more than 10 members. We evaluate the model actual skill by making predictions with inputs from real-world observations. We evaluate the performance separately for each of the 1200 instances (lag times, averaging times, and initialization months) of both models to assess the model performance at different timescales. Anomaly persistence is used as a benchmark in both the perfect model and observational test cases. We evaluate persistence by persisting the SIE anomaly from the initialization month to September. The persistence forecast is deterministic; therefore, we cannot use it for benchmarking the reliability metric, which requires probabilistic distributions.

We also apply two other performance metrics that are commonly used in the sea ice prediction literature to allow direct comparisons between previous studies, such as Bushuk et al. (2024). These are ACC and $RMSE$, and are given by

$$ACC = \frac{\sum_{t=1}^{M}\left(y(t)-\overline{y(t)}^{t}\right)\left(o(t)-\overline{o(t)}^{t}\right)}{\sqrt{\sum_{t=1}^{M}\left(y(t)-\overline{y(t)}^{t}\right)^{2}}\sqrt{\sum_{t=1}^{M}\left(o(t)-\overline{o(t)}^{t}\right)^{2}}}, \tag{3}$$

and

$$RMSE = \sqrt{\frac{\sum_{t=1}^{M}(y(t)-o(t))^{2}}{M}}, \tag{4}$$

where t is time, $o(t)$ is the observation, $y(t)$ is the model prediction, and M is the number of years. The ACC measures the strength and direction of the linear relationship between the predicted and observed anomalies. It is computed as the covariance between the prediction and observation, scaled by the product of their standard deviations. The ACC ranges from −1 to 1, where 1 indicates perfect positive correlation, −1 indicates perfect negative correlation, and 0 indicates no linear relationship. The ACC is a measure of how well the model explains the phase variability in the data, while the $RMSE$ measures the overall error in model predictions, including both phase and amplitude components (Thorpe et al., 2013).

3.4. Model Behavior

We investigate the behavior of the TO and NN models to understand the predictions. First, we investigate how the model predictions respond to a range of initial conditions that lie within the range of the model bins (i.e. −3 to 3). We note that these initial conditions are not linked to any particular output value, and therefore, we cannot calculate the model performance on these predictions. We apply this analysis to both the TO (varying the input SIE from −3 to 3) and the NN (varying both the input SIE and SIAt from −3 to 3) and analyze the model response in terms of the likelihood that the model predicts September SIE to be lower than the climatological mean for each combination of initial conditions.

We also apply explainable artificial intelligence (XAI) methods for a more formal analysis of the model behavior, with particular interest in understanding which of the inputs in the 2DNN (i.e., SIE or SIAt) is more relevant for the model when making predictions of September SIE. Here, XAI methods are applied to the training data for a case in which the model was known to have skill and reliability (i.e., initialized in January–December, 1-year averaging time, and 1-year lag time). We apply two types of XAI: perturbation and deep Taylor decomposition. Perturbation provides information on how the model prediction changes in response to a perturbation in the inputs (Ivanovs et al., 2021; Sinha & Abernathey, 2021). We perturb each input feature separately by adding a fraction of the standard deviation ($+0.5\sigma$), while keeping the other input fixed. The model is then run on each perturbed set to make a prediction. We calculate the RMSD between the model prediction on a control case (nonperturbed) and each of the perturbed inputs. A particular input is considered to be highly relevant if the RMSD is large (i.e., if the model prediction changed as a result of perturbing that particular input). We also compute the 95% confidence intervals for the perturbation relevance of each of input to determine whether their differences are statistically significant. The deep Taylor decomposition (Montavon et al., 2017) provides information on the relevance of each of the input features by propagating backward through the NN and redistributing the activation score of the output layer to the first layer. This method shows how relevant a particular input feature is for the NN to place the highest probability in the correct SIE output bin. We use the iNNvestigate package (Alber et al., 2019) and apply deep Taylor decomposition to the training data. We then calculate the mean relevance for each of the inputs, as well as the p-value to determine whether a particular input feature is significantly more relevant than the other.

4. Results

This section presents results from two predictive models: the TO model, which forecasts September SIE based on SIE from prior months, and the NN model, which incorporates both SIE and SIAt from prior months to predict September SIE. The results of our statistical predictions are based on several assumptions. The first is that knowledge of SIE or the combination of SIE and SIAt is enough to predict September SIE, while in reality, it is established that other drivers (e.g., ocean heat transport) are important predictors (Årthun et al., 2021). The second assumption is present in how we define the signal that represents external forcing for the observations, where Section 3.1 discusses a few choices for what to use as the external forcing to calculate the residual. Lastly, we assume that the statistics are stationary in a changing climate, which is of course not the case in reality.

4.1. Model Performance

We assess the model performances first for the perfect model case (Figure 3) and then using the observed internal variability as inputs (Figure 4). We compare the coefficient of determination (Equation 1) and reliability (Equation 2) of the TO and the NN for different averaging times and lag times. Hatched regions show where anomaly persistence is more skillful than either the TO (Figures 3a and 4a) or the NN (Figures 3c and 4c). Figures 3e and 4e show the difference between the coefficient of determination of the NN and the TO. Red indicates that the NN is

10.1029/2025JH000669

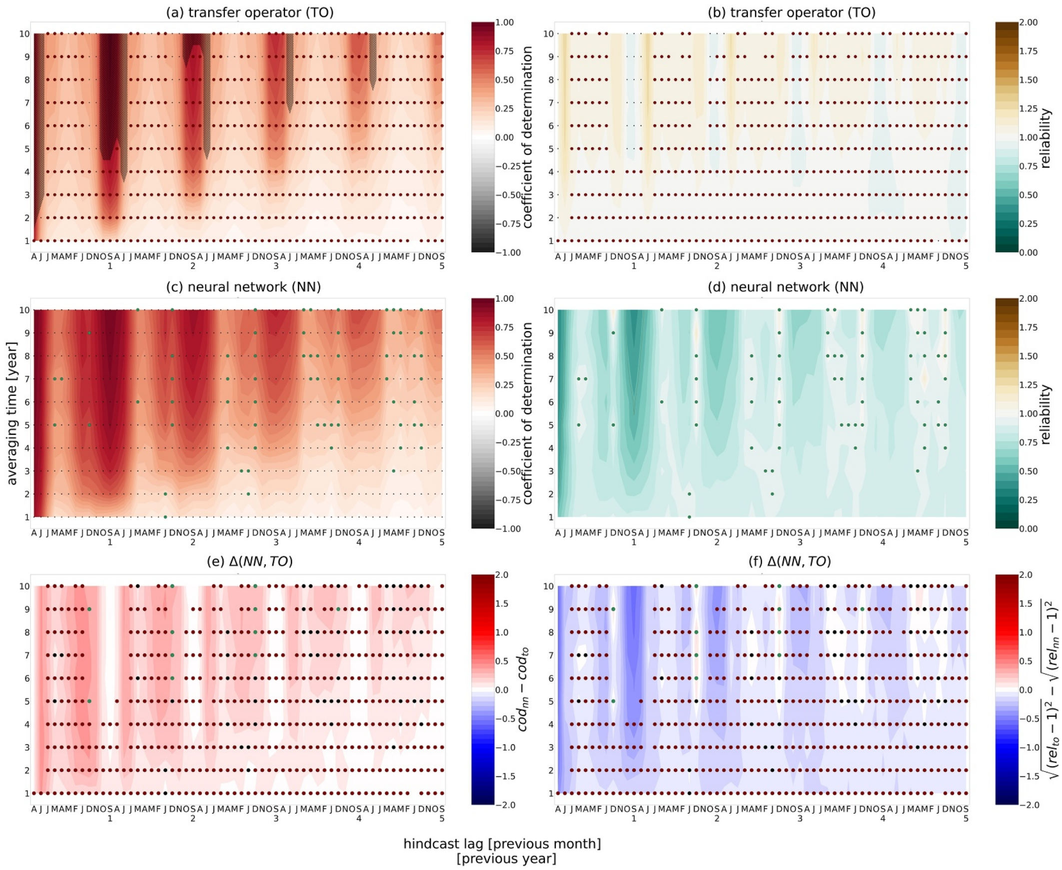

Figure 3. Model performance metrics for interannual hindcasts of sea ice extent using CMIP6 output for various averaging times (y-axis) and hindcast lag months and years (x-axis, top and bottom). The target month is always September; hence, the x-axis runs backward and time as the initialization months become more distant from September. (a and c) Skill of the prediction measured by the coefficient of determination between CMIP6 outputs and the prediction made by (a) a transfer operator and (c) a neural network. Persistence is used as a benchmark, and hatched regions show where the coefficient of determination for persistence is higher than that for the given statistical model. (e) The difference between the coefficient of determination of the NN and the TO; red indicates the NN outperforms the TO, while blue is where the TO outperforms the NN. (b and d) Reliability of the (b) transfer operator and (d) the neural network. (f) The difference between the root mean square distance between the reliability of each model and one (perfectly reliable); red indicates that the NN is more reliable than the TO, while blue shows where the TO is more reliable. Stippling in (a–d) represents where either (a and b) the TO or (c and d) the NN is skillful (a and c) or reliable (b and d). (a–d) Colored dots represent overlapping skill and reliability. Dots in (e–f) show where either there is skill and reliability for either the TO (red), the NN (green), or both models (black).

more skillful than the TO, and blue is where the TO outperforms the NN. Figures 3f and 4f compare how far each model type is away from a perfect reliability of 1 using the metric $\sqrt{(rel_{TO} - 1)^2} - \sqrt{(rel_{NN} - 1)^2)}$. Here, red indicates that the NN is more reliable (closer to 1) and blue is where the TO is more reliable than the NN.

As discussed in Section 3.3, a model is considered skillful if the coefficient of determination is greater than that of a persistence forecast and has a lower-level confidence interval that is greater than zero. A model is considered reliable if a reliability value of 1 is within the lower and upper confidence intervals. These instances of skill or reliability are represented by stippling and dots in Figures 3 and 4. The stippling in Figures 3 and 4a–4d represents where either the TO (a and b) or the NN (c and d) is skillful (a and c) or reliable (b and d). The colored dots represent models that are both skillful and reliable. The different colored dots in Figures 3e, 3f, 4e, and 4f indicate which particular model performs well for the given lag and averaging time, where the TO is represented in red, the NN in green, and the overlap of both models in black.

