Extended-Range Arctic Sea Ice Forecast with Convolutional Long Short-Term Memory Networks

YANG LIU,[a,b] LAURENS BOGAARDT,[c] JISK ATTEMA,[a] AND WILCO HAZELEGER[d]

[a] *Netherlands eScience Center, Amsterdam, Netherlands*
[b] *Meteorology and Air Quality Group, Wageningen University, Wageningen, Netherlands*
[c] *Department for Statistics, Informatics and Modelling, National Institute for Public Health and the Environment, Utrecht, Netherlands*
[d] *Faculty of Geosciences, Utrecht University, Utrecht, Netherlands*

(Manuscript received 10 April 2020, in final form 27 February 2021)

ABSTRACT: Operational Arctic sea ice forecasts are of crucial importance to science and to society in the Arctic region. Currently, statistical and numerical climate models are widely used to generate the Arctic sea ice forecasts at weather time scales. Numerical models require near-real-time input of relevant environmental conditions consistent with the model equations and they are computationally expensive. In this study, we propose a deep learning approach, namely convolutional long short-term memory networks (ConvLSTM), to forecast sea ice in the Barents Sea at weather to subseasonal time scales. This is an unsupervised learning approach. It makes use of historical records and it exploits the covariances between different variables, including spatial and temporal relations. With input fields from reanalysis data, we demonstrate that ConvLSTM is able to learn the variability of the Arctic sea ice and can forecast regional sea ice concentration skillfully at weekly to monthly time scales. It preserves the physical consistency between predictors and predictands, and generally outperforms forecasts with climatology, persistence, and a statistical model. Based on the known sources of predictability, sensitivity tests with different climate fields as input for learning were performed. The impact of different predictors on the quality of the forecasts are evaluated and we demonstrate that the surface energy budget components have a large impact on the predictability of sea ice at weather time scales. This method is a promising way to enhance operational Arctic sea ice forecasting in the near future.

KEYWORDS: Sea ice; Statistical forecasting; Deep learning; Machine learning

1. Introduction

As one of the most noticeable frontiers with visible changes due to global warming, the Arctic has received more and more attention in recent decades. This is accompanied with increased commercial and scientific activities as a result of sea ice melting. This drives a demand for reliable operational sea ice forecasts, especially for shipping companies and related stakeholders (Gascard et al. 2017; Stephenson and Pincus 2018). Therefore, it is of crucial importance to improve operational Arctic sea ice forecasts at weather time scales.

The physical interactions between atmospheric and oceanic conditions and Arctic sea ice provide a basis for forecasting sea ice characteristics. Predictability of Arctic sea ice at different time scales in different seasons has been explored extensively in many studies. Blanchard-Wrigglesworth et al. (2011a) investigated the temporal evolution of Arctic sea ice in observations and in ensemble climate model output. They found a summer to summer and a melt season to growth season reemergence effect in sea ice which potentially serves as a good predictor for Arctic sea ice forecasts at monthly to annual time scales. Mohammadi-Aragh et al. (2018) studied the potential predictability of Arctic sea ice in winter, including the deformation and concentration of sea ice at weather time scales. They noticed that the sea ice concentration (SIC) is predictable throughout a 10-day forecast period.

In addition there are studies on longer range forecasts. Guemas et al. (2016) reviewed progress on sea ice forecasts and showed that predictability of Arctic sea ice at seasonal to decadal time scales mainly originates from persistence or advection of sea ice anomalies, air–sea interaction, and changes in radiative forcing. Krikken and Hazeleger (2015) analyzed the natural variability of Arctic sea ice from an energy budget perspective. They found strong correlations between the Arctic energy balance components and the reemergence of sea ice anomalies from the melt season to the growth season, which extends the theory proposed by Blanchard-Wrigglesworth et al. (2011a) and further confirms the essential role of the energy budget in sea ice forecasts. Another key element to the predictability of sea ice is the sea ice thickness (SIT). Bonan et al. (2019) explore the role of SIT on the summer predictability of

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-20-0113.s1.

Corresponding author: Yang Liu, y.liu@esciencecenter.nl

DOI: 10.1175/MWR-D-20-0113.1

sea ice and found that similar skill of SIT forecasts can be obtained as a perfect model experiment. They also discussed the predictability barrier in late spring and made suggestions for the initialization of forecasts.

Cruz-García et al. (2019) examined seasonal-to-interannual sea ice predictability with multiple climate models and revisited the essential role of the reemergence effect of sea ice anomalies. They observed that SIC anomalies in the Barents Sea have a strong negative correlation with the local sea surface temperature (SST) anomalies. Moreover, Onarheim et al. (2015) emphasized that ocean heat transport (OHT) variations play an important role in the observed winter sea ice variance in the Barents Sea. They claimed an increase in forecast skill at annual times scales up to 2 years using OHT. The knowledge of sea ice predictability associated with different physical processes provided by these studies underlines the importance of choosing relevant predictors related to the target location and time scale.

Currently, many operational Arctic sea ice forecasts are produced by numerical climate models (e.g., Van Woert et al. 2004; Metzger et al. 2014; Hebert et al. 2015; Smith et al. 2013, 2016). These numerical models are built upon physical linkages in the climate system and are able to generate accurate sea ice forecasts, but they are computationally expensive, due to the need for relatively high spatial and temporal resolutions, the implementation of ensemble approaches to address uncertainties, their dependency on the real-time input of observed conditions for the data assimilation processes, and the calibration of model output. Moreover, dynamical models are imperfect and many processes have to be parameterized. Specifically for sea ice there are several modeling challenges. For instance, rheology, ice thickness distribution, wave–ice interaction, landfast ice, melt ponding, and floe size distribution (Leppäranta et al. 2020). Many studies have shown that the forecast skill strongly relies on the initialization (Blanchard-Wrigglesworth et al. 2011b; Goessling et al. 2016), target location and time scales (Cruz-García et al. 2019).

Some operational sea ice forecasts are generated by statistical models (Howell et al. 2015; Yuan et al. 2016; Wang et al. 2019). Most of these statistical models are linear models, thus they are not suited to learn nonlinear relations between variables in the Arctic climate system. Given the importance of nonlinear feedback mechanisms in the atmosphere, ocean and sea ice coupled system in the Arctic, we may need nonlinear approaches to forecast Arctic sea ice with a statistical model. This brings contemporary machine learning techniques into scope.

Machine learning approaches, especially deep learning, are widely embraced by many fields and are increasingly used to deal with problems like clustering, classification, and regression (LeCun et al. 2015). Benefiting from large volumes of data of Earth system (Knüsel et al. 2019), those deep learning methods may be appropriate for the weather and climate domain (Reichstein et al. 2019). Although these applications still have limitations, for example, they rely on the data from numerical weather forecast or reanalysis for the training process and this could be computationally very expensive depending on the configuration as well as the tuning procedure, there are many successful use cases. These cases are, for instance, the representation of physical processes (e.g., Rasp et al. 2018), weather and climate forecasts (e.g., Salman et al. 2015; Ham et al. 2019), and extreme events detection (e.g., Gope et al. 2016). Deep learning based techniques could potentially be used as an alternative method or an auxiliary approach for the current state-of-the-art forecast systems, or even as a preliminary and fast forecast system. It can be viewed as an enhancement or supplement to our existing tools.

In this study, we consider sea ice forecasts at weekly time scales and we perform extended-range sea ice forecasts in the Barents Sea with a complex deep neural network (DNN), namely the convolutional long short-term memory networks (ConvLSTM). These intricate neural networks are built on top of the basic structures, like convolutional neural networks (CNN) and recurrent neural networks (RNN). Early studies have shown that even these basic neural networks (NN) are able to reproduce both the short-term evolution and the long-term statistics of dynamical systems like a Lorenz system (e.g., Chattopadhyay et al. 2020), which provides a basis for learning the nonlinear relations between meteorological fields and predicting the evolution of a weather-like system in a chaotic regime.

To work with such complex spatial–temporal sequence problems, ConvLSTM are useful (Xingjian et al. 2015). As a novel combination of CNN and Long Short-Term Memory (LSTM) networks (Fukushima 1980; Hochreiter and Schmidhuber 1997), ConvLSTM were first introduced by Xingjian et al. (2015) when dealing with precipitation nowcasting. Until now, many studies have shown that ConvLSTM are suitable for weather forecasts at different time scales, like precipitation forecasts (Xingjian et al. 2015; Kim et al. 2017), hurricane tracking and forecasting (Kim et al. 2019), sea ice concentration (Kim et al. 2020) and sea ice motion forecasts (Petrou and Tian 2019). However, most of those studies only incorporate a few variables and were mostly data-driven without physical insights. In this paper, except for the sea ice forecasts with ConvLSTM using multiple predictors, we also conducted a sensitivity analysis of predictors and focus on the physical consistency between sea ice and other meteorological fields in the trained and forecasted output.

The paper is organized as follows: The methodology and the datasets used in this study are described in section 2. The results are shown in section 3, including constrained forecasts, sensitivity tests of predictors and operational forecasts with ConvLSTM. The discussion and a brief summary of this study are given in sections 4 and 5, respectively.

2. Data and methodology

A detailed elaboration on the deep neural networks and datasets used in this study is given in this section. In addition, a brief summary about the hyper-parameter tuning of our neural networks and an overview of the evaluation metrics is included at the end of this section.

a. Convolutional long short-term memory networks

To enhance LSTM networks to include learning and forecasting spatial information, Xingjian et al. (2015) embedded convolutional cells into LSTM cells and created a new

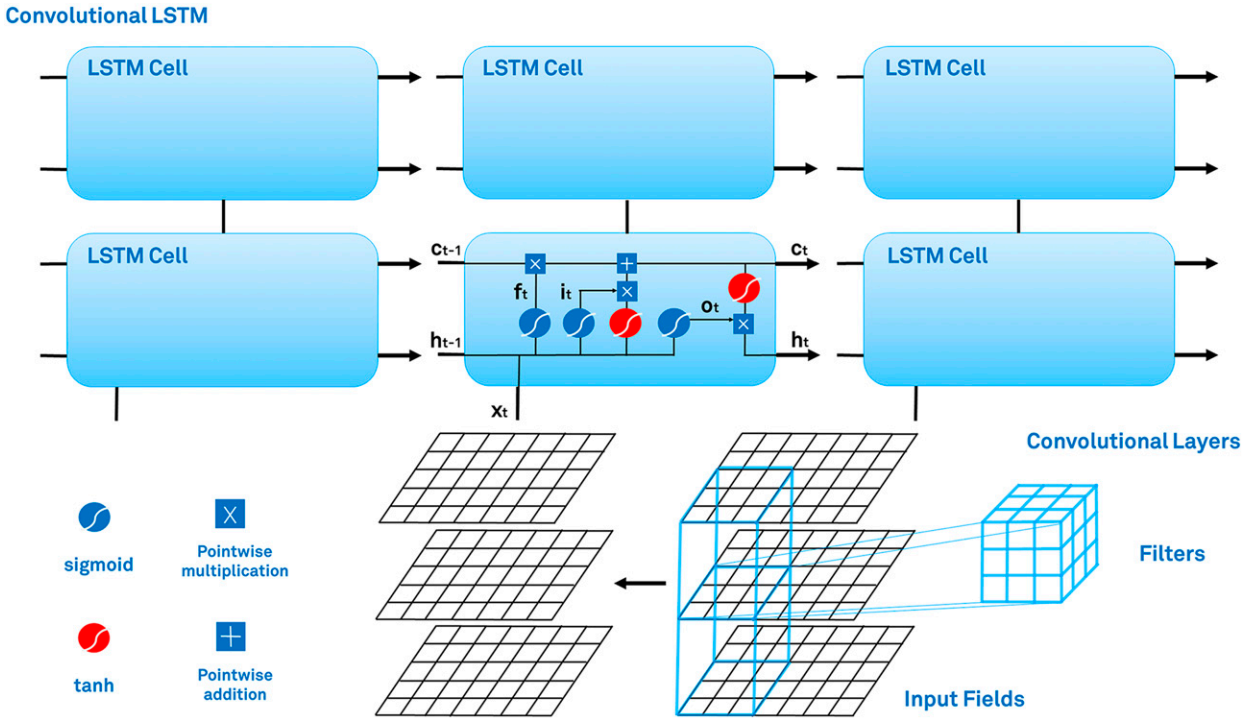

FIG. 1. Structure of the convolutional long short-term memory neural networks.

neural network structure coined ConvLSTM. Consequently, ConvLSTM inherits the ability of LSTM to ''remember'' and ''forget,'' which is achieved by the design of memory cells and multiple gates that control the flow of information (Hochreiter and Schmidhuber 1997). Also, the spatial awareness of a convolutional network is added to a LSTM. These aspects of the structure are relevant for weather and climate problems. The structure of ConvLSTM can be defined and explained by the following equations (Xingjian et al. 2015):

$$
\begin{aligned}
i_t &= \sigma(\mathbf{W}_{xi} * x_t + \mathbf{W}_{hi} * h_{t-1} + \mathbf{W}_{ci} \circ c_{t-1} + b_i),\\
f_t &= \sigma(\mathbf{W}_{xf} * x_t + \mathbf{W}_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f),\\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(\mathbf{W}_{xc} * x_t + \mathbf{W}_{hc} * h_{t-1} + b_c),\\
o_t &= \sigma(\mathbf{W}_{xo} * x_t + \mathbf{W}_{ho} * h_{t-1} + \mathbf{W}_{ct} \circ c_t + b_o),\\
h_t &= o_t \circ \tanh(c_t),
\end{aligned}
\tag{1}
$$

with i_t the input gate, f_t the forget gate, c_t the cell state, o_t the output gate, h_t the hidden state, $\mathbf{W}$ the weight matrix, x the input, b the bias, * the convolutional operation, $\circ$ the element-wise product, σ the sigmoid function, and tanh the hyperbolic tangent function. The subscripts describe the correspondence of the weight matrix to different gates and states. For instance, $\mathbf{W}_{xi}$ indicates the weight matrix of input values related to the input gate, while $\mathbf{W}_{hf}$ represents the weight matrix of hidden states corresponded to the forget gate. The subscript t indicates the time step and will be elucidated in section 2c.

The structure of the ConvLSTM network is illustrated in Fig. 1. At each time step, convolutions over the input fields (e.g., data of the Arctic climate system) are performed. Then, at each grid point, those values are fed into an LSTM cell. The LSTM cells only differ in their input, and hence their memory, but share all other parameters. This way, the number of parameters in the ConvLSTM is vaster than that of a convolutional network that takes input as all fields at all time steps. Multiple layers can be stacked to further increase the complexity of the network if needed.

ConvLSTM networks are powerful tools for intricate spatial–temporal sequence prediction problems. They are likely suitable for sea ice forecasts. Physically, the use of filters inside convolutional layers accounts for the local interactions between multiple fields (e.g., temperature, wind) which affect the formation of sea ice, and the advance and retreat of sea ice at neighboring grid points. The temporal evolution of sea ice, including the communication of neighbor points within the convolutional cells, are tracked by the LSTM structure of the network through its recurrence feature. Moreover, this approach is unsupervised learning and it can make use of historical records of weather and climate states.

In this study, we perform many-to-one prediction, which means sequences with spatial structure are taken as input and spatial maps of one time step ahead will be the output by the networks. The numerical processes, including training and testing of ConvLSTM, are elaborated upon in detail in the appendix. The ConvLSTM used in this study are constructed on top of the Pytorch library, and our script is published on Github (https://github.com/geek-yang/DLACs).

b. *Reanalysis datasets*

We train ConvLSTM and evaluate its capability to forecast sea ice using reanalysis datasets, namely, ERA-Interim and ORAS4.

ERA-Interim is a global atmospheric reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Dee et al. 2011), which covers the data-rich period since 1979. It employs the cycle 31r2 of ECMWF's Integrated Forecast System (IFS) and generates atmospheric state estimates using 4D-Var data assimilation with a T255 (~79 km) horizontal resolution on 60 vertical levels (Berrisford et al. 2009). We use the surface fields, including SIC, 2 m temperature (T2M), sea level pressure (SLP), net surface turbulent and radiation flux (SFlux), 10 m zonal and meridional wind (UV10m), and geopotential height at 850 hPa (Z850) and 500 hPa (Z500), with a 0.75° × 0.75° horizontal resolution (~ 28.6 km × 28.6 km, at 70°N). We take 6-hourly data with a range from 1979 to 2016 and the data are averaged to weekly time scales.

Given a lack of detailed verification studies of the SIC quality in ERA-Interim, we further verify the SIC field in ERA-Interim with a satellite-based product, namely the NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration (version 3) (Peng et al. 2013). A point-to-point comparison of the SIC in the Barents Sea between ERA-Interim and the chosen satellite product is shown in Fig. S1 in the online supplemental material. They look similar in most of the areas, except for some places close to the continent as a result of land–sea mask and interpolation since the native grids are different. To ensure the robustness of our results, we also perform our training and testing with the SIC field from the chosen satellite-derived product and the results are almost the same as those based on the SIC in ERA-Interim (see Fig. S2).

The Ocean Reanalysis System 4, in short (ORAS4), is the replacement of the reanalyses system ORAS3 used by the ECMWF (Balmaseda et al. 2013). It implements Nucleus for European Modelling of the Ocean (NEMO) as ocean model (Madec 2008; Ferry et al. 2012) and uses NEMOVAR as the data assimilation system (Mogensen et al. 2012). The model is forced by atmosphere-derived daily surface fluxes, from ERA-40 from 1957 to 1989 and ERA-Interim from 1989 onward. ORAS4 produces analyses with a 3D-Var FGAT assimilation scheme and spans from 1958 to the present. ORAS4 runs on the ORCA1 grid, which is associated with a horizontal resolution of 1° in the extratropics and a refined meridional resolution up to 0.3° in the tropics. It has 42 vertical levels, 18 of which are located in the upper 200 m. We use the monthly mean temperature on the native model grid from 1979 to 2016 to calculate the ocean heat content (OHC) from the sea surface to 300 m. Given the long memory effect in the ocean, OHC at weekly time scales is interpolated from monthly fields.

We chose this combination because both datasets are ECMWF reanalysis products and ORAS4 takes surface forcing from ERA-Interim (Balmaseda et al. 2013). This combination is promising to provide a physically consistent picture of the interaction between the atmosphere-, ocean-, and sea ice–related processes. The selected fields are potential predictors for the sea ice variations due to their physical relationships in the Arctic (Krikken and Hazeleger 2015; Guemas et al. 2016).

In this study, we only focus on the Barents Sea. Following the same definition of the Arctic regions as Walsh et al. (2019), our domain is covered by the ERA-Interim grid with 24 (latitudinal) × 56 (longitudinal) points. It is noteworthy that this regional focus has a negative impact on the performance of ConvLSTM. The sampling of convolutional layers is affected by the cutoff of data close to the boundary of the Barents Sea. This partially explains the relatively bad forecast quality in the boundary regions, as we will show, and provides room for improvement if a larger area is included in the future. However, this comes with a computational cost.

c. Evaluation metrics

Two types of lead-time-dependent forecast were performed in this study. One is called *constrained forecast*, which takes input fields from the future, excluding SIC only, to test the maximum expected predictability given the chosen forecast methods and input fields. It can be described with the equation shown below:

$$\mathrm{SIC}_{\mathrm{pred}[t_{n+l+1}]} = \mathrm{ConvLSTM}\Big(\mathrm{SIC}_{\mathrm{obs}[t_1,t_2,\ldots,t_n]+\mathrm{pred}[t_{n+1},t_{n+2},\ldots,t_{n+l}]}, \mathrm{OHC}_{\mathrm{obs}[t_1,t_2,\ldots,t_n,t_{n+1},t_{n+2},\ldots t_{n+l}]}, \ldots\Big). \quad (2)$$

Here $\mathrm{OHC}_{\mathrm{observe}[t_{n+l}]}$ is the observed OHC for the l leading week. An extended analysis of the contributions from several predictors was conducted based on the constrained forecast formulation.

The other setup is called *operational forecast*, which uses only historical records and the forecasts at a specific lead time are based on the predicted fields of all variables (e.g., the week $n + 2$ forecast is made with all variables of the reanalysis time series considered until the current step n and the week $n + 1$ forecast). The procedure can be explained by the equation below:

$$\mathrm{SIC}_{\mathrm{pred}[t_{n+l+1}]}, \mathrm{OHC}_{\mathrm{pred}[t_{n+l+1}]}, \ldots = \mathrm{ConvLSTM}\Big(\mathrm{SIC}_{\mathrm{obs}[t_1,t_2,\ldots,t_n]+pred[t_{n+1},t_{n+2},\ldots,t_{n+l}]}, \mathrm{OHC}_{obs[t_1,t_2,\ldots,t_n]+pred[t_{n+1},t_{n+2},\ldots,t_{n+l}]}, \ldots\Big) \quad (3)$$

Here $\mathrm{SIC}_{\mathrm{obs}[t_1,t_2,\ldots,t_n]}$ and $\mathrm{OHC}_{\mathrm{obs}[t_1,t_2,\ldots,t_n]}$ are the time series of observed SIC and OHC until the current time step, respectively; and $\mathrm{SIC}_{\mathrm{pred}[t_{n+l}]}$ and $\mathrm{OHC}_{\mathrm{predict}[t_{n+l}]}$ are the predicted SIC and OHC for the l leading week, respectively. We can assess the performance of ConvLSTM with reforecasts, i.e., performing forecasts over the reanalysis period as if they were actual operational forecasts.

The configurations of constrained forecasts and operational forecasts are illustrated in Fig. 2. The major differences between these two setups are as follows: the constrained forecasts use predictors from the future to predict SIC, whereas the operational forecasts use forecast predictands as predictors to predict future SIC. The value of considering the constrained forecasts is to gain insight into what forecasts of SIC would look like if the ConvLSTM models were able to perfectly forecast the predictors that are then used in the operational forecasts.