10.1029/2025JH000669

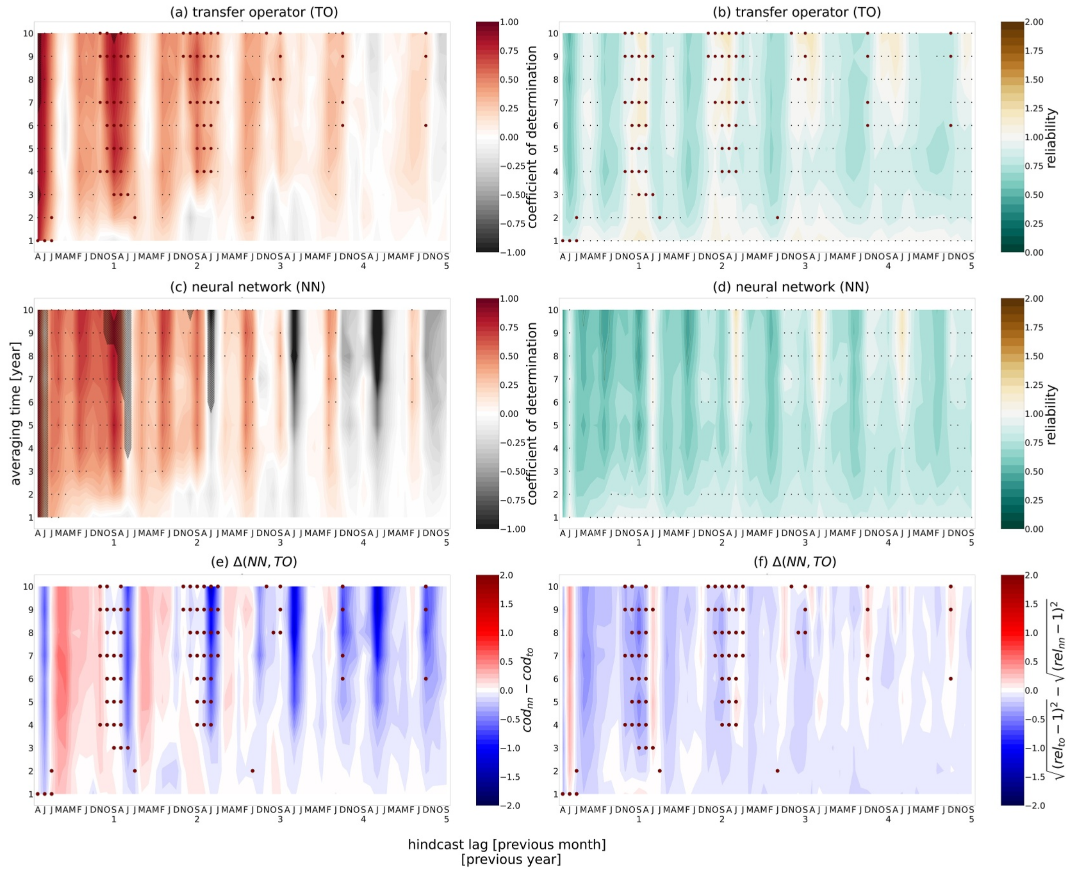

Figure 4. Same as Figure 3, but observations are used as model inputs.

4.1.1. Performance for the Perfect Model Case

We assess the model performances for the perfect model case, where we use testing data and methods as described in Sections 3.1 and 3.3, respectively (Figure 3). The TO and the NN show skill for all cases in the perfect model evaluation; however, only 83% and 11% of these predictions are both skillful and reliable for the TO and the NN, respectively (percentage of lag and averaging times with colored dots in Figures 3a and 3c). Both models have higher skill for longer averaging times and shorter lag times. The skill of both models exhibits seasonality: skill is higher in summer and fall and lower in winter and spring. The TO is reliable for almost all cases, with the exception of some lag and averaging times in July, November, and December. On the other hand, the NN is reliable for a few selected lag and averaging times, which are typically models initialized in winter and spring. The NN reliability also exhibits seasonality: the NN is less reliable during the summer and fall months (Figure 3d).

When comparing the TO and NN (Figures 3e and 3f), we find that for all lag and averaging times, the NN is more skillful (red in Figure 3e) and the TO is more reliable (blue in Figure 3f). This contrast in behavior between the two performance metrics applied to these two model types is expected based on their inherent design. Transfer operators are designed to model the evolution of distributions that reflects the underlying statistics in the data, which lends them toward reliable predictions. Neural networks are designed to capture nonlinear, hidden dynamics within the data, which lends toward a high predictive skill.

4.1.2. Performance on Observations

We assess the model performances when using the observed internal variability to compute hindcast predictions over the 1979–2024 period in Figure 4. For clarity, we begin by discussing the performances of the two model types individually, and follow with comparisons between the two types of models.

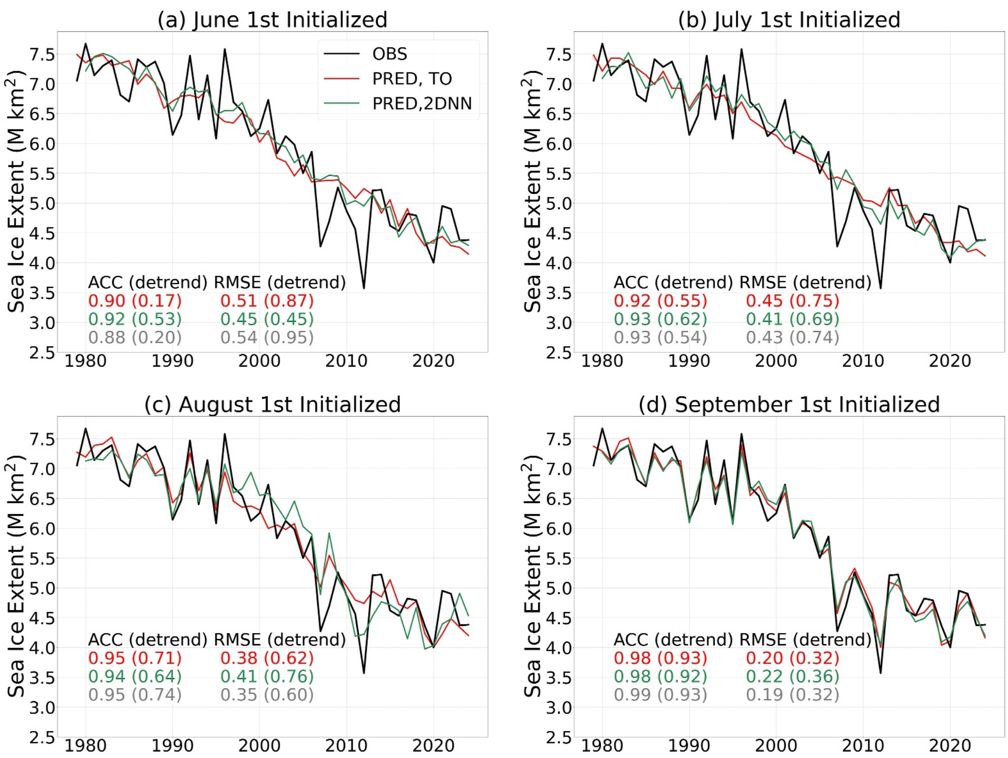

Figure 5. Predictions of September Arctic sea ice extent initialized on (a) June 1, (b) July 1, (c) August 1, and (d) September 1. Predictions from a transfer operator (red) and a neural network (green) are compared to observations (black) for the 1980–2024 period. The corresponding skill metrics (*ACC* and *RMSE*, Equations 3 and 4) are shown in the text, with the detrended skill in parenthesis. We also show these metrics for persistence in gray text. This figure was constructed to be similar to the multimodel analysis in Figure 2 of Bushuk et al. (2024) for ease of comparison.

The TO method is capable of producing skillful and reliable predictions for 13% of the tested lags and averaging times (colored dots in Figures 4a and 4b). There is seasonality in the performance of the TO models. The TO models that are both skillful and reliable are typically those initialized in summer, fall, and winter months (June–November). The TO has lower skill for models initialized in March–May. At longer lag times ($\tau > 3$ years), TO models for July to September transitions tend to lose skill, but well performing TOs are still found in September and December–February for some averaging times. In general, and as expected, the skill is higher for shorter hindcast lags and longer averaging times (red in Figure 4a). Conversely to the skill metric, the TO method tends to be more reliable (reliability closer to 1) for shorter averaging times and longer lag times. The TO generally produces reliable predictions from observations, except for some winter and spring months (e.g., January–March for averaging times greater than 1 year). The TO also shows more skill than the baseline persistence for most cases (minimal hatching in Figure 4a), with exceptions occurring only for long averaging times (9–10 year moving mean) at short lag times (1 month and 1 year).

In contrast to the TO, the NN method does not produce predictions that are simultaneously skillful and reliable for any of the tested lags and averaging times (lack of colored dots in Figures 4c and 4d). While there are cases of skillful or reliable NNs (stippling in Figures 4c and 4d), they do not overlap. Similar to the TO method, the NN exhibits lower skill for longer lag times, and a reliability closer to 1 for longer lag times. In contrast to the TO method, there are several instances where the NN is outperformed by the baseline persistence model (hatching in Figure 4c), particularly for models initialized in June–August. As in the perfect model case, the TO method is typically less skillful (red in Figure 4e), but more reliable (blue in Figure 4f) than the NN when tested on observations. However, there are a few instances where the TO shows more skill and less reliability than the NN.

In summary, the TO alone is able to make both reliable and skillful predictions from the observations, and tends to perform well for summer-to-summer transitions. We will discuss how this is linked to reemergence mechanisms in Section 5.1. Although the NN lacks initialization times that produce predictions with overlapping skill and reliability, it is interesting to note that NNs that predict spring-to-summer transitions (i.e., March–May initializations) exhibit a high skill (R^2). Showing skill in this time frame is unique, as it represents the spring predictability barrier during which models (both numerical and statistical) struggle to make skillful predictions of sea ice. It is likely that the additional input parameter in this NN (i.e., SIAt) contributes to the improved performance in this time period, as it was not found for an NN with SIE as the only input (Figures S4 and S5 in Supporting Information S1). In Section 5, we will discuss the ability of the NN and the TO to produce both skillful and reliable predictions, as well as future features that could be implemented in the NN to improve the model reliability. Lastly, we note that these performance metrics are dependent on the time period over which they are calculated. Here, we use the 1979–2024 hindcast period for observations, but we show skill metrics over different time periods in the following sections.

4.2. Hindcast Predictions

4.2.1. Sea Ice Outlook (SIO) Initialization Months

We now assess the ability of the TO and the NN to predict September SIE when initialized in the summer months (1 June, 1 July, 1 August, and 1 September) to compare to predictions submitted to the Sea Ice Outlook (SIO; Stroeve et al. (2014)). The SIO collects and analyzes predictions from a large international community of polar scientists who use a mixture of dynamical, statistical, and heuristic models to predict the September SIE. We discuss the performance of our statistical models in this section and then compare them to results from a comprehensive study of SIO predictions from Bushuk et al. (2024) in Section 5.2.

Figure 5 shows time series of the observed NSIDC SIE (black) and predictions made by the TO (red) and NN (green) for models initialized from monthly May–August monthly mean values (labeled as June 1st–September 1st initialized). The *ACC* and *RMSE* values for the models are indicated in each panel. We show the performance metrics for the case where the trend is included (i.e., the native observations versus the predicted residuals with the forced component added back in), and for the detrended case (i.e., the observations with the forced component removed versus the predicted residuals). We also show these metrics for anomaly persistence in gray, where to represent the full signal, we persist the monthly anomaly and add back in the September fit.