To evaluate the performance of sea ice forecast by the ConvLSTM, several scores are calculated. The root-mean-square

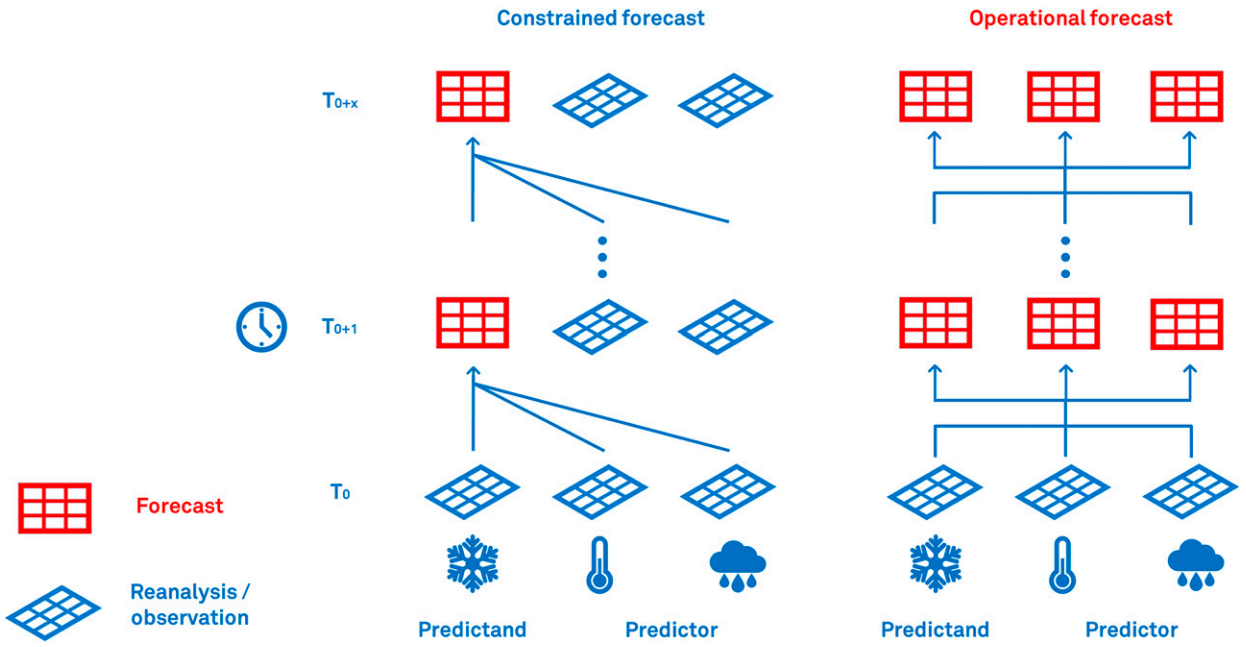

FIG. 2. Configurations of constrained forecast and operational forecast at starting time step T_0 and lead time step T_{0+x}.

error (RMSE) is used to evaluate the sea ice forecast in the chosen area. To evaluate the predicted data with both temporal and spatial information, we define the RMSE in the following way:

$$\mathrm{RMSE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{\left\{\sum_{x=1}^{X}\sum_{y=1}^{Y}\left[a_{x,y}\mathrm{SIC}_{x,y,t}^{(\mathrm{predict})} - \mathrm{SIC}_{x,y,t}^{(\mathrm{observe})}\right]\right\}}{\sum_{x=1}^{X}\sum_{y=1}^{Y}a_{x,y}}}. \quad (4)$$

Here x is the number of points in the longitudinal direction; y is the number of points in latitudinal direction; X and Y are the total number of gridcell length in longitudinal and latitudinal direction, respectively; t is the number of time step; N is the total time steps; $a_{x,y}$ is the area of grid cell denoted by indices x and y; and $\mathrm{SIC}_{x,y,t}^{(\mathrm{predict})}$ and $\mathrm{SIC}_{x,y,t}^{(\mathrm{observe})}$ are the predicted and observed SIC, respectively.

The mean absolute error (MAE) is used to assess the forecasts on a pointwise basis. The spatial structure is preserved (see Fig. 4). MAE is defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}|a_{x,y}[\mathrm{SIC}_{x,y,t}^{(\mathrm{predict})} - \mathrm{SIC}_{x,y,t}^{(\mathrm{observe})}]|. \quad (5)$$

For several applications, such as shipping and navigation in the Arctic, it is also necessary to know if a certain area is open water or covered by sea ice. This requires a binary forecast. Following Van Woert et al. (2004) and Walsh et al. (2019), the grid boxes with sea ice concentration less than 15% are regarded as open water areas. Based on this criterion, we introduce the integrated ice-edge error score (IIEE), which is a verification metric first proposed by Goessling et al. (2016), to better represent the performance of binary forecasts. The score is defined as the area where the forecast and the "truth" disagree on the ice concentration being above or below 15% and it can be further decomposed into two components, the overestimated (O) and underestimated (U) local sea ice extent, respectively:

$$\begin{aligned}\mathrm{IIEE} &= O + U,\\ O &= \int_A \max(c_f - c_t, 0)\, dA,\\ U &= \int_A \max(c_t - c_f, 0)\, dA,\\ c_f, c_t &= \begin{cases} 1, & \mathrm{SIC} > 15\% \\ 0, & \mathrm{SIC} < 15\% \end{cases},\end{aligned} \quad (6)$$

where A is the area of interest, and subscripts f and t denote the forecast and the truth. By definition, overestimated local sea ice extent means the failure of predicting a sea ice free area as ice covered, while underestimated local sea ice extent is the failure of predicting a sea ice–covered area as ice free.

Many studies assess their sea ice forecast systems against persistence and climatology at weather time scales (Van Woert et al. 2004; Metzger et al. 2014; Hebert et al. 2015; Smith et al. 2013, 2016). The reason is that, at weekly to submonthly time scales, the persistence of sea ice anomalies is very high (Blanchard-Wrigglesworth et al. 2011a; Guemas et al. 2016). Therefore, it is very challenging to beat persistence at these time scales. For instance, Van Woert et al. (2004) provide daily ice analyses and 5-day forecasts with their polar ice forecast system but this has almost no skill in winter against persistence. Consequently, in this study, forecasts with the ConvLSTM using different input fields will also be evaluated against

persistence and climatology. The persistence is defined as the SIC anomaly at lead time step 0 added to the climatology at each lead time. We use the climatology based on a 10-yr sliding window preceding the forecast time (e.g., Zampieri et al. 2018) in order to take changes in the climatology into account.

To compare the ConvLSTM with numerical model-based forecasts, we include two ensemble forecast datasets from the subseasonal to seasonal prediction project (Vitart et al. 2017). We use forecasts from the National Centers for Environmental Prediction (NCEP) and the ECMWF ensemble forecasts. These datasets were chosen because of their active sea ice model, available time range (January 2015–December 2016), and the forecast frequency.

The NCEP global ensemble forecast system generates real time forecasts using the NCEP Climate Forecast System, version 2 (CFSv2) (Saha et al. 2014). It consists of 16 ensemble members and the forecast length is 45 days. The atmospheric model has 64 model levels and its horizontal resolution is T126 (~100 km), which is lower than the ERA-Interim (T255, ~80 km). Its ocean model is GFDL MOM4 (Pacanowski et al. 1991). It has a spatial resolution in the zonal direction of 0.5° and in the meridional direction, 0.25° from 10°S to 10°N, progressively decreasing to 0.5° from 10° to 30°, and is fixed at 0.5° beyond 30° in both hemispheres. There are 40 levels in vertical. The system is coupled to an active sea ice model, which is part of the Modular Ocean Model (MOM4) (Pacanowski et al. 1991).

Based on the Integrated Forecasting System (IFS), version CY46R1, the ECMWF global ensemble forecast system has 51 members and it runs twice a week up to day 46. The atmospheric component of the system has a horizontal resolution of about 16 km up to day 15 and a relatively coarse horizontal resolution of about 32 km after day 15. Vertically, the atmospheric model has 91 model levels. The ocean model is NEMO3.4.1 with a 0.25° horizontal resolution and 75 vertical levels (Madec 2008; Ferry et al. 2012). The system is coupled to the Louvain-la-Neuve Sea Ice Model (LIM2) (Rousset et al. 2015).

Note that the output from the forecast systems in the S2S project have been regridded to the same model grid. To enable a direct comparison, we interpolate our ConvLSTM results and ERA-Interim sea ice data to the model grid used in the S2S project and weigh the results the same as the area weight applied throughout the paper. More information about these experiments can be found on the homepage of the S2S project (https://confluence.ecmwf.int/display/S2S/Models).

Sensitivity tests of the contribution from each predictor to the skill of sea ice forecasts are conducted with the ConvLSTM using SIC plus one extra predictor. To assess the change of forecast skill with different predictors, we introduce a dimensionless score with the definition given below:

$$\text{Relative forecast skill score} = \frac{\text{RMSE}_{\text{sic}} - \text{RMSE}_{\text{predictor}}}{\text{RMSE}_{\text{sic}}}. \quad (7)$$

With RMSE_{sic} the RMSE of forecast with ConvLSTM using only SIC, and $\text{RMSE}_{\text{predictor}}$ the RMSE of forecast with ConvLSTM using SIC and one extra predictor. More details are provided in the section 3b.

d. Training and hyperparameter tuning

The networks are trained with reanalysis data from 1979 to 2008 (1440 weeks). Given the spatial resolution of input fields, the training set includes $24 \times 56 \times 1440$ points, thus 1440 points for each node in the convolutional layer. Data from 2009 to 2012 is used for cross validation, allowing to implement an early stop module and to avoid overfitting. Data from 2013 to 2016 is taken as the test set for evaluation. The weight matrix of the ConvLSTM is updated by optimizing the loss function, for which we use mean square error (MSE):

$$\text{MSE} = \frac{1}{N}\sum_{t=1}^{N} \frac{\left\{\sum_{x=1}^{X}\sum_{y=1}^{Y} a_{x,y}\left[\text{SIC}_{x,y,t}^{(\text{predict})} - \text{SIC}_{x,y,t}^{(\text{observe})}\right]\right\}^2}{\sum_{x=1}^{X}\sum_{y=1}^{Y} a_{x,y}^2}. \quad (8)$$

Physically, this loss function measures the difference between the actual and forecasted SIC.

The training time varies from 8 to 12 h on a single GPU, depending on the choices of hyperparameters (e.g., number of epochs, filter size, number of layers). A brief summary of the hyperparameter tuning of the ConvLSTM used in this study is given in Table S1 in the supplemental material. With an assessment based on the RMSE of sea ice forecasts for the first leading week, the results show that a combination of a learning rate equal to 0.01, three stacked ConvLSTM layers, a filter size of 3×3, and 1500 epochs is the best. The learning curve obtained with this combination of hyperparameters is shown in Fig. S3 in the supplemental material. Although the loss stops decreasing after 600 epochs, the model's skill for anomalies continues to increase.