Both model types exhibit skill for the lead times shown when the metrics are calculated with the SIE trend included; however, the detrended predictions (i.e., the native form of the model predictions) exhibit more considerable performance decreases for increasing lead times. The *ACC* values for the trended predictions are 0.90 or greater for both models at all lead times, whereas the detrended *ACC* values range from 0.17 to 0.93 for the TO and 0.53–0.92 for the NN. The *RMSE* values for the predictions with the trend included are smaller than the standard deviation of the detrended observations (0.51 M km^2) for both models, with the exception of the June 1st initialized case for the TO. The detrended *RMSE* values are only smaller than the standard deviation of the detrended observations for TO and NN predictions initialized on September 1st, and NN predictions initialized on June 1st. For both model types, the performance improves for decreasing lead times. For shorter lead times, the TO makes better predictions, while the NN predictions are slightly more skillful than the TO for models initialized on July 1st and June 1st. Although the *ACC* and *RMSE* of our statistical models are close to those of persistence for August 1st or September 1st initializations, neither model outperforms persistence for these times. However, the NN shows improved performance compared to persistence for the initializations in June and July, and the TO outperforms persistence for predictions initialized in June.

We note that the performance metric used here (*ACC*) is different from the one used during the overall model evaluation in Section 4.1 (R^2). Although our models show a lower *ACC* than persistence when initialized in August and September, the R^2 for these months is higher than persistence. The difference between these two metrics suggests that in comparison to persistence, the models initialized in these months have skill beyond predicting the climatological mean (higher R^2), but struggle with the timing in the variability of the anomalies (lower *ACC*). While these performance metrics suggest the models make skillful predictions, it is evident that difficulties remain in predicting extreme events at lag times greater than one month, as only models that are initialized on September 1st (from the August mean) are able to capture extreme events such as the 2012 minimum (Figure 5). We will further discuss the model predictions of extreme events in Section 4.2.3.

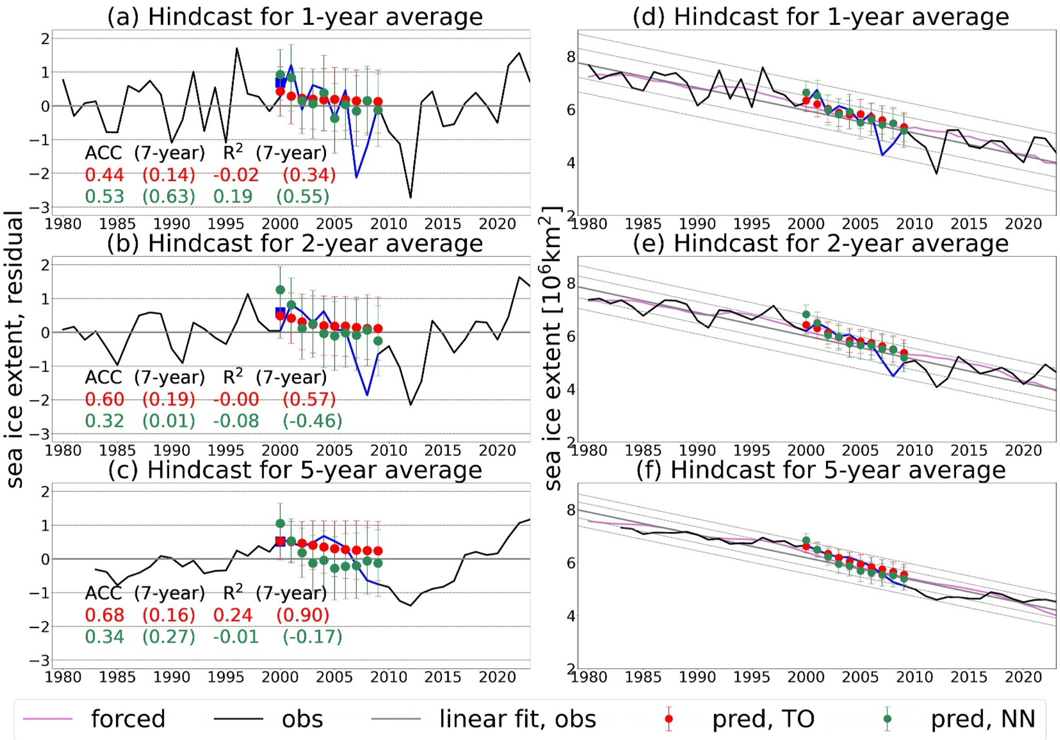

Figure 6. Predictions of the 2000–2009 accelerated sea ice retreat observed through sea ice extent (SIE) anomalies. Observations and predictions of SIE anomalies averaged over (a) 1 year, (b) 2 years, and (c) 5 years. Prediction means and standard deviations are obtained from the transfer operator (red) and the neural network (green) initialized in July 2000 (blue squares). The blue line highlights SIE for the 10-year prediction time frame. Gray lines indicate zero mean and ±1 and 2 standard deviations. (d–f) Same as (a–c) except with the forced term (purple) added back into the prediction. Here, the gray lines represent the linear trend in the observations over the 1979–2014 time period, with ±1 and 2 standard deviations.

4.2.2. Periods of Accelerated Sea Ice Retreat

Next, we test the TO and the NN on predictions of the 2000–2009 period of accelerated sea ice retreat to determine their usefulness for predicting low sea ice conditions that have persisted for several years. The predictions are made from models initialized from the July mean because TO and NN models with this prediction time frame are consistently shown to be reliable (dots in Figures 4b and 4d). Figure 6 shows the detrended and raw SIE from NSIDC observations (black), TO predictions (red), and NN predictions (green) of SIE. The prediction period is represented by the blue line. The forced term that is added back into the predicted residuals of Figures 6a–6c is indicated by the purple line in Figures 6d–6f, and the gray lines represent the linear trend plus or minus one and two standard deviations (this trend is zero for the detrended case in Figures 6a–6c). We show the *ACC* and R^2 performance metrics for the detrended predictions (Figures 6a–6c), where the first number indicates the metric over the 10-year period (indicated by the blue line). The metric in parenthesis is calculated for the first 7-year of this period, when the internal variability in the observations is greater than zero. We show the metric this way because of the inherent behavior of the TO and NN where predictions tend toward zero mean for increasing lag times. In other words, the probability density function of the prediction converges to the total density distribution with zero mean for increasing lag time (Figures S7–S10 in Supporting Information S1). This asymptotic convergence of the forecast distribution to the climatological distribution for long lag times is a typical behavior of forecast systems, and can be indicative of the model's predictability timescale (DelSole & Tippett, 2018).

The coefficient of determination $\left(R^2\right)$ for the predictions over these time periods suggest that for the TO, up to 24% of the annual, 2-year, and 5-year variations are accurately predicted for the 10 year period, and 34%–90% over the 7 year period. There is little difference in these ranges for the NN, which show −1% to 19% of the variations are accurately predicted for the 10-year period, and −46% to 55% for the 7-year period. The TO tends to exhibit a higher R^2 for hindcasts with 2-year and 5-year averaging, while the NN has a higher R^2 for hindcasts

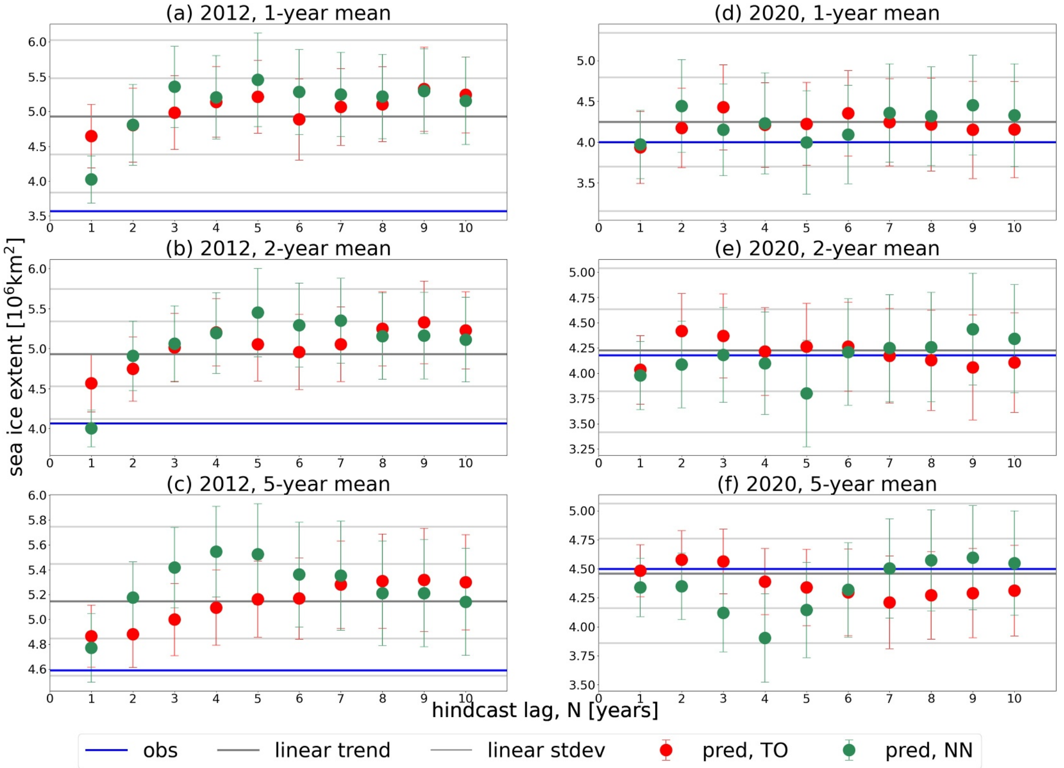

Figure 7. Predictions of September sea ice extent (SIE) for the two lowest years on record, (a–c) 2012 and (d–f) 2020. Predictions are made by the transfer operator (red) and the neural network (green) for moving means of (a and d) 1 year, (b and e) 2 years, and (c and f) 5 years and are initialized in July for hindcast lags of 1–10 years (*x*-axis). Predictions have been postprocessed by adding the forced term back. The blue line represents the observed September SIE. The gray lines represent the SIE for the given year as predicted by the linear trend in the observations over the 1979–2014 time period, with ±1 and 2 standard deviations.

with 1-year averaging (Figure 6). This 2000–2009 accelerated sea ice retreat is within one standard deviation of the predictions made by both the TO and the NN for only the 2000–2006 period for hindcasts with 1- and 2-year averaging, and the 2000–2008 period for the 5-year average. However, we note that the 2007 SIE represents an extreme minimum, which presents difficulties for predictions of this nature.

We also find that the predictions of the TO and the NN are able to capture the 2012–2021 slowdown in sea ice decline (Figure S6 in Supporting Information S1), which remains within one standard deviation of the TO and NN predictions for 2013–2020 and 2012–2021, respectively. Both models perform quite well during this prediction time: the TO exhibits an *ACC* from 0.72 to 0.97 and an R^2 ranging from 0.24 to 0.52, while the for the NN, *ACC* ranges from 0.56 to 0.92 and the R^2 ranges from 0.30 to 0.85, depending on averaging time. We note that the ability of the models to capture both the 2000–2009 accelerated decline in SIE and the 2012–2021 slowdown in sea ice decline maybe linked to the inherent behavior of the models toward asymptotic convergence to the mean over long prediction lag times (Figures S7–S10 in Supporting Information S1).

4.2.3. Extreme Events: Case Studies for 2012 and 2020

We assess the ability of the TO and NN models to predict extreme events during individual years. Figure 7 shows TO (red) and NN (green) predictions for the two lowest September SIE years on record in the observations (blue line): 2012 (Figures 7a–7c) and 2020 (Figures 7d–7f). The gray lines represent the linear trend and the first two standard deviations. Predictions are from the 1-year, 2-year, and 5-year averaged models initialized with the July mean for hindcast lags of 1–10 years prior to 2012 and 2020. We show the full SIE for the events in Figure 7, and have added the forced term back in to the predicted residual. The predictions for the third and fourth lowest September SIE years on record (2007 and 2019) are shown in Figure S11 in Supporting Information S1.

The 2012 event represents the lowest September SIE on record, and also has the lowest residual from the forced trend (Figure 1c). For a 1-year lag and 2- and 5-year moving mean, the NN is able to predict the 2012 SIE within one standard deviation. Additionally, TO and NN models initialized in July 2012 predict a negative SIE anomaly and a high likelihood of low SIE for September 2012 (Figure S12 in Supporting Information S1). For lags greater than 1 year, both the TO and the NN are unable to predict the 2012 SIE within one standard deviation. The TO predictions are closer to observations than those from the NN, particularly for lead times less than 9 years.

The 2020 September SIE is the second lowest recorded in the observations. However, the residual from the forced term for the 2020 event is not as low as that for the 2012 event because the forced term itself continues to decrease after 2012. The TO and NN are able to reproduce this minimum within one standard deviation for almost all averaging times and lag times (Figures 7d–7f), with the exception of predictions made using the NN for the 3- and 4-year lag times with the 5-year moving mean. The mean predictions from the TO and the NN are the same as the observed value in 2020 for the case of 1-year average and 1-year moving mean. For lags greater than 1 year, we do not necessarily see increasing divergence from the observations. Furthermore, the TO and NN models both predict a high likelihood that 2020 September SIE is lower than the climatological mean of the natural variability for 1- and 2-year moving means (Figure S13 in Supporting Information S1).

The difference between the model's ability to reproduce the 2020 versus the 2012 extreme suggests that the model predictions are dependent on the initial conditions, which are in the form of standardized residuals. Models tend to be better at representing sea ice conditions that exhibit a less extreme residual from the forced term, and therefore exhibit difficulties in capturing extreme events. This type of model behavior will be further discussed in Section 4.4.

4.3. Predictions for 2024–2028

After thoroughly analyzing the performance of the models in hindcast predictions, we turn to making real-time forecasts of the September SIE for the 2024–2028 period using different TO and NN models. We use the July-initialized models for predictions, and therefore cannot yet make predictions initialized in 2025 (July 2025 has not happened yet). Nevertheless, we have shown that the July-initialized TO can still make reliable predictions at 5-year lead times (Figure 4). Figure 8 shows forecasts for the TO for 1-, 2-, and 5-year averages (rows). This forecast is initialized from the July 2024 SIE. The first column (Figures 8a–8c) shows the forecast trajectory for the predicted interannual variability over the next five time steps. The second column shows the probabilistic forecast, and indicates whether the forecast for September SIE will be moderately ($\pm 1\sigma$), intensely ($\pm 2\sigma$), or extremely ($\pm 3\sigma$) higher or lower than the climatological mean of the natural variability. These intensity classifications are further represented in the final column (the background colors), where we show the difference between the predicted and climatological distributions (red minus gray bars in Figures 8d–8f). We note that these three models (1-, 2-, and 5-year lag and averaging times with July initialization) were all deemed reliable for predictions over the 1979–2024 time period (red dots in Figure 4a). While there is skill for only the model at a 1-year lag and averaging time, we show these future predictions in the form of probabilistic forecasts and therefore are more concerned with reliability than skill.

For all three averaging times, the TO shows a high likelihood of higher SIE (Figure 8d). The TO predicts that the 2024 *SIE* has an 89.5% probability of being higher than what is predicted by the forcing alone (i.e., the climatological mean of zero). Based on the intensity classifications, 2024 will likely exhibit a moderately high September SIE (within one standard deviation of the mean). This is further supported by Figure 8g, which confirms that high SIE has the largest change of occurrence compared to climatology (red bars in Figure 8g), while low SIE has an occurrence decrease with respect to climatology (blue bars in Figure 8g). The prediction of September 2024 SIE is confirmed by observations from NSIDC $\left(SIE_{SEP,2024} = 4.35 \text{ Mkm}^2\right)$, which represents a standardized residual SIE that is greater than the climatological mean (i.e., $+0.91$).

Additionally, the TO shows a higher likelihood of higher September SIE for the 2024–2025 and 2024–2028 periods. The TO predicts that September SIE has 89.9% and 85.7% chance of being higher than what is predicted by the forcing for the 2024–2025 and 2024–2028 periods, respectively (Figures 8e and 8f). Moderately positive anomalies in SIE show the highest likelihood in compared to climatology for the 2024–2025 and 2024–2028 periods, respectively (Figures 8h and 8i). We note that this positive anomaly is relative to the continued

10.1029/2025JH000669

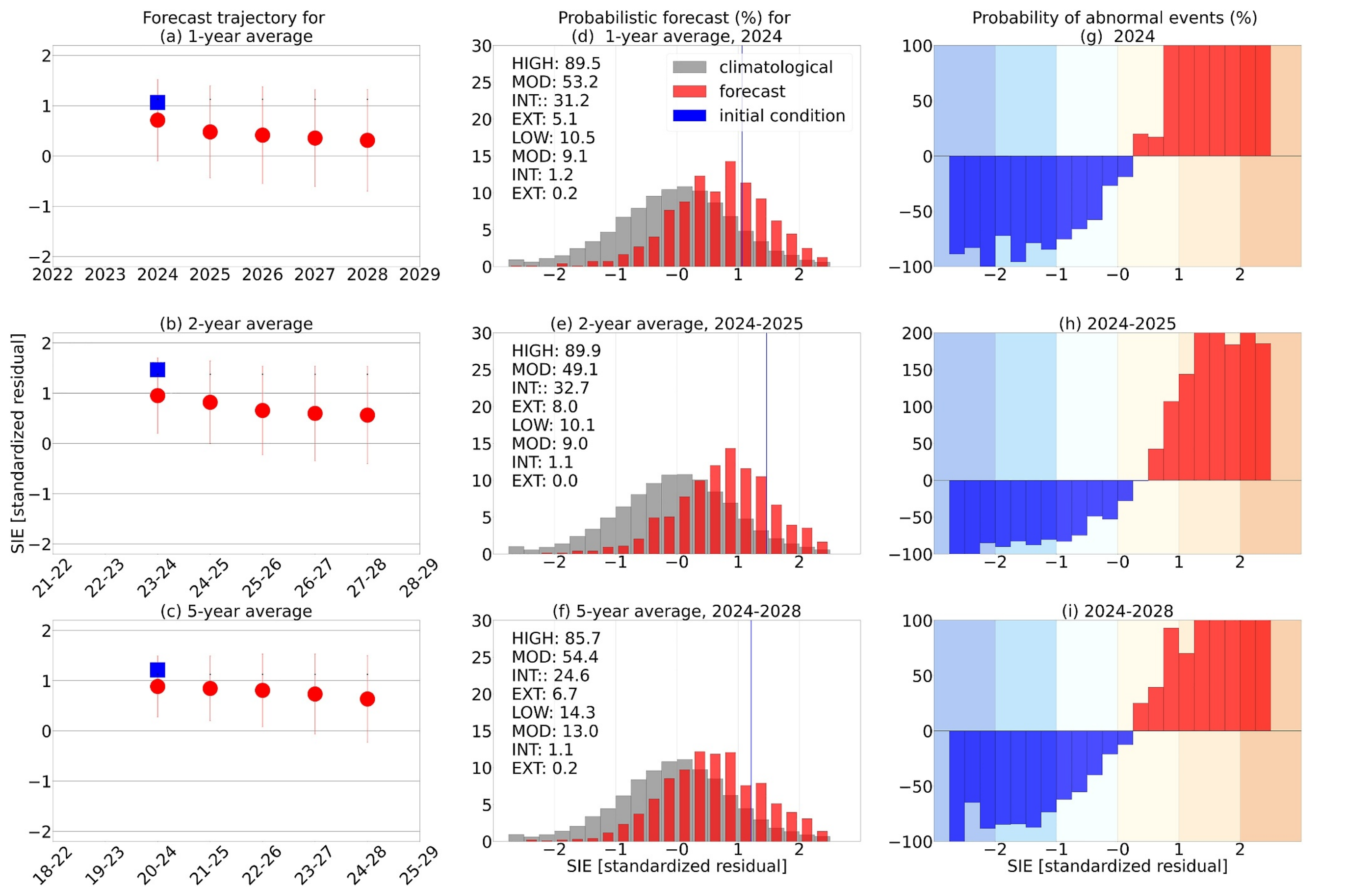

Figure 8. Interannual probabilistic forecast of standardized September sea ice extent (SIE) anomalies for 2024–2028. Predictions are made using a transfer operator for averaging times of (a, d, and e) 1 year, (b, e, and h) 2 years, and (c, f, and i) 5 years. (a–c) Mean and standard deviation predictions of September SIE anomaly for 5 years (red) initialized from the July anomaly (blue square). (d–f) Predictions of SIE distribution (%, red) for 0, 1, and 4 years in advance with respect to July 2024. The gray region represents the climatological distribution of the anomaly, the blue line represents the initial condition from July 2024, and the gray lines show the climatological mean and ±1 and 2 standard deviations. Text indicates the percentage of the predicted distribution that is higher (HIGH) or lower (LOW) than the climatological mean, with further divisions of moderate (MOD), intense (INT), and extreme (EXT) indicating 0–1, 1–2, and 2+ standard deviations away from the climatological mean, respectively. (g–i) Distribution of probability changes with respect to the climatological distribution, where background colors are consistent with moderate, intense, and extreme events in (d–f) following Sévellec and Drijfhout (2018).

downward trend in the forced component, and does not represent an increase in the full SIE signal, just the internal variability.

We also make predictions of the full September SIE signal from the TO and the NN initialized in July for 1-, 2-, and 5-year averaging periods (Figure S14 in Supporting Information S1). While both models are reliable for these lag and averaging times, they exhibit skill only for the 1-year lag and 1-year moving mean (first points in Figure S14a in Supporting Information S1). Deterministic predictions past this point should therefore be interpreted with caution. However, the TO and NN models predict a September 2024 SIE of 4.30 ± 0.47 Mkm^2 and 4.13 ± 0.54 Mkm^2, respectively (Figure S14 in Supporting Information S1), compared to the observed value of 4.35 Mkm^2 from the NSIDC. Both predictions capture this observed value well within one standard deviation.