It is worthwhile emphasizing the importance of the convolutional filter size. In each convolutional layer, the filter size controls the exchange of information between neighboring points. Physically, it accounts for the influence of for instance SIC, OHC or SLP from adjacent regions on the selected area (node). Given the physical consistency between regional atmospheric and oceanic fields, and the advection of sea ice anomalies (Blanchard-Wrigglesworth et al. 2011a; Guemas et al. 2016), this feature of the convolutional layer should improve the forecast skill. Similarly, the number of stacked ConvLSTM layers also relates to the communication of neighboring points because of the filtering in each layer with convolutions.

e. Baseline statistical model analysis

To set a baseline for our forecasts with ConvLSTM with less complex statistical models and to provide insight into its ability to account for the nonlinearity between multiple physical fields, we also fit a generalized linear model with a "logit" link function. For conciseness, it is referred to as the baseline statistical model in this paper. This method is similar to the logistic regression, which means the nonlinear properties of the input fields are covered. In this approach, each spatial location

is modeled separately. While the fitted parameters vary over space, the structure of the linear regression is the same everywhere. In particular, the SIC is predicted using an autoregressive model which includes the value of the previous three weeks at that location, as well as the previous values of the T2M and OHC. In addition, the SIC of all neighboring locations one week prior is used to establish whether spatial drift is a relevant factor. For all terms, only linear parts are included, although the percentage of sea ice content is first transformed using the "logit" function. Finally, the model is fit using ridge regression, where the optimal amount of regularization is determined using fivefold cross validation over the training set. The predictors are selected in terms of the balance between their expected source of predictability and the cost of training. Similarly, it uses the "constrained forecast" configuration as all the input fields are from the reanalysis. Mathematically, this generalized linear model with a logit link function can be expressed as

$$
\begin{aligned}
\mathrm{SIC}(t) &= \beta_t\, t + \beta_{\sin}\sin(t) + \beta_{\cos}\cos(t) + \beta_{\mathrm{SIC}}\cdot\mathbf{SIC}(t) + \beta_{T2M}\cdot\mathbf{T2M}(t)\\
&\quad + \beta_{\mathrm{OHC}}\cdot\mathbf{OHC}(t) + \beta_{\mathrm{neighbours}}\cdot\mathbf{SIC}_{\mathrm{neighbours}}(t), \text{ where}\\
\mathbf{SIC}(t) &= \begin{bmatrix} \mathrm{logit}(\mathrm{SIC}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}_{t-2,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}_{t-3,\mathrm{detrend}}) \end{bmatrix}, \quad \mathbf{T2M}(t) = \begin{bmatrix} \mathrm{logit}(\mathrm{T2M}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{T2M}_{t-2,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{T2M}_{t-3,\mathrm{detrend}}) \end{bmatrix},\\
\mathbf{OHC}(t) &= \begin{bmatrix} \mathrm{logit}(\mathrm{OHC}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{OHC}_{t-2,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{OHC}_{t-3,\mathrm{detrend}}) \end{bmatrix}, \quad \text{and} \quad \mathbf{SIC}_{\mathrm{neighbours}}(t) = \begin{bmatrix} \mathit{logit}(\mathrm{SIC}^{\mathrm{N}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{NE}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{E}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{SE}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{S}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{SW}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{W}}_{t-1,\mathrm{detrend}}) \\ \mathrm{logit}(\mathrm{SIC}^{\mathrm{NW}}_{t-1,\mathrm{detrend}}) \end{bmatrix}, \qquad (9)
\end{aligned}
$$

where t is the current step; SIC(t) is the sea ice forecast of current step; sin(t) and cos(t) are cycles with one year period; β is the trainable weight for each term' $\mathrm{SIC}_{t-x,\mathrm{detrend}}$, $\mathrm{T2M}_{t-x,\mathrm{detrend}}$, and $\mathrm{OHC}_{t-x,\mathrm{detrend}}$ are the detrended SIC, T2M, and OHC at the previous x time step, respectively; and $\mathrm{SIC}^{\mathrm{direction}}_{t-1,\mathrm{detrend}}$ is the detrended SIC of neighboring point at previous time step at certain location (i.e., N indicates north, NE indicates northeast, etc.), excluding land pixels.

3. Results

a. Constrained predictability

Before implementing the novel deep neural network for sea ice forecasts and performing retrospective analysis of skill, it is worthwhile examining its capability in an ideal setup. We performed lead-time-dependent constrained forecasts of SIC with the ConvLSTM network using different combinations of input fields. The RMSE of these forecasts against those given by persistence and climatology with a lead time up to 6 weeks is shown in Fig. 3a (details in Table S2). For all the forecasts (except for climatology as it does not vary over time, by definition), RMSE increases with increased lead time. It is observed that most of the forecasts with the ConvLSTM using SIC and one extra field, such as OHC, Z500, SLP, and SFlux, can outperform the forecasts with persistence and climatology. The increase of RMSE as a function of lead time from these forecasts with the ConvLSTM is smaller than that with persistence and climatology. However, this is not true for some combinations of input fields, for instance, SIC with T2M or Z850. To obtain more insight on the impact of several predictors, an extended constrained SIC forecast with lead time up to 16 weeks is shown in Fig. 7. It is found that with multiple input fields, the performance of ConvLSTM is also stable at long lead times. This time scale is beyond the scope of this study though.

Forecasts with a ConvLSTM using only SIC also provides better results than forecasts from persistence and climatology. The nonlinearity introduced by the ConvLSTM effectively contributes to the skill of the forecast. A comparison between forecasts with our baseline statistical model and forecasts using the ConvLSTM shows that the ConvLSTM produces slightly better forecasts. They both significantly outperform a persistence forecast.

As SIC has a strong seasonal cycle and the predictability is known to be strongly seasonal dependent, we further evaluate the forecasts by examining the error in each month. The RMSE of the constrained forecast of SIC for the first week in each

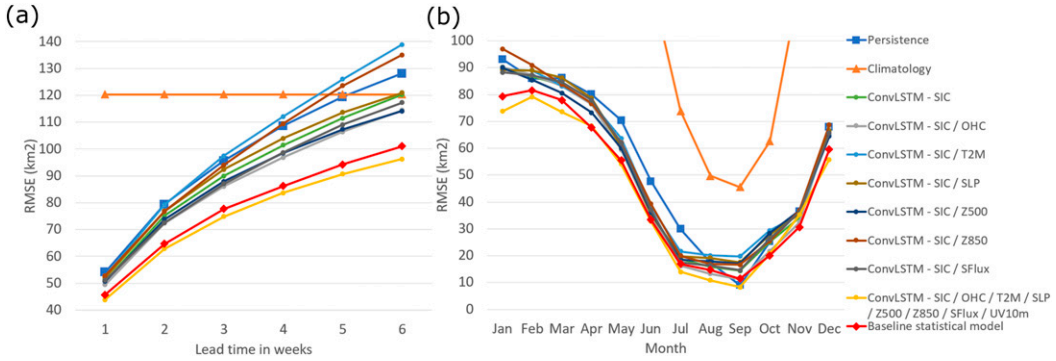

FIG. 3. RMSE of (a) the constrained forecast of SIC with a lead time up to 6 weeks and (b) the constrained forecast of SIC for the first week in each month with ConvLSTM using different predictors against persistence, climatology, and the baseline statistical model. The unit is square kilometers per grid cell.

month with different predictors and methods is given in Fig. 3b (details in Table S3). In general, the forecast error is larger in winter than that in summer, which is similar to operational forecasts with numerical models (Smith et al. 2016). This can be explained by the large year-to-year variations of sea ice in winter and a large open water area in summer (Perovich and Richter-Menge 2009). Compared to the sea ice forecasts with persistence, the ConvLSTM provides more skillful forecasts in

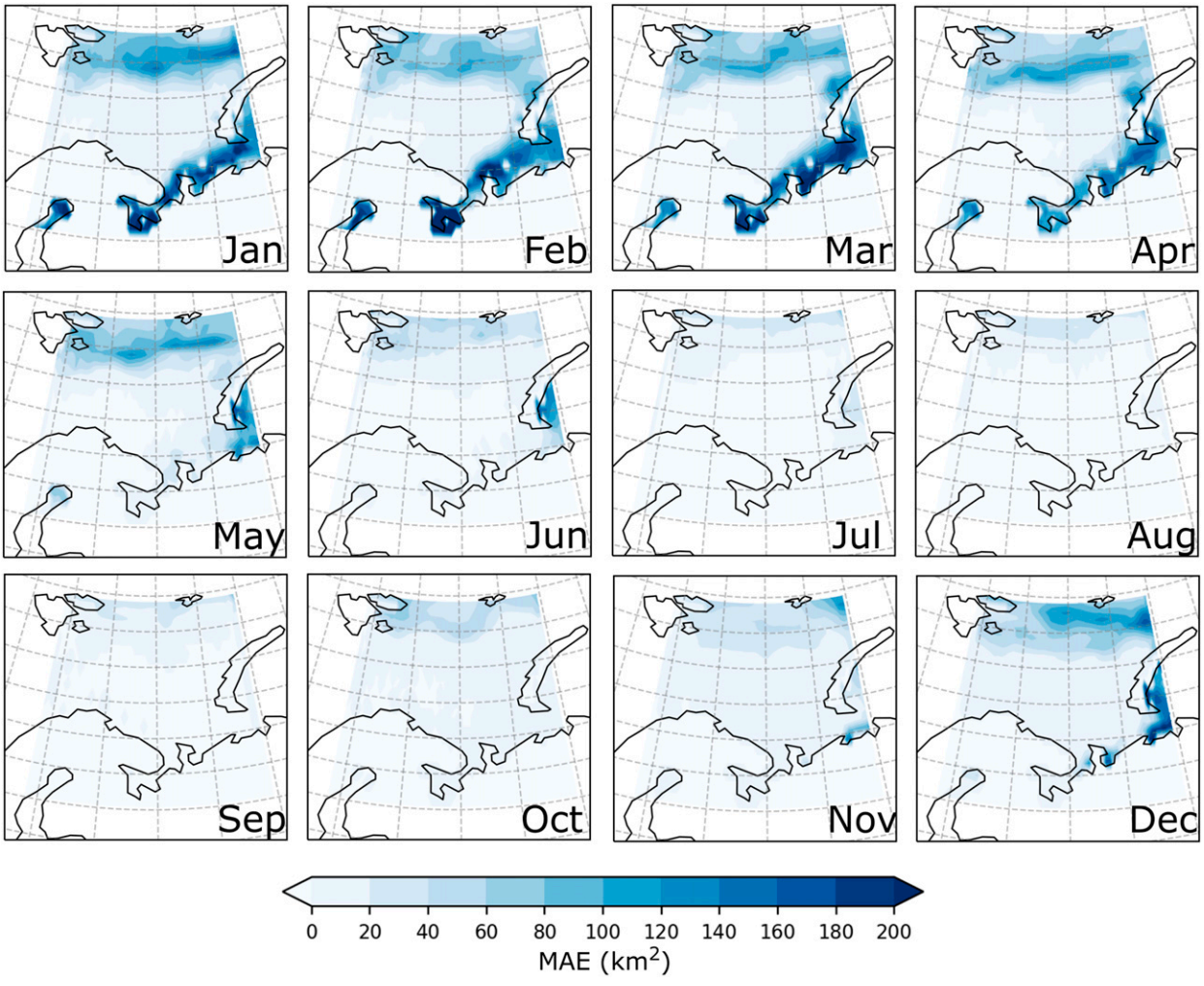

FIG. 4. MAE of the constrained forecast of SIC for the first week in each month with ConvLSTM using SIC and OHC.