4.4. Model Behavior

This section aims at understanding the behavior of the two model types, with an emphasis on how the information in the inputs is used by the models to predict September SIE. Predictions made by the TO and NN models are based on the input conditions, which are in the form of standardized residuals. The TO behavior is straightforward and as expected (Figure S15 in Supporting Information S1): negative initial conditions for SIE tend to produce negative predictions, and positive initial conditions for SIE lead to positive predictions. The larger the magnitude

10.1029/2025JH000669

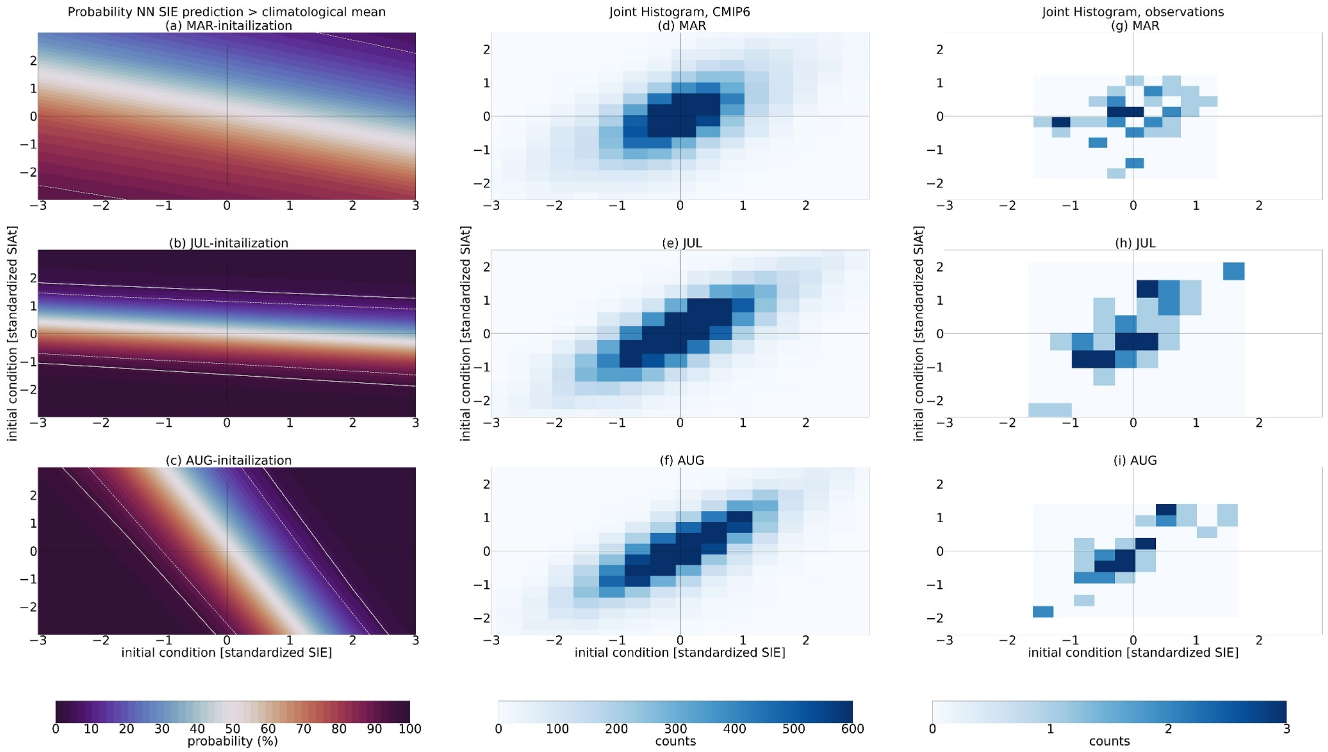

Figure 9. (a–c) Probability that the NN will predict September sea ice extent (SIE) to be lower than the climatological mean for given initial conditions. Probabilities are shown for (a) March, (b) July, and (c) August-initialized predictions from the neural network for a 1-year moving mean and 1-year lag time. The solid and dashed white contour lines show the upper and the lower 95th and 90th percentiles, respectively. (d–i) Joint histograms of the input conditions of SIE and the SIAt from (d–f) the CMIP6 training data and (g–i) observations.

of the input condition, the farther the probability is from 50%. For example, an initial condition of −3.0 shows that the TO has almost 100% chance of predicting that SIE will be lower than the climatological mean (Figure S15a in Supporting Information S1). This response is exaggerated with increasing the time length of the moving mean, where the white lines indicating 95% chance of a prediction either lower or higher than the climatological mean cover a larger range of initial conditions and lag times (Figure S15 in Supporting Information S1). We also see that for increasing lag times, the predictions have a more balanced chance of being lower or higher than the climatological mean for any input condition. This suggests that the prediction converges to the climatological distribution for longer lag times (as discussed in Section 4.2.2 and confirmed in Figures S7–S10 in Supporting Information S1), which is a typical behavior of forecasts (DelSole & Tippett, 2018).

Figure 9 shows the sensitivity of the NN predictions to the initial conditions, in the form of the standardized residual of the inputs (SIE and SIAt). Results are shown for the 1-year average, the 1-year lag NN initialized in March (Figure 9a), July (Figure 9b), and August (Figure 9c). The initial SIE and SIAt conditions ranges from −3 to 3, which represents ±3 standard deviations from the mean. The colors represent the probability that the model will predict September SIE to be lower than the expected forced response (i.e., zero) for a given initial condition. Red indicates that there is a greater than 50% probability that the given input condition will produce a prediction that is lower than the climatological mean, and blue shows that this probability is lower than 50%. The solid and dashed white lines represent the 95% and 90% probability that the NN will predict a value less than (overlapping red) or greater than (overlapping blue) the climatological mean. We also show the joint histograms of SIE and SIAt conditions from the training data in CMIP6 (Figures 9d–9f) and from observations (Figures 9g–9i), to provide a reference for when the tested initial conditions occur in the data.

The behavior of the NN is slightly more complex than that of the TO because it has two input variables (SIE and SIAt) that control the model predictions. Figures 9a–9c shows the March-, July-, and August-initialized NN behavior as a function of these two inputs for a 1-year moving mean and lag time. For the March and July-initialized models (Figures 9a and 9b), the NN seems to be more sensitive to SIAt than SIE. For example, for

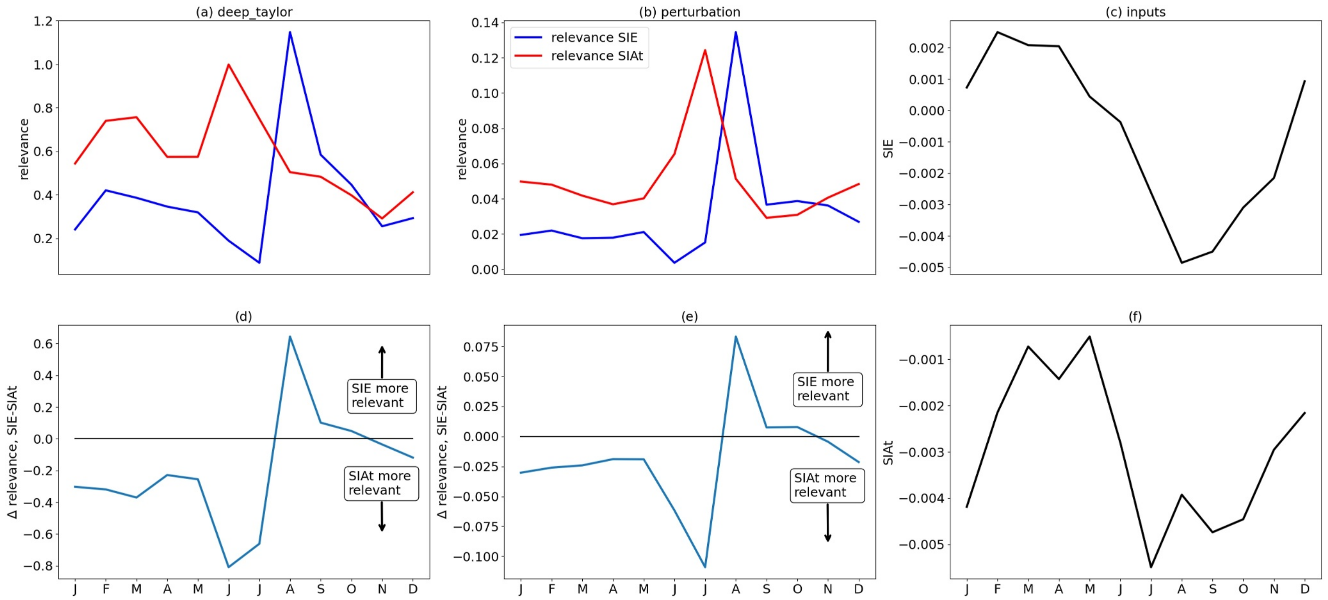

Figure 10. The relevance of each of the input features (sea ice extent (SIE) and SIAt) to the NN for predicting September SIE has the highest probability of occurring in the correct bin. These relevance values are calculated for the 1-year lag time and 1-year average models initialized in January–December (x-axis). Relevance is shown for the (a and d) deep Taylor decomposition and (b and e) perturbation methods. The first row shows the overall relevance for both input features (a and b), while the second row shows the difference in relevance between the two features (d and e). (c and f) The seasonal cycle of the standardized residual (c) SIE and (f) SIAt from CMIP6 training data.

an input SIAt of 2, the NN always has a less than 50% probability of predicting September SIE to be lower than the climatological mean, no matter what the input SIE. On the other hand, for any particular SIE, the probability of predicting September SIE to be lower than the climatological mean changes with SIAt. This behavior is more exaggerated for the NN initialized in July than in March. Conversely, the August-initialized NN is more sensitive to SIE (Figure 9c). For example, for an input SIE of 2, the NN always has a less than 50% probability of predicting September SIE to be lower than the climatological mean, no matter what the input SIAt is. On the other hand, for any particular SIAt, the probability of predicting September SIE to be lower than the climatological mean changes with SIE. To explore this behavior further, we apply explainable AI (XAI) methods to the NN in the form of a perturbation analysis and a deep Taylor decomposition (Figure 10).

Figure 10 shows how relevant each of the input features is for the NN in predicting September SIE for two different explainable AI methods (deep Taylor decomposition in Figures 10a and 10d; perturbation in Figures 10b and 10e) using models with 1-year lag and averaging times initialized in January–December (x-axis). The deep Taylor decomposition method shows the relevance of each input for predicting September SIE falls in the correct bin, while the perturbation method shows relevance in terms of the sensitivity of the overall SIE prediction to perturbations in each of the input features. We show the absolute relevance of each input (Figures 10a and 10b) and the difference between the relevance of the two inputs (Figures 10c and 10d). A negative difference indicates that SIAt is more relevant than SIE, and a positive difference indicates SIE is the more relevant input feature. For comparison, we show the seasonal cycle in the standardized residual of each input from the CMIP6 training data (Figures 10c and 10f). We calculate p-values (for deep Taylor decomposition) and 95% confidence intervals (for perturbation), which show that the relevance values for each input feature are statistically significantly different.

The outputs from both XAI methods show similar results: SIE and SIAt are both more relevant for models initialized in summer months, when SIE and SIAt are low (Figure 10a). Taylor decomposition shows that SIE relevance peaks in August and SIAt relevance peaks in June, while the perturbation method shows that SIE relevance peaks in August and SIAt relevance peaks in July (Figures 10a and 10b). We find that whether SIE or SIAt is the most relevant input feature depends on the initialization month (x-axis), but not the XAI method (deep Taylor vs. perturbation). For both XAI methods, SIE is the most relevant predictor only for models initialized in August–October (Figures 10d and 10e). Otherwise, SIAt is the most relevant input feature. These two methods agree that SIE is most relevant in the summer months, when SIE and SIAt are low (August–October), and when both input features show a high relevance for both methods (Figures 10a and 10b). The input feature SIAt is the most relevant in July for the perturbation method and June–July for Taylor decomposition (Figures 10d and 10e).