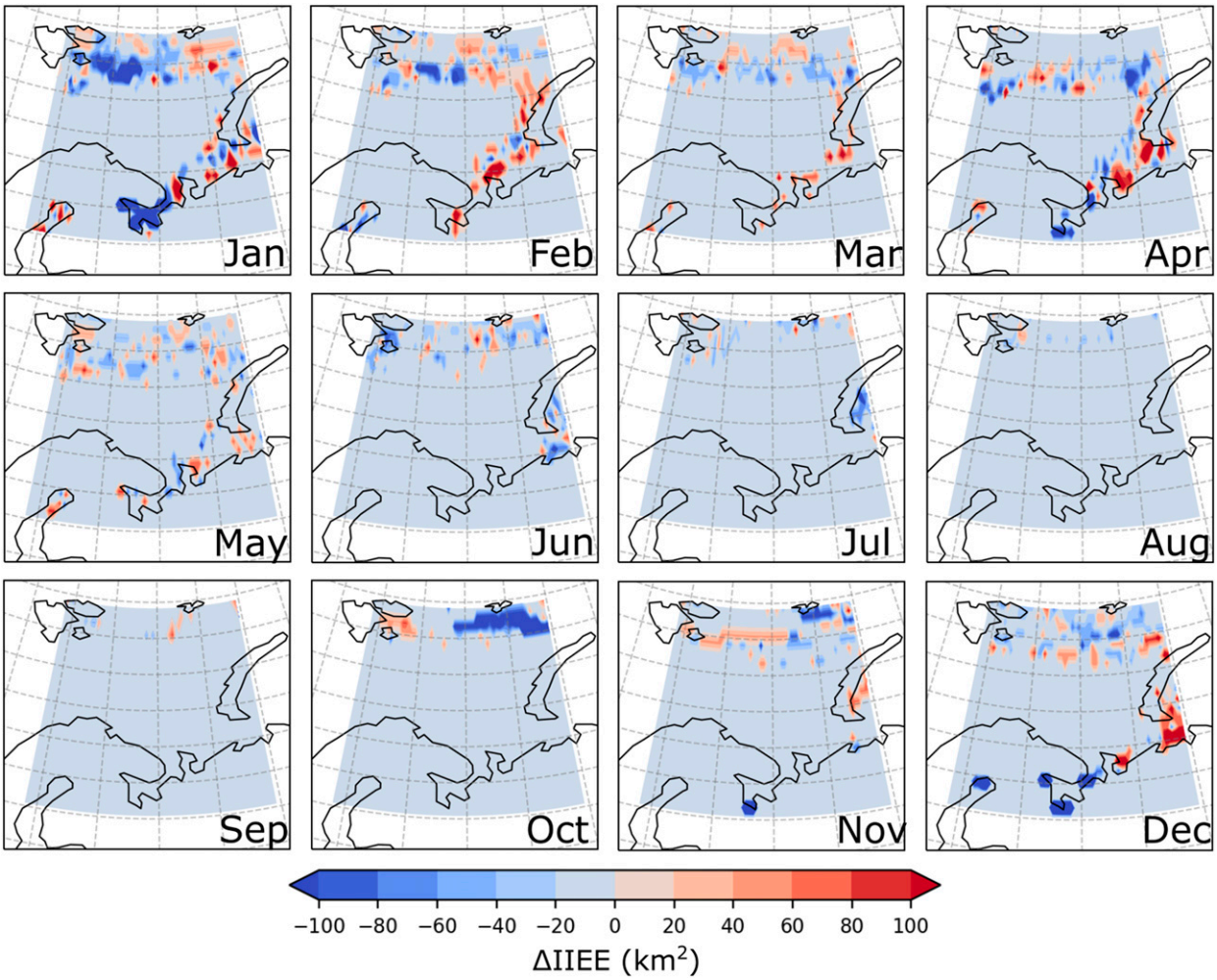

FIG. 5. Difference of the IIEE score of the constrained forecast of SIC for the first week in each month between ConvLSTM and persistence ($IIEE_{ConvLSTM} - IIEE_{persistence}$). The SIC forecast with ConvLSTM uses SIC and OHC fields.

winter and spring, but not for summer and early autumn. Similarly, it was demonstrated by Goessling et al. (2016) that the predictability of sea ice edge is relatively low from summer to autumn at monthly time scales over the whole Arctic. However, for the Barents Sea, it was shown by some studies that the initialized seasonal forecasting models could show monthly skill in the autumn (NDJ) up to 9 months in advance (e.g., Bushuk et al. 2017). The better performance of the sea ice forecast with the ConvLSTM in winter and during the transition seasons shows that the ConvLSTM is able to capture the intricate sea ice variability in these seasons. This is not so surprising given the significant role of SSTs on sea ice predictability in this region in fall and winter (e.g., Guemas et al. 2016). However, given the large ice free area in summer from 2013 to 2016 in the Barents Sea, the worse performance of the ConvLSTM than persistence seems to imply that ConvLSTM tend to overpredict sea ice in summer. Similar results are provided by forecast with the baseline statistical model and forecasts using the ConvLSTM. Considering the large difference between RMSE in each calendar month, it is not surprising that the spread of RMSE in Fig. 3a is very large (see Table S2 in the supplemental material).

The seasonal differences in forecast skill at more lead times for the constrained forecasts are shown in Figs. S4–S6 in the supplemental material. Similar to Fig. 3b but with different lead times, these figures show that, in general, the error increases with the increase of lead time, especially for the early spring around March. For the ConvLSTM forecast with all the chosen predictors, the reduction of skill is relatively smaller than the others. Forecasts with the ConvLSTM cannot beat persistence in autumn starting from October regardless of the lead time. But they are always better than the persistence in the transition, from spring to summer. At lead time step 6, forecasts with the ConvLSTM behave similar as the climatology in terms of the loss.

To understand the source of forecast error in each month, it is insightful to examine the spatial structure of the error. The spatial distribution of the forecast error with the ConvLSTM

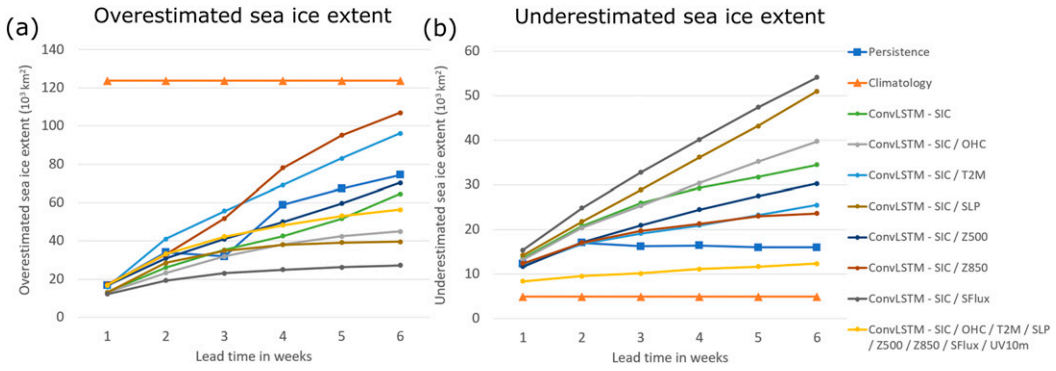

FIG. 6. (a) Overestimated and (b) underestimated local sea ice extent of the constrained forecast of SIC with ConvLSTM using different predictors against persistence and climatology. The unit is square kilometers per grid cell.

using SIC and OHC is plotted in Fig. 4. In winter, the error mainly comes from the coastal area close to the Eurasian continent. Since the boundary sea ice dynamics is relatively complex, it is difficult to forecast, which is also a challenge for most of the operational numerical sea ice forecast systems (e.g., Smith et al. 2013, 2016). In almost all seasons there is also a contribution to the total error from regions with rapid SIC variations near the northern boundary. Physically, processes like sea ice advection, sea ice advance and retreat, oceanic heat transport, and polar air–sea interaction, make forecasting in this area extremely difficult (Årthun et al. 2012; Smith et al. 2016). Note that the cutoff of data around the selected boundary also influences the performance of the ConvLSTM, which was discussed in section 2b.

In practice, it is useful to know whether a region is ice free or not, for instance for shipping and related activities. A binary evaluation was carried out based on the forecast made by the ConvLSTM, based on persistence and climatology. We computed the IIEE score of these forecasts and found that the results are comparable among the forecasts with different methods, except for that with climatology. To assess the forecast skill locationwise, we plot the difference of IIEE score of the SIC constrained forecast between the ConvLSTM (using SIC and OHC) and persistence in Fig. 5. Since the IIEE indicates either an overestimation or underestimation of sea ice forecast, areas with lower scores have better forecast skill. In general, the ConvLSTM has better forecast skill than persistence in almost all months. During the transition time from summer to winter, the ConvLSTM gives more skillful forecasts than persistence, especially around the northern boundary of the Barents Sea. In winter and spring, these two methods have skill in different regions and the ConvLSTM is slightly better. In summer, the differences are small.

We can learn more about the forecast skill of each method by analyzing the overestimated and underestimated sea ice extent separately. The overestimated and underestimated sea ice extent of the constrained forecasts of SIC with the ConvLSTM using different predictors against persistence, and climatology are shown in Fig. 6. Given the definition of these two components of IIEE (section 2c), in combination they can be interpreted as a trade-off between overpredicting and underpredicting. Climatology tends to overpredict the sea ice in this area ($\sim 120 \times 10^3\ \mathrm{km}^2$), but at the same time it has the least underestimated sea ice extent ($\sim 5 \times 10^3\ \mathrm{km}^2$) than the other methods. Almost all the forecasts with the ConvLSTM provide better overestimated component of IIEE score but slightly worse underestimated component of IIEE score than persistence. It reflects that the ConvLSTM learns the temporal evolution of sea ice, especially from an ice-free period to an ice-covered period.

To conclude, the ConvLSTM is able to outperform persistence and climatology when conducting sea ice forecasts at weekly to submonthly time scales. However, a careful selection of input fields should be made in terms of the physical consistency between predictors and sea ice at chosen time scales. An irrelevant predictor could "confuse" the neural network and hence reduce the performance. It is noteworthy that with more training data, the forecast skill of the ConvLSTM can be improved dramatically. This might indicate that ConvLSTM is good at finding nonlinear relations between variables which could eventually contribute to the predictability of the entire system, as long as the network is complicated enough, and the training data are sufficient.

b. *Sensitivity analysis of predictors*

In the previous section, we found that the ConvLSTM is skillful for SIC forecasts using multiple climate fields. The generally better performance of the ConvLSTM than the baseline statistical model indicates that ConvLSTM learns the nonlinear relations between input fields, and sea ice forecast systems for weekly time scales can benefit from such representation of nonlinearity. However, the performance could also be attributed to the change of signal to noise ratio and there is a possibility that the model only learns the noise from these predictors. To understand whether these predictors contribute to the forecast or they only pollute the prediction, we apply Monte Carlo reshuffling to all the chosen predictors with respect to the time sequences and use these reshuffled predictors to forecast SIC. Note that the input SIC sequence is not reshuffled. The result is shown in Fig. 7. It can be noticed

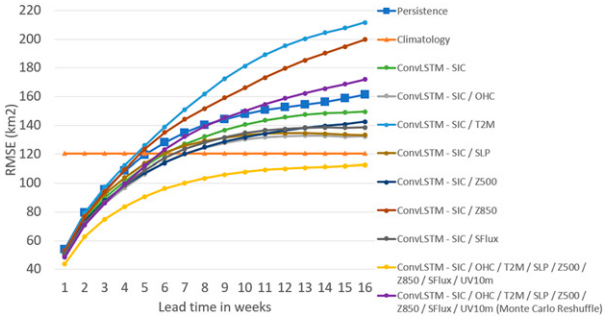

FIG. 7. RMSE of the constrained forecast of SIC with a lead time up to 16 weeks with ConvLSTM using different predictors against persistence and climatology. A ConvLSTM forecast with Monte Carlo reshuffled predictors is included for comparison. The unit is square kilometers per grid cell.

that forecasts with reshuffled predictors are much worse than those without reshuffling, especially for a long lead time. This indicates that these predictors provide useful information for the forecast. For a short period, for instance, up to lead weeks 1 and 2, forecasts with randomly reshuffled series of chosen predictors do not differ much compared to those without reshuffling. This indicates that, in the first 2 weeks, forecasts mainly rely on the memory of sea ice. This also explains why persistence is difficult to beat at weekly time scales. Unfortunately, the information we have is not sufficient to prove whether the improved forecasts benefit from the nonlinear relations between these variables or an increase of signal to noise ratio. Given that these predictors contribute to the predictability of SIC, we can further evaluate the contribution from different predictors to sea ice forecasts by comparing the forecasts using the same structure of the ConvLSTM, but trained with different input fields.