These XAI relevance results are in agreement with results from behavioral studies with initial conditions (Figures 9a and 9b), which indicate that SIAt is the most relevant input feature in March and July (i.e., for a given SIAt, changes in SIE do not change the likelihood that the NN will predict September SIE to be lower than the climatological mean in Figure 9b) and the least relevant feature in August (i.e., for a given SIE, changes in SIAt do not change the likelihood that the NN will predict September SIE to be lower than the climatological mean in Figure 9c). We present these results with the caveat that the two input features (SIE and SIAt) are highly correlated, which could lead to difficulties in disentangling their comparative relevance when applying explainable AI methods (Flora et al., 2024).

5. Discussion

We highlighted the performance and behavior of two types of statistical models (TO and NN) for predicting September SIE. We move forward with a discussion of how these results are related to historical findings on the predictability of September SIE, and how they compare to predictions made by other models in the literature.

5.1. How Skillful Are Statistical Model Predictions of September Arctic SIE?

Monthly mean Arctic SIE is considered to have statistically significant potential predictability continuously for 1–2 years, and intermittently for up to 4 years (Blanchard-Wrigglesworth, Bitz, & Holland, 2011; Day et al., 2014; Guemas et al., 2016; Tietsche et al., 2014) based on estimates from dynamical models. The predictability limit of SIE is related to persistence and reemergence mechanisms. Blanchard-Wrigglesworth, Armour, et al. (2011) highlighted two mechanisms: the reemergence from melt to freeze associated with persistence of ocean heat content anomalies, and the freeze-to-melt reemergence associated with persistent thickness anomalies. The first gives rise to increased predictability from winter preconditioning, while the second gives rise to summer-to-summer reemergence of predictability. These mechanisms are linked to the spring predictability barrier (Day et al., 2014; Tietsche et al., 2014) whereby correlations between monthly and September SIE are highest for June–August, drop substantially during the springtime melting months (March–May), and then reemerge during the winter and fall months. Additionally, Chevallier and Salas-Mélia (2012) discuss a winter preconditioning mechanism, whereby the area covered by relatively thick sea ice (SIT $>$ 0.9–1.5 m) is a better predictor of summer SIA than is the total SIA. For example, Chevallier and Salas-Mélia (2012) found that the March SIAt (SIA where SIT $>$ 0.9 m) explains 30% of the variability in September SIA, and June SIAt explains 76% of the variability in September SIA. We discuss how these mechanisms are apparent in the performance of our TO and NN models.

Our performance analysis includes both skill (coefficient of determination) and reliability metrics for a more complete evaluation of the usefulness of the model predictions. We show that the TO and the NN are able to predict state transitions with a higher skill than anomaly persistence for most cases. Both models tend to produce reliable predictions. However, the TO is generally more reliable than the NN, and the TO is the only model to make predictions from observations that have overlapping skill and reliability. Our analysis shows the usefulness of these models for making seasonal to interannual predictions of Arctic SIE. However, we note that both model types, unsurprisingly, have difficulty capturing extremes.

For the perfect model case, both models show overall decreases in skill and increases in reliability for increasing hindcast lags on yearly timescales. Additionally, both TO and NN models exhibit seasonality for the reliability and skill metrics. Both the TO and the NN produce skillful predictions for all initialization months, with the most skillful predictions coming from models initialized in summer (June, July, and August) months. The TO predictions are less skillful during all other seasons, but the NN only loses a bit of skill for the spring (March, April, and May) initializations. The TO is generally reliable for all initialization months. However, the NN tends to produce unreliable predictions for the perfect model case, except for the case of models initialized in spring and winter months. The decrease in predictive skill for models initialized in spring months (March–May) reflects the spring predictability barrier for sea ice (Day et al., 2014). Additionally, the rise in correlation for summer months at hindcast lags greater than 1 year is characteristic of the summer-to-summer reemergence (Blanchard-Wrigglesworth, Armour, et al., 2011).

For models run with the observations as inputs, the TO is reliable for most cases, with slightly less reliability for winter (January and February) initialized models. The TO produces more skillful predictions in the summer and fall months, and shows no skill for spring initialized models. This pattern is largely consistent with the spring predictability barrier, where there is a substantial loss of predictive skill for models initialized in spring that reemerges for models initialized in summer (Bushuk et al., 2020; Day et al., 2014; Tietsche et al., 2014). The NN is generally less reliable than the TO, and tends to produce reliable predictions from observations only during the time where the predictions are not also skillful. Conversely, the NN tends to be more skillful than the TO, except during summer months. Interestingly, the NN with two inputs (SIE and SIAt) shows skill for springtime initializations at 1- and 2-year lag times. This is not the case for the TO or the NN with only one input (Figures 4 and S2 and S3 in Supporting Information S1), indicating SIAt is an important predictor for SIE. This provides evidence for the mechanism of winter preconditioning of summer sea ice cover from Chevallier and Salas-Mélia (2012), and represents a unique case of overcoming the spring predictability barrier.

5.2. How Well Do Models Perform for Hindcast Predictions?

The skill and reliability of the TO and NN models suggests that they are useful for probabilistic forecasts. The reliability of the TO and the NN also suggests that the CMIP6 models used to train these models can capture the statistical patterns found in the observations, and are therefore useful for understanding Arctic sea ice as a part of the climate system despite their inherent limitations. However, we note that our model predictions are not reliable for all initialization months or hindcast lags, which could be further indicators of where the models used from CMIP6 are limited in representing observed Arctic SIE. We discuss the TO and NN performances for predicting September SIE for the initialization months used by the SIO, for hindcasts of the 2000–2009 slowdown in SIE decline, and for the prediction of the extreme low SIE events in 2012 and 2020.

We apply TO and NN models to predict historical hindcasts of September SIE based on initializations in June, July, August, and September of the same year (Figure 5). We present this information in a manner that is consistent with Figure 2 in Bushuk et al. (2024) for ease of comparison with the various dynamical and statistical models that were evaluated by these authors. For June–September initialization, the TO and NN predictions seem to fall within the interquartile range of the models tested in Bushuk et al. (2024).

We also compare the performance metrics (*ACC* and *RMSE*) shown in Figure 5 to those in Figure 3 of Bushuk et al. (2024). We use data provided by Bushuk et al. (2024) to reproduce their Figure 3, with the results of our model performances overlaid (Figure S17 in Supporting Information S1). Our TO and NN models show a similar *ACC* to both persistence and the multimodel analysis of Bushuk et al. (2024) for predicting the full signal, but are lacking in performance for the detrended *ACC*. The TO has a higher *ACC* than persistence for the full signal and detrended case for June 1st initializations, and the NN outperforms persistence for June and July 1st initializations (Figures S17a and S17b in Supporting Information S1). The *RMSE* for our TO and NN predictions is higher (worse) than the multimodel analysis of Bushuk et al. (2024), with the exception of the June 1st initialized model in the full and detrended case (Figures S17c and S17d in Supporting Information S1). Additionally, both the TO and the NN have a lower (better) *RMSE* than persistence for June 1st initialized models in the full and detrended case.

We note that our models make predictions from monthly mean inputs, whereas many of the models in Bushuk et al. (2024) use daily inputs. Hence, our prediction for the June 1st initialized case (Figure 5a) is actually made using the models for predicting May-to-September transitions, which we show to have poor performance relative to other initialization months. Additionally, we calculate the performance metrics over the 1980–2024 period, whereas Bushuk et al. (2024) show them for the 1993–2021 period in Figure 2 and the 2001–2020 period in Figure 3. These factors could lead to differences in the performance metrics discussed above.

We also show that the TO and NN models are reasonably skillful for predicting the 2000–2009 accelerated sea ice decline (Figure 6) and the 2012–2021 slowdown in sea ice decline (Figure S6 in Supporting Information S1). The coefficient of determination is typically higher when analyzed over the first 7 years of the 2000–2009 accelerated sea ice decline than over the entire 10-year period. This is because the model predictions tend toward the climatological mean at longer hindcast lags, and the latter years of this accelerated sea ice decline include the 2007 extreme low SIE, which models struggle to reproduce (Figure 7). However, we note that the ability of the models to capture either the accelerated decline (2000–2009) or the slowdown in sea ice decline (2012–2021) could be linked to the inherent behavior of the model predictions to converge toward the climatological mean at increasing lag times (DelSole & Tippett, 2018).

Table 1
Likelihood of Anomalously Low/High September Sea Ice Extent for the Next Decade

Forecast period	TO prediction (low/high)	NN prediction (low/high)
1 year: 2024	11%/**89%**	4%/**96%**
2 years: 2024–2025	10%/**90%**	5%/**95%**
3 years: 2024–2026	7%/**93%**	35%/**65%**
4 years: 2024–2027	11%/**89%**	**58%**/42%
5 years: 2024–2028	14%/**86%**	**56%**/44%
6 years: 2024–2029	16%/**84%**	50%/50%
7 years: 2024–2030	28%/**72%**	**73%**/27%
8 years: 2024–2031	20%/**80%**	**72%**/28%
9 years: 2024–2032	32%/**68%**	**66%**/34%
10 years: 2024–2033	25%/**75%**	**63%**/37%

Note. Bold text indicates if SIE is more likely to be higher or lower than the climatological mean.

Although the models have limitations in reproducing extremes, their predictive performance for such events is influenced by the magnitude of the input values, which are in the form of a standardized residual that depends on the definition of the forced term (i.e., the multimodel ensemble mean from CMIP6). The TO and NN models are better at representing sea ice conditions that are closer to the forced term. However, both the TO and the NN predicted that 2012 and 2020 SIE would likely be lower than the climatological mean of the forced trend (Figures S12 and S13 in Supporting Information S1).

5.3. What Predictions Are Made for Future September SIE?

We also used the TO and NN initialized in July to make interannual predictions of future Arctic SIE (Table 1). We find that predictions from the TO show that September 2024 has an 89.5% probability of a high SIE anomaly, which is supported by the decreased likelihood of a low SIE. For the 2024–2025 period, the TO predicts an 89.9% chance of anomalously high SIE, and on longer timescales, an 85.7% chance of anomalously high SIE over the 2024–2028 time period. The July-initialized NN similarly shows a high chance of a high SIE anomaly for September 2024 (96% chance) and the 2024–2025 period (95% chance). Conversely, the NN predicts that there is a 56% chance of anomalously low SIE over the 2024–2028 time period.

5.4. How Do the Input Conditions Impact Predictions of September SIE?

We find that both the TO and the NN are sensitive to the input conditions. The TO tends to predict a September SIE lower than the climatological mean for negative initial SIE conditions, and higher than the climatological mean. This reflects a behavior of persistence. The NN has two input features (SIE and SIAt), and the predictions are sensitive to changes in both of these. We find that the relative importance of the two input features is dependent on the model's initialization month. The SIAt input is typically more relevant for models initialized in December–July, while SIE is more relevant to models initialized in August–November. Explainable AI methods also show that the SIAt is more relevant than SIE for models initialized in spring months (March–June), which could explain why the NN can produce skillful and reliable predictions during the spring predictability barrier where the TO fails to skillfully predict September SIE. We note that outputs from explainable AI maybe influenced by the fact that the two input features (SIE and SIAt) are highly correlated, and suggest that this relationship be teased out in future work for a better understanding of the comparative relevance of these two input features.