Following the definition of our dimensionless relative forecast skill scores in section 2c, the sensitivity analysis of predictors is shown in Fig. 8. Similar to section 3a, we analyze the forecast skill based on lead-time-dependent constrained forecasts up to 6 weeks, and for the first week in each month. Positive scores indicate an improvement in forecast skill. Note that the skill is relative to the ConvLSTM model that only uses SIC as a predictor. In Fig. 8a, it can be observed that OHC, SFlux, and Z500 add skill to the sea ice forecasts at chosen time scales, while SLP, T2M, and Z850 reduce the skill. Predictability from OHC can be attributed to the long memory of the ocean and the crucial role of OHC in the energy budget, which is also claimed by Guemas et al. (2016); Cruz-García et al. (2019). Furthermore, it is shown in Fig. 8b that OHC has a significant contribution to the forecast skill in summer. Another component of the energy budget, SFlux, also plays an important role here since it has a direct relation to sea ice, that is, the surface energy balance between open ocean and sea ice–covered ocean is very different. We notice that it contributes to the predictability with a lead time more than a week, which is consistent with the lead–lag relations between sea ice melting and the variability of surface fluxes found by Krikken and Hazeleger (2015).

At weekly to submonthly time scales, Fig. 8 shows that some surface fields, like SLP and T2M, and some near-surface atmospheric fields, do not contribute to improving sea ice forecasts with the ConvLSTM. It was reported by Onarheim et al. (2015) and Mohammadi-Aragh et al. (2018) that the chaotic behavior of the atmosphere causes the low predictability of the near surface wind divergence and vorticity, which explains the weak relationship between sea ice variability and SLP. The bad performance of sea ice forecast with the ConvLSTM using SIC and T2M suggests that the surface temperature field is not directly related to the variation of sea ice, especially in summer as shown in Fig. 8b. This is the same for Z850.

In summary, for the extended-range sea ice forecasts in the Barents Sea, meteorological and oceanic fields involved in the energy budget (e.g., SFlux, OHC) can enhance the forecast skill of the sea ice forecasts made by deep neural networks. The same holds for the fields representing the free troposphere (e.g., Z500). However, some surface fields (e.g., T2M, SLP) and lower atmospheric fields (e.g., Z850) do not improve the forecast quality in the selected region. Use of the ConvLSTM can potentially help us understand the nonlinear relationships between multiple selected variables, but limited physical information will be provided by this method. Note that these conclusions about the predictability of sea ice are drawn

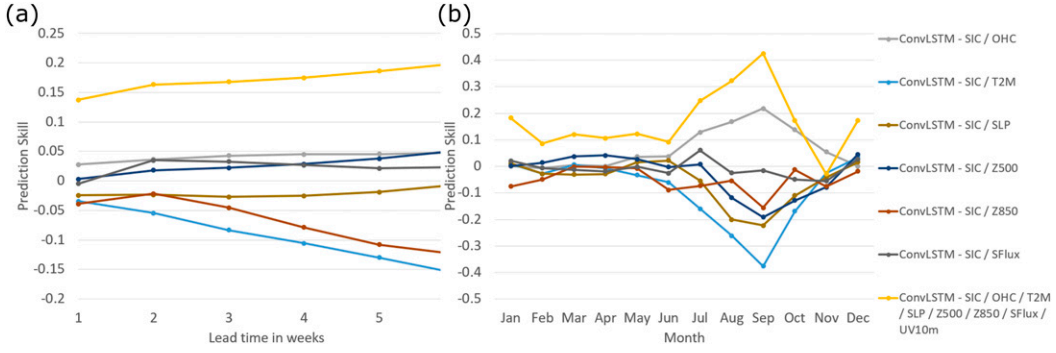

FIG. 8. RMSE-based relative forecast skill improvement of (a) the constrained forecast of SIC with a lead time up to 6 weeks and (b) the constrained forecast of SIC for the first week in each month with ConvLSTM using different predictors.

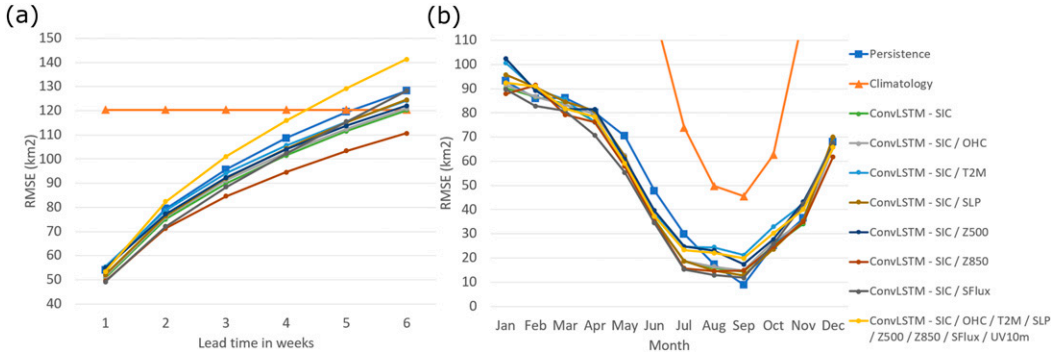

FIG. 9. RMSE of (a) the operational forecast of SIC with a lead time up to 6 weeks and (b) the operational forecast of SIC for the first week in each month with ConvLSTM using different predictors against persistence and climatology. The unit is square kilometers per grid cell.

for specific time scales and region. From this study it is unclear whether it can be generalized to other time scales and locations.

c. *Lead-time-dependent operational forecasts*

In contrast to constrained forecasts, operational weather forecasts can only proceed from a known state to the future state with predicted fields. Therefore, to achieve a "more than one step" sea ice forecast, we have to predict all the input fields and use the predicted fields for further forecasts, specifically with the ConvLSTM. Similar to evaluations of the constrained forecasts given in section 3a, an assessment of the lead-time-dependent forecasts with the ConvLSTM using different combinations of input fields against persistence and climatology are presented in this section. We emphasize that from now on the presented forecasts with the ConvLSTM generate all fields (the same variables as the input) and the loss function also includes these fields. The presented forecasts are reforecasts and these retrospective forecasts can be analyzed using observations.

We first show the RMSE of lead-time-dependent forecasts with different predictors and methods in Fig. 9a (details in Table S4). In general, RMSE increases with the increased lead time, which is similar to Fig. 3a. However, this time, almost all the forecasts with the ConvLSTM outperform the forecasts with persistence, except for the ConvLSTM forecast with all the selected input fields. The ConvLSTM forecast are better than the climatology with lead time from week 1 to week 5, in general. The best forecast is given by ConvLSTM with SIC and Z850, which is very different from the constrained forecast (Fig. 3a). Considering the RMSE in each month (Fig. 9b, details in Table S5), most of the forecasts with the ConvLSTM are better than the persistence forecast in spring and autumn. All the forecasts are comparable in winter but persistence is much better in summer. Also, the forecasts with the ConvLSTM using all given input fields are worse than the ConvLSTM forecasts using fewer variables. The seasonal differences in forecast skill at more lead times for the operational forecasts are shown in Figs. S7–S9 in the supplemental material. The results are analogous to those given by the constrained forecasts. The only difference is that the ConvLSTM forecasts with one extra predictor (e.g., Z850, SFlux) show better skill in sea ice forecasting for most of the time compared to that with all chosen predictors, which is consistent with our analysis regarding Fig. 9.

The decrease in performance of the ConvLSTM with more input variables than that with fewer variables can be attributed to the training process of deep neural networks and the setup of operational forecast. Based on this configuration, the number of input fields are equal to the number of output fields. More input variables mean more output variables and therefore more model parameters, which in turn requires more training data. Consequently, the operational forecast with ConvLSTM is a trade-off between skill gain from predictors and skill loss due to the difficulty in learning. Therefore, it is necessary to choose the right combination of input fields for the ConvLSTM, instead of using all the data in a purely data-driven manner.

The spatial distribution of the forecast error for the first week with the ConvLSTM using SIC and OHC is plotted in Fig. 10. It is very similar to the result in Fig. 4. It is difficult for the system to predict sea ice in winter and spring, and large forecast errors are mainly found in the coastal area and the northern boundary. In summer and autumn, the forecast errors are relatively small. We now consider the binary forecast of sea ice. The difference of the IIEE score of SIC forecasts between the ConvLSTM (using SIC and OHC) and persistence is shown in Fig. 11. Given the similarity of the spatial distribution of the forecast errors between the constrained (Fig. 4) and operational (Fig. 10) forecasts, it is not surprising to find that the ConvLSTM has better forecast skill than persistence in most of the regions in the Barents Sea in almost all months. More explanation on this can be found in the section 3a. Note that starting from the second week the spatial distribution of the forecast errors (RMSE and binary) becomes different for the constrained and operational forecasts with ConvLSTM, but the regions with high forecast skill with the ConvLSTM in these two cases are still the same (not shown).