5.5. What Insights Did We Obtain About Model Structure and Behavior?

5.5.1. The TO Model Performs Better With a Single Input (SIE), Whereas the NN Benefits From the Inclusion of an Additional Input (SIAt)

For both model types, we train an array of models to predict September SIE for various averaging times (1–10 years), hindcast lags (1–10 years), and initialization months. The TO is designed to predict state transitions in SIE, and the NN predicts state transitions in SIE based on SIE and SIAt at the previous state. The NN predictions benefit from additional information for certain initialization months, hindcast lags, and averaging times (i.e., the 2D–NN can sometimes outperform the 1D–NN, Figures S4 and S5 in Supporting Information S1). On the other hand, the TO works better if there is only one input. This is because the TO is built by creating bins of the input and output states. Each additional input predictor exponentially increases the number of input bins. This decreases the number of input states in within each bin, creating bins that do not have a reliable number of data points to obtain a full understanding of the transition from one state to the next. Interestingly, Bushuk et al. (2024) find that statistical models that use only sea ice predictors tend to outperform those that also include atmospheric and oceanic inputs. Including additional predictors, even those that are physically relevant, may lead to overfitting and hinder the predictive skill.

5.5.2. The Model Predictions Depend on the Initial Conditions, Which Are in Turn Influenced by the Definition of the Forced Term

We are also interested in the behavior of the models as a function of different initial conditions. In Figure 9, we show that models are more likely to have predicted distributions that are greater than the climatological mean for positive inputs, and lower than the climatological mean for negative inputs. We clarify that these results are based on arbitrary values of inputs that lie with the range from which the statistical models are trained to make predictions, and are not specifically linked to SIE observations, although the values do lie within the distribution of observed SIE conditions.

The initial conditions taken from observations of SIE are in the form of a standardized residual, which is dependent on how the forced term is defined. Therefore, results from Section 4.4 indicate that the TO predictions of SIE are sensitive to the definition of the forced term. We tested several different definitions of the forced term (not shown), including the linear trend in observed SIE, a fit between observed SIE and global mean temperature, a combination historical (1850–2014) and SSP5-8.5 (2015–2100) CMIP6 simulations. Using these terms to define the forcing did not lead to substantial differences in the performance or predictions of TO and NN models. However, in one case, we applied forcing as a combination of historical CMIP6 simulations (1850–2014) and a T + 5-year moving mean of the observations (where T is the moving mean that the model in question is trained on). For this case, we found that the forcing term did not have a continued downward trend, and leveled out in response to the 2012–2024 slowdown in sea ice decline. The model performance and predictions with this forced term differed substantially from the others (not shown). We opted to use the forcing that exhibited a continued downward trend because it is representative of the known interplay between SIE and global mean temperature, and our goal was not to predict this trend, but the interannual variability around it.

5.5.3. TO Models Yield More Reliable Predictions, While NNs Generate More Skillful Ones

We find that TO predictions tend to be more reliable, while NN predictions are more skillful. As discussed in Section 4.1, this is expected because of the nature of these two model types. Transfer operators are specifically designed to represent the evolution of the probability distribution over time, inherently prioritizing reliability due to their probabilistic framework. On the other hand, neural networks are designed to learn patterns in the data, and achieve skill because of their ability to capture information about nonlinear interactions in the system. This reflects a general trade-off between model complexity and reliability: simpler models tend to be more conservative in their predictions, while more flexible models like neural networks often require explicit regularization or calibration to avoid overconfidence. In this study, we implement features in the NN to improve its function in making probabilistic predictions (i.e., a softmax layer and the KL divergence loss function), but other methods exist for further improvement. Future work could explore ways to modify the NN architecture to incorporate constraints or regulation terms that allow them to conserve the statistical dynamics of the data. Some possible methods include introducing a loss function to penalize deviation from the probability distribution during training, implementing the NN in the form of a Bayesian neural network that explicitly models uncertainty, or combining the TO and the NN by using the learned TO to constrain a loss function in the NN training. Additionally, a useful extension of this work could be to employ a simpler probabilistic model, such as logistic regression, using the same inputs and outputs as our NN (i.e., inputs of SIE and SIAt; output as a probability distribution of SIE). While we expect logistic regression maybe less skillful than an NN due to its limited capacity to capture nonlinear relationships, it may offer improved reliability.

6. Conclusion

There is an increased need for skillful predictions of Arctic sea ice as it faces rapid decline. The relationship between SIE and greenhouse gas emissions lends to partial predictability of the downward trend; however, uncertainties remain in predicting the internal variability of Arctic SIE that manifests most clearly on seasonal to interannual timescales. The interannual timescales have been given lower attention than the seasonal and multidecadal timescales. Dynamical model-based predictions face known issues for prediction at interannual timescales, including initialization shocks, drift, and development of transient biases. Data-driven approaches offer an alternative pathway to explore predictive skill. In this work, we train statistical models (TO and NN) to predict variability in September SIE on interannual to decadal timescales using output from CMIP6 simulations. Overall, we find that statistical models can make skillful and reliable predictions of Arctic SIE with high computational efficiency. However, we do note that there are many cases in which these metrics do not overlap, and a model is alone either reliable or skillful. Additionally, TO predictions show that the anomalously high September SIE is likely to continue for the next decade (Table 1), while NN predictions show a likely transition to a state of anomalously low September SIE within the next 4 years. These anomalies are relative to the forced trend, which exhibits a projected future decline in SIE. We also show that SIAt is an important predictor of SIE, highlighting the need for improved measurements of sea ice thickness.

We recommend the use of transfer operators as efficient, skillful, and reliable statistical models for making probabilistic predictions of state changes in Arctic sea ice. We highlight the unique ability of neural networks that include information about the area covered by thick ice to overcome the spring predictability barrier. However, we note that in future work, we would improve upon the neural network design by using a learned transfer operator to constrain a loss function during training, which would likely improve the reliability of forecasts made by neural networks. In future work, we plan to further explore using neural networks trained to predict interannual to decadal Arctic September SIE from maps of properties related to sea ice, the atmosphere, and the ocean. We will also apply explainable artificial intelligence (XAI) methods to these models to understand the relevance of the various model inputs for predicting SIE and gain insight into the drivers of variability for predicting September SIE in the Arctic.

Appendix A: Neural Network Architecture

We provide an overview of the data used for training and testing of the transfer operator (TO) and neural network (NN) models used in this study in Table A1. The TO and NN are trained using data from CMIP6 historical simulations, and are tested using observations from NSIDC and PIOMAS. For both the TO and NN, the output is a probability distribution of September SIE. For the TO, the input is SIE from a given month; for the NN the input layer receives both SIE and SIAt for a given month. Table A1 also provides an overview of the NN architecture and parameters. The input layer is fed to two hiddn layers, each containig 10 nodes with ReLU activation ($\alpha = 0.01$). The output layer contains 24 nodes (bins of SIE) and a softmax function that remaps the output values so that they sum to one and represent the confidence for the prediction that SIE falls into each bin. The model is trained to optimize the KL divergence loss function using an adam optimizer and L2 regularization ($\lambda = 0.01$). The model is run for 10 epochs with a batch size of 200.

Table A1
Details About Training and Testing Data for the NN and TO Models, and Model Architecture (i.e., Parameters and Hyperparameters) for the NN

Training data	Inputs	Sea ice extent from CMIP6, SIE [TO and NN] area where sea ice thickness >1.25 m from CMIP6, $SIAt$ [NN]
	Outputs	September sea ice extent from CMIP6, SIE_{SEP}
Testing data	Inputs	Sea ice extent from the NSIDC, SIE [TO and NN] area where sea ice thickness >1.25 m from the PIOMAS, $SIAt$ [NN]
	Outputs	September sea ice extent from the NSIDC, SIE_{SEP}
Model architecture	Layer 1	Dense and nodes = 10
	Layer 2	Dense and nodes = 10
	Layer 3	Dense, bins, and softmax activation
	Optimizer	Adam
	Activation function	ReLU (slope coefficient, $\alpha = 0.01$)
	Loss function	KL Divergence
	Regularizer	L2 with $\lambda = 0.01$
	Epochs	10
	Batch Size	200

Data Availability Statement

We acknowledge all sources of publicly available data that were used in this study. The data from the Coupled Model Intercomparison Project Phase 6 (CMIP6) can be found on the Earth System Grid Federation (ESFF) nodes. The DOI for each model simulation can be found in the Supplementary Materials (Table S1 in Supporting Information S1). The National Snow and Ice Data Center (NSIDC) Sea Ice Index can be accessed from Fetterer et al. (2017). The Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS) thickness data can be accessed via the Polar Science Center of the University of Washington from Schweiger et al. (2011) and Zhang and Rothrock (2003). The GISS Surface Temperature Analysis data can be accessed via the NASA Goddard Institute for Space Studies from GISTEMP Team (2025). All scripts and data used for the processing in this document can be accessed through Hoffman (2025).

Acknowledgments

This research study was funded by the European Union (ERC, Arctic WATCH, 101040858). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. We thank our reviewers for their helpful feedback.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems (Version 2.8.0). [Software]. https://www.tensorflow.org/

Alber, M., Lapuschkin, S., Seegerer, P., Hagele, M., Schutt, K. T., Montavon, G., et al. (2019). iNNvestigate neural networks. *Journal of Machine Learning Research*, *20*. http://jmlr.org/papers/v20/18-540.html

Amo, A., Montero, J., Biging, G., & Cutello, V. (2004). Fuzzy classification systems. *European Journal of Operational Research*, *156*(2), 495–507. https://doi.org/10.1016/S0377-2217(03)00002-X

Årthun, M., Onarheim, I. H., Dörr, J., & Eldevik, T. (2021). The seasonal and regional transition to an ice-free arctic. *Geophysical Research Letters*, *48*(1), e2020GL090825. https://doi.org/10.1029/2020GL090825

Blanchard-Wrigglesworth, E., Armour, K. C., Bitz, C. M., & DeWeaver, E. (2011). Persistence and inherent predictability of Arctic sea ice in a GCM ensemble and observations. *Journal of Climate*, *24*(1), 231–250. https://doi.org/10.1175/2010JCLI3775.1

Blanchard-Wrigglesworth, E., Bitz, C. M., & Holland, M. M. (2011). Influence of initial conditions and climate forcing on predicting Arctic sea ice. *Geophysical Research Letters*, *38*(18). https://doi.org/10.1029/2011GL048807

Bonan, D. B., Lehner, F., & Holland, M. M. (2021). Partitioning uncertainty in projections of Arctic sea ice. *Environmental Research Letters*, *16*(4), 044002. https://doi.org/10.1088/1748-9326/abe0ec

Bushuk, M., Ali, S., Bailey, D. A., Bao, Q., Batté, L., Bhatt, U. S., et al. (2024). Predicting September Arctic sea ice: A multimodel seasonal skill comparison. *Bulletin of the American Meteorological Society*, *105*(7), E1170–E1203. https://doi.org/10.1175/BAMS-D-23-0163.1

Bushuk, M., Winton, M., Bonan, D. B., Blanchard-Wrigglesworth, E., & Delworth, T. L. (2020). A mechanism for the Arctic sea ice spring predictability barrier. *Geophysical Research Letters*, *47*(13), e2020GL088335. https://doi.org/10.1029/2020GL088335