Compared to the constrained forecast, small differences are found in the overestimated and underestimated components of the IIEE score for the operational forecasts with the ConvLSTM. The overestimated and underestimated local sea

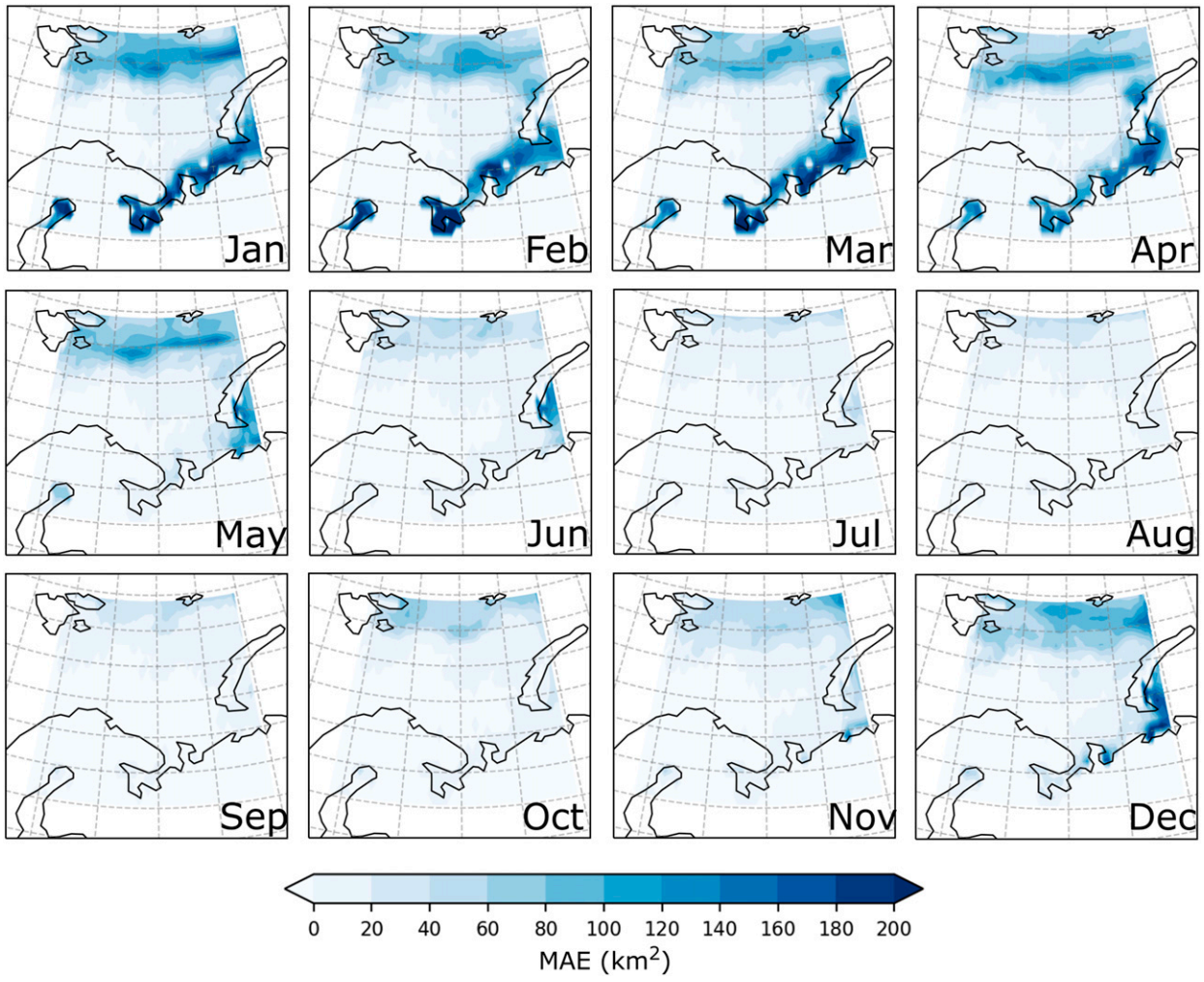

FIG. 10. MAE of the operational forecast of SIC for the first week in each month with ConvLSTM using SIC and OHC.

ice extent of the lead-time-dependent operational SIC forecast with the ConvLSTM using different predictors against persistence and climatology are shown in Fig. 12. Still, most of the forecasts with the ConvLSTM provide better overestimated component of the IIEE score than persistence, except for the ConvLSTM forecast using SFlux and all available predictors. However, for the underestimated component of the IIEE score, persistence has always more skill starting from the third week. Considering both components of the IIEE score, it can be noticed that the forecast with the ConvLSTM using all the input fields significantly overpredicts the sea ice. Again, it shows a caveat that rather than blindly using all the available data, a smart selection of input fields is necessary to improve the sea ice forecast with the current structure of the ConvLSTM.

In addition, we compare the ConvLSTM forecasts with the NCEP and ECMWF ensemble forecasts considering period 2015–16. The results are shown in Fig. 13 (and spatial distribution of error in Fig. S10). It can be noticed that the NCEP ensemble forecast is worse than most of the ConvLSTM forecasts within 4 lead weeks. This may originate from the initialization since the NCEP ensemble forecast is initialized by sea ice conditions from Climate Forecast System Reanalysis (CFSR), and not from the ERA-Interim or any satellite-based observations (e.g., NSIDC/NOAA Passive Microwave SIC). The ECMWF ensemble forecast performs much better than all the ConvLSTM forecasts for more than 2 weeks ahead. Note that since the forecasts are evaluated against ERA-Interim reanalysis and ECMWF ensemble forecasts use the same IFS and similar configurations as ERA-Interim, it is not completely a fair comparison. The forecast error grows slower for the ECMWF ensemble forecast than the ConvLSTM forecasts. Since for the first 2 weeks memory in sea ice has significant impact on its variations, it reflects that forecasts with ConvLSTM considerably rely on the memory of sea ice, and the chosen numerical model can preserve the physical consistency and therefore may provide better forecasts even for large lead times. In general, at the lead time up to 2 weeks, the forecasts with ConvLSTM are comparable to the best NWP-based forecasts in S2S project.

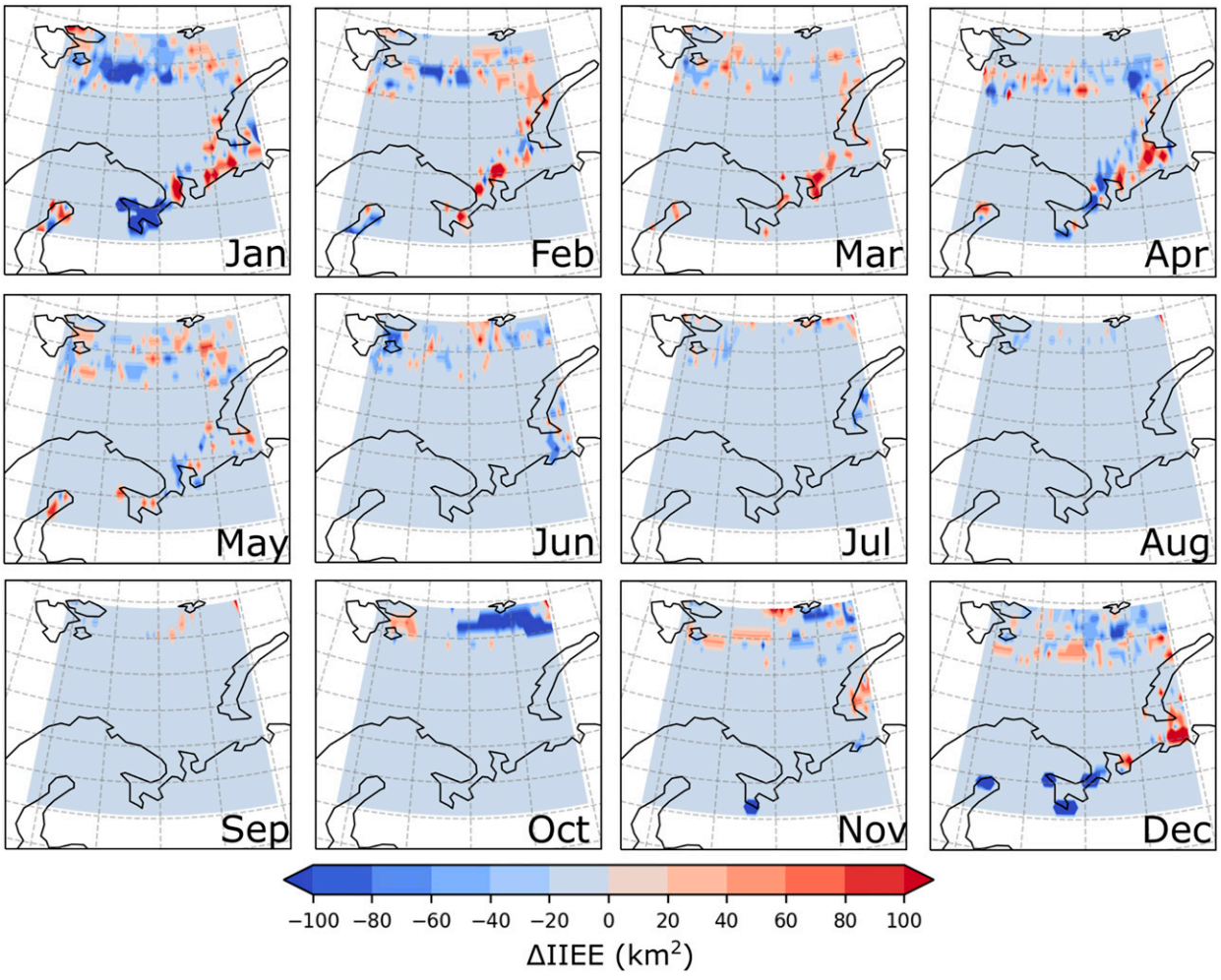

FIG. 11. Difference of the IIEE score of the operational forecast of SIC for the first week in each month between ConvLSTM and persistence ($IIEE_{ConvLSTM} - IIEE_{persistence}$). The SIC forecast with ConvLSTM uses SIC and OHC fields. The unit is square kilometers per grid cell.

So far we have not discussed the forecast quality for various start dates. It is useful to know if either there is a substantial dependence of forecast errors on climate variability and change, or whether the variations between forecast cases (start dates) are large in comparison to the differences between forecast models (ConvLSTM forecasts with different combination of input fields). The former one can also be interpreted as whether the whole temporal consistency is needed or not for a ConvLSTM.

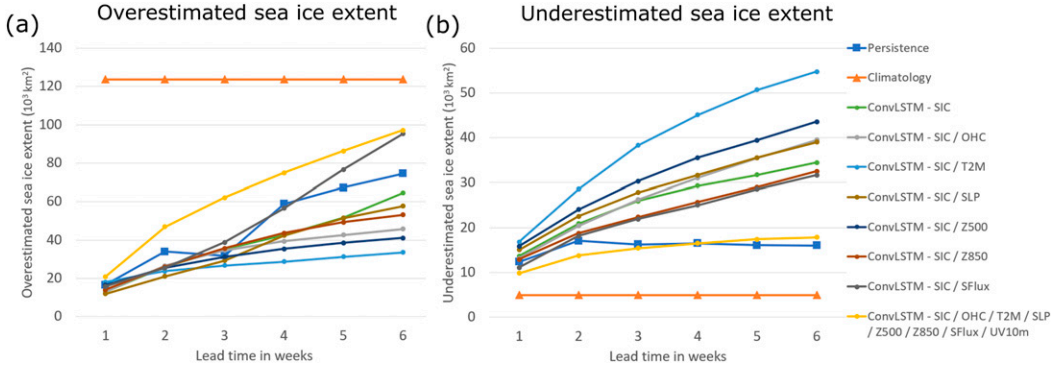

FIG. 12. (a) Overestimated and (b) underestimated local sea ice extent of the operational forecast of SIC with ConvLSTM using different predictors against persistence and climatology. The unit is square kilometers per grid cell.