Chevallier, M., & Salas-Mélia, D. (2012). The role of sea ice thickness distribution in the Arctic sea ice potential predictability: A diagnostic approach with a coupled GCM. *Journal of Climate*, *25*(8), 3025–3038. https://doi.org/10.1175/JCLI-D-11-00209.1

Chi, J., Bae, J., & Kwon, Y.-J. (2021). Two-stream convolutional long- and short-term memory model using perceptual loss for sequence-to-sequence arctic sea ice prediction. *Remote Sensing*, *13*(17), 3413. https://doi.org/10.3390/rs13173413

Day, J. J., Tietsche, S., & Hawkins, E. (2014). Pan-Arctic and regional sea ice predictability: Initialization month dependence. *Journal of Climate*, *27*(12), 4371–4390. https://doi.org/10.1175/JCLI-D-13-00614.1

DelSole, T., & Tippett, M. K. (2018). Predictability in a changing climate. *Climate Dynamics*, *51*(1–2), 531–545. https://doi.org/10.1007/s00382-017-3939-8

Dörr, J. S., Bonan, D. B., Årthun, M., Svendsen, L., & Wills, R. C. J. (2023). Forced and internal components of observed Arctic sea-ice changes. *The Cryosphere*, *17*(9), 4133–4153. https://doi.org/10.5194/tc-17-4133-2023

Eyring, V., Bony, S., Meehl, G., Senior, C., Stevens, B., Ronald, S., & Taylor, K. (2015). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organisation. *Geoscientific Model Development Discussions*, *8*, 10539–10583. https://doi.org/10.5194/gmdd-8-10539-2015

Fetterer, F., Knowles, K., Meier, W. N., Savoie, M., & Windnagel, A. K. (2017). Sea Ice Index, (G02135, version 3) [Dataset]. *National Snow and Ice Data Center*. https://doi.org/10.7265/N5K072F8

Flora, M. L., Potvin, C. K., McGovern, A., & Handler, S. (2024). A machine learning explainability tutorial for atmospheric sciences. *Artificial Intelligence for the Earth Systems*, *3*(1), e230018. https://doi.org/10.1175/AIES-D-23-0018.1

Froyland, G. (2001). Extracting dynamical behavior via Markov models. In A. I. Mees (Ed.), *Nonlinear dynamics and statistics* (pp. 281–321). Birkhäuser Boston. https://doi.org/10.1007/978-1-4612-0177-9_12

GISTEMP Team. (2025). GISS Surface Temperature Analysis (GISTEMP), version 4 [Dataset]. *NASA Goddard Institute for Space Studies*. https://data.giss.nasa.gov/gistemp/

Goosse, H., Arzel, O., Bitz, C. M., de Montety, A., & Vancoppenolle, M. (2009). Increased variability of the Arctic summer ice extent in a warmer climate. *Geophysical Research Letters*, *36*(23). https://doi.org/10.1029/2009GL040546

Guemas, V., Blanchard-Wrigglesworth, E., Chevallier, M., Day, J. J., Déqué, M., Doblas-Reyes, F. J., et al. (2016). A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 546–561. https://doi.org/10.1002/qj.2401

Hamilton, J. D. (1994). Modeling time series with changes in regime. In *Time series analysis* (pp. 677–703). Princeton University Press. Retrieved from http://www.jstor.org/stable/j.ctv14jx6sm.25

Hawkins, E., Tietsche, S., Day, J. J., Melia, N., Haines, K., & Keeley, S. (2016). Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 672–683. https://doi.org/10.1002/qj.2643

Hoffman, L. (2025). Data for "Probabilistic forecasts of September Arctic sea ice extent at the interannual timescale with data-driven statistical models" [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.16143257

Holland, M. M., Bitz, C. M., & Tremblay, B. (2006). Future abrupt reductions in the summer Arctic sea ice. *Geophysical Research Letters*, *33*(23). https://doi.org/10.1029/2006GL028024

Holland, M. M., & Hunke, E. C. (2022). A review of Arctic sea ice climate predictability in large-scale Earth system models. *Oceanography*, *35*, 20–27. https://doi.org/10.5670/oceanog.2022.113

Holland, M. M., Landrum, L., Bailey, D., & Vavrus, S. (2019). Changing seasonal predictability of Arctic summer sea ice area in a warming climate. *Journal of Climate*, *32*(16), 4963–4979. https://doi.org/10.1175/JCLI-D-19-0034.1

Ionita, M., Grosfeld, K., Scholz, P., Treffeisen, R., & Lohmann, G. (2019). September Arctic sea ice minimum prediction – A skillful new statistical approach. *Earth System Dynamics*, *10*(1), 189–203. https://doi.org/10.5194/esd-10-189-2019

IPCC. (2023). *Climate change 2021 – The physical science basis: Working group I contribution to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press.

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

Kay, J. E., Holland, M. M., & Jahn, A. (2011). Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters*, *38*(15). https://doi.org/10.1029/2011GL048008

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Labe, Z., Magnusdottir, G., & Stern, H. (2018). Variability of Arctic sea ice thickness using PIOMAS and the CESM large ensemble. *Journal of Climate*, *31*(8), 3233–3247. https://doi.org/10.1175/JCLI-D-17-0436.1

Lee, Y. J., Watts, M., Maslowski, W., Kinney, J. C., & Osinski, R. (2023). Assessment of the pan-Arctic accelerated rate of sea ice decline in CMIP6 historical simulations. *Journal of Climate*, *36*(17), 6069–6089. https://doi.org/10.1175/JCLI-D-21-0539.1

Lenssen, N., Schmidt, G., Hendrickson, M., Jacobs, P., Menne, M., & Ruedy, R. (2024). A GISTEMPv4 observational uncertainty ensemble. *Journal of Geophysical Research: Atmospheres*, *129*(17), e2023JD040179. https://doi.org/10.1029/2023JD040179

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M., & Barriat, P.-Y. (2012). Constraining projections of summer Arctic sea ice. *The Cryosphere*, *6*(6), 1383–1394. https://doi.org/10.5194/tc-6-1383-2012

Massonnet, F., Vancoppenolle, M., Goosse, H., Docquier, D., Fichefet, T., & Blanchard-Wrigglesworth, E. (2018). Arctic sea-ice change tied to its mean state through thermodynamic processes. *Nature Climate Change*, *8*(7), 599–603. https://doi.org/10.1038/s41558-018-0204-z

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Notz, D., & Stroeve, J. (2016). Observed arctic sea-ice loss directly follows anthropogenic co_2 emission. *Science*, *354*(6313), 747–750. https://doi.org/10.1126/science.aag2345

Notz, & Community, S. (2020). Arctic sea ice in CMIP6. *Geophysical Research Letters*, *47*(10), e2019GL086749. https://doi.org/10.1029/2019GL086749

Onarheim, I. H., Eldevik, T., Smedsrud, L. H., & Stroeve, J. C. (2018). Seasonal and regional manifestation of Arctic sea ice loss. *Journal of Climate*, *31*(12), 4917–4932. https://doi.org/10.1175/JCLI-D-17-0427.1

Petty, A., Keeney, N., Cabaj, A., Kushner, P., & Bagnardi, M. (2023). Winter Arctic sea ice thickness from ICESat-2: Upgrades to freeboard and snow loading estimates and an assessment of the first three winters of data collection. *The Cryosphere*, *17*(1), 127–156. https://doi.org/10.5194/tc-17-127-2023

Roach, L. A., & Meier, W. N. (2024). Sea ice in 2023. *Nature Reviews Earth & Environment*, *5*(4), 235–237. https://doi.org/10.1038/s43017-024-00542-0

Schweiger, A., Lindsay, R., Zhang, J., Steele, M., Stern, H., & Kwok, R. (2011). Uncertainty in modeled Arctic sea ice volume. *Journal of Geophysical Research*, *116*(C8), C00D06. https://doi.org/10.1029/2011JC007084

Serreze, M. C., Holland, M. M., & Stroeve, J. (2007). Perspectives on the Arctic's shrinking sea-ice cover. *Science*, *315*(5818), 1533–1536. https://doi.org/10.1126/science.1139426

Sévellec, F., & Drijfhout, S. S. (2018). A novel probabilistic forecast system predicting anomalously warm 2018-2022 reinforcing the long-term global warming trend. *Nature Communications*, *9*(1), 3024. https://doi.org/10.1038/s41467-018-05442-8

Sinha, A., & Abernathey, R. (2021). Estimating ocean surface currents with machine learning. *Frontiers in Marine Science*, *8*, 672477. https://doi.org/10.3389/fmars.2021.672477

Sticker, A., Massonnet, F., Fichefet, T., DeRepentigny, P., Jahn, A., Docquier, D., et al. (2024). Seasonality and scenario dependence of rapid Arctic sea ice loss events in CMIP6 simulations. *EGUsphere*, *2024*, 1–21. https://doi.org/10.5194/egusphere-2024-1873

Stroeve, J., Hamilton, L. C., Bitz, C. M., & Blanchard-Wrigglesworth, E. (2014). Predicting September sea ice: Ensemble skill of the search sea ice outlook 2008–2013. *Geophysical Research Letters*, *41*(7), 2411–2418. https://doi.org/10.1002/2014GL059388

Stroeve, J., & Notz, D. (2018). Changing state of Arctic sea ice across all seasons. *Environmental Research Letters*, *13*(10), 103001. https://doi.org/10.1088/1748-9326/aade56

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E., & Jahn, A. (2015). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, *5*(2), 86–89. https://doi.org/10.1038/nclimate2483

Thorpe, A., Bauer, P., Magnusson, L., & Richardson, D. (2013). An evaluation of recent performance of ECMWF's forecasts. *ECMWF Newsletter*, *137*, 15–18. https://doi.org/10.21957/hi1eektr

Tietsche, S., Day, J. J., Guemas, V., Hurlin, W. J., Keeley, S. P. E., Matei, D., et al. (2014). Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophysical Research Letters*, *41*(3), 1035–1043. https://doi.org/10.1002/2013GL058755

Weisheimer, A., & Palmer, T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of The Royal Society Interface*, *11*(96), 20131162. https://doi.org/10.1098/rsif.2013.1162

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zhang, J., & Rothrock, D. A. (2003). Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates. *Monthly Weather Review*, *131*(5), 845–861. https://doi.org/10.1175/1520-0493(2003)131<0845:mgsiwa>2.0.co;2

References From the Supporting Information

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., et al. (2018). *IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.5195

Dix, M., Bi, D., Dobrohotoff, P., Fiedler, R., Harman, I., Law, R., et al. (2019). *CSIRO-ARCCSS ACCESS-CM2 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4271

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). *CCCMA CANESM5 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.3610

10.1029/2025JH000669

Tatebe, H., & Watanabe, M. (2018). *MIROC MIROC6 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.5603

Yukimoto, S., Koshiro, T., Kawai, H., Oshima, N., Yoshida, K., Urakawa, S., et al. (2019). *MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.6842

Ziehn, T., Chamberlain, M., Lenton, A., Law, R., Bodman, R., Dix, M., et al. (2019). *Csiro ACCESS-ESM1.5 model output prepared for CMIP6 CMIP historical*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4272