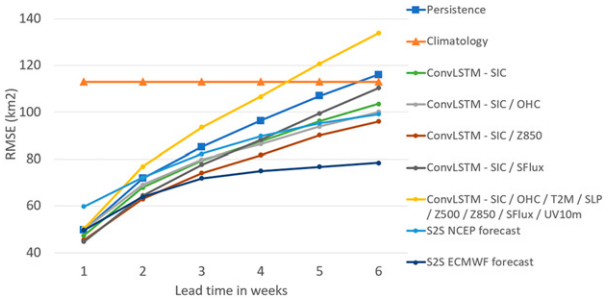

FIG. 13. RMSE of the operational forecast of SIC between 2015 and 2016 with a lead time up to 6 weeks with ConvLSTM using different predictors against persistence, climatology, NCEP ensemble forecast, and ECMWF ensemble forecast. The NCEP real-time forecast and ECMWF ensemble forecast are provided by the subseasonal to seasonal prediction project (S2S archive). The unit is square kilometers per grid cell.

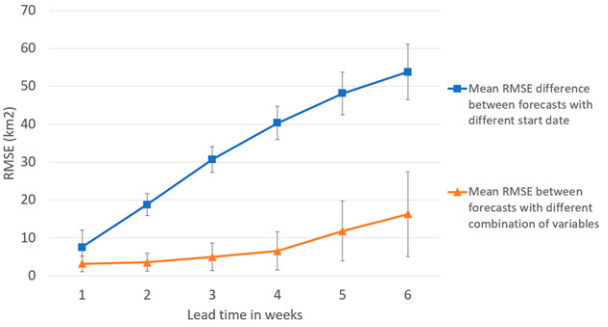

FIG. 14. Comparison between the mean difference of errors among forecasts with different start dates and that among forecasts with different combination of input variables based on reversed operational forecast 1979–82 and operational forecast 2013–16, as a function of lead time. The mean difference of errors between forecasts with different start dates were computed as the mean absolute difference of RMSE between reversed operational forecast 1979–82 and operational forecast 2013–16 with the same input fields $[(1/7)\sum_{n=1}^{7}(|\mathrm{RMSE}_{\mathrm{ConvLSTM-SIC/field(2013-16)}} - \mathrm{RMSE}_{\mathrm{ConvLSTM-SIC/field(1979-82)}}|)$, with SIC/field indicating different combinations of input fields shown in Fig. 9 and Fig. S11], while the mean difference of RMSE between forecasts with different input fields were obtained from reversed operational forecast 1979–82 and operational forecast 2013–16 $[(1/14)\sum_{n=1}^{14}(|\mathrm{RMSE}_{\mathrm{ConvLSTM-SIC/field}} - \mathrm{RMSE}_{\mathrm{ConvLSTM-SIC}}|)$, concerning the forecasts in Fig. 9a and Fig. S11]. The standard deviation at each lead week is included.

To address this we conducted an extra experiment in which we performed Monte Carlo subsampling of the reanalysis data with a 4-yr period. We randomly selected five periods (20 years of data) to train the network and four periods (4 years for each) as valid date. The results indicate that, compared to the climatology and persistence, these ConvLSTM forecasts do not show any skill considering each start date at all lead weeks (not shown). This reflects that this particular ConvLSTM was not able to determine the state of the system and therefore could not provide reliable forecasts. Indeed, this is not so surprising as forecasts with the ConvLSTM are strongly dependent on memory and the temporal order in the data must be preserved during training and predicting. To gain more insight into the dependency of forecasting on the temporal consistency, we performed a follow-up test in which we provided the model with 20 years of data following temporal order and then applied Monte Carlo subsampling with the same setup. Also, this model gained no skill (not shown). Only when the full time series preceding the valid date were fed to the model, the ConvLSTM started to generate meaningful forecasts. This again demonstrates that forecasts using ConvLSTM rely on the correct temporal order to reproduce and skillfully forecast the state of the system. This indicates that the common workflow used by NWP systems to determine skill for a range of start dates, and therefore an error estimate of forecast quality for a specific forecast system, cannot be transferred meaningfully to ConvLSTM, given the insufficient length of training data. This would only be possible with sufficient length of training data such that the memory can be maintained in the samples.

There is another way to test the state dependency and address the significance of the conclusions. Given the susceptibility of forecasts to the length of the training set, and a strict requirement on the temporal order of input fields during training and forecasting, we launched another experiment with a reversed time series of the reanalysis data, and chose data from 1979 to 1982 as testing set, data from 1983 to 1986 as cross-validation set, and the rest as training set. The results are shown in Fig. S11 in the supplemental material. Despite slightly different skill compared to those shown in Fig. 9a, ConvLSTM forecasts with different combinations of input fields exhibit similar errors. From now on we will refer to this experiment as ''reversed operational forecast 1979–82'', and the former experiment with time series following the correct temporal order as ''operational forecast 2013–16'' (Fig. 9a).

By comparing these two operational forecasts, we can evaluate the skill of ConvLSTM for different start dates (between 1979–82 and 2013–16) and compare it to the skill of ConvLSTM related to the choices of input fields (e.g., between ConvLSTM-SIC/OHC and ConvLSTM-SIC in Fig. 9a) in a relatively fair manner, as we can use the same amount of training data and both systems show comparable forecast skill, although the setups are slightly different. We highlight this comparison quantitatively in Fig. 14. The mean difference of errors between forecasts with different start dates were computed as the mean absolute difference of RMSE between reversed operational forecast 1979–82 and operational forecast 2013–16 with the same input fields $[(1/7)\sum_{n=1}^{7}(|\mathrm{RMSE}_{\mathrm{ConvLSTM-SIC/field(2013-16)}} - \mathrm{RMSE}_{\mathrm{ConvLSTM-SIC/field(1979-82)}}|)$, with SIC/field indicating different combinations of input fields shown in Figs. 9a and S11], while the mean difference of RMSE between forecasts with different input fields were obtained from reversed operational forecast 1979–82 and operational forecast 2013–16 $[(1/14)\sum_{n=1}^{14}(|\mathrm{RMSE}_{\mathrm{ConvLSTM-SIC/field}} - \mathrm{RMSE}_{\mathrm{ConvLSTM-SIC}}|)$, concerning the forecasts in Figs. 9a and S11]. These two cases have errors in skill that are similar. Therefore, these experiments indicate that the ConvLSTM forecasts when

properly trained are skillful and their forecast quality is state-dependent just like forecasts with numerical models. The mean differences between forecasts with different input fields are smaller than those related to different start dates, but in general they are of the same amplitude. Note that the differences in skills between the forecasts of periods 1979–82 and 2013–16 may also originate from the data richness and quality linked to these two different eras.

To conclude, similar to constrained forecasts, the operational forecasts with the ConvLSTM provide better results than persistence in most of the cases. However, unlike the constrained forecasts, the coherence between the way of learning (loss function and an increase in the number of trainable parameters) and the way of predicting (forecast based on predicted fields) with a neural network places a requirement for appropriate choices of input variables, which may a priori not be clear. However, in practice, it is natural to select predictors related to their physical consistency with the predicted variables. The results confirm this, with higher contributions from OHC and SFlux and less clear contributions from more atmospheric characteristics. It is noteworthy that forecasts with ConvLSTM shows state-dependency and the start dates have impact on the forecast quality, just like forecasts using numerical models.

d. *Physical consistency of ConvLSTM forecasts*

Deep learning techniques are often considered as brute force approaches or black-box methods. There is a valid concern that physical laws may be violated and physical relationships may not hold in the neural network. There is no constraint from first principle in its formulation. With this in mind, we are interested in the physical interpretation of our forecasts with the ConvLSTM. Our first impression comes from the mathematical description of the ConvLSTM [Eq. (1)], which includes both linear and nonlinear operations. The convolutional operations are all linear (e.g., $\mathbf{W}_{xi}{*}x_t$), and this will maintain the physical consistency of input fields, at least their linear relations. However, the nonlinear behavior caused by the use of sigmoid functions and hyperbolic tangent functions could potentially introduce further physical inconsistency in the forecasts.

To evaluate whether the physical links between predicted fields are preserved by the forecasts with the ConvLSTM, we performed singular value decomposition (SVD) on the covariance map of two predicted fields (e.g., SIC and OHC) within training sets (1979–2008), testing sets (2013–16), and forecast data (2013–16) (Bretherton et al. 1992). Since this method searches for the maximum covariance between given variables, it is also known as the maximum covariance analysis (MCA) (Frankignoul et al. 2011). We first show three SVD modes of the covariance map between SIC and OHC in Fig. 15. These fields are expected to be related and OHC varies relatively slowly. The first modes explain over 99% of covariance, and they mainly represent the trend and climatology, which can be regarded as fairly trivial. Therefore, it is worthwhile to evaluate the second and third modes. In general, the forecast data show similar patterns for both SIC and OHC as training sets and testing sets in all three SVD modes. This indicates that forecasts with ConvLSTM are able to preserve the physical links between given fields, which reflects the source of predictability within chosen fields and the forecast skill of chosen neural networks. This is further confirmed by the projection of SVD modes of the covariance map between SIC and OHC on the actual time series of SIC and OHC, which is shown in Fig. S12 in the supplemental material.

Similar to the SVD of covariance map between SIC and OHC, we also inspected other variables, for instance, the SVD and SVD projection of the covariance map between SIC and Z500 (Fig. 16 and S13). In this case, physically interpretable results are obtained in the training and testing datasets. However, the forecasts with the ConvLSTM fails to provide similar coupled patterns of Z500 and SIC. In particular Z500 seems like a spurious pattern. Given the chaotic behavior of the atmospheric circulation and the related lack of predictability at extended range, it is not so surprising that the predictability of Z500 at weekly time scales is relatively low in the ConvLSTM (e.g., Hohenegger and Schar 2007). Also, the distorted shape of the patterns suggests that the cutoff of input fields around the boundary also influences the physical consistency learned by the ConvLSTM (see the SVD mode patterns of Z500 around the boundary in Figs. 16 and S13).

We also checked the SVD of the covariance map between SIC and SFlux, and that between SIC and Z850. The results are shown in Figs. S14 and S15 in the supplemental material. Since these decompositions were all based on the forecast of lead week 1, the results do not differ substantially.

To conclude, depending on the physically interpretable predictability of chosen meteorological and oceanic fields, the ConvLSTM is able to preserve a realistic physical consistency between various predictors and predictands during forecasts. It should be noted that such analysis based on SVD can only account for the linear relations of the covariance between given fields. To further evaluate the physical consistency during forecast with deep neural networks, a choice of meteorological and oceanic fields could be made based on some lead–lag analysis between potential fields, similar to Krikken and Hazeleger (2015), but at shorter time scales than they considered. Also, other clustering techniques can be used that are not limited to linear relationships.

4. Discussion

In this study, we introduce a deep neural network to predict the sea ice in the Barents Sea. With the intention to adapt this approach to the Arctic sea ice forecast problem and further assess its capabilities, all the computations were performed with reanalysis datasets. It is noteworthy that our results with the ConvLSTM are comparable to those with the state-of-the-art numerical climate models (Van Woert et al. 2004; Metzger et al. 2014; Hebert et al. 2015; Smith et al. 2013, 2016). For instance, our forecast errors with the ConvLSTM using SIC and OHC shown in Fig. 10 are comparable to those with the latest Global Ice Ocean Prediction System (GIOPS) employed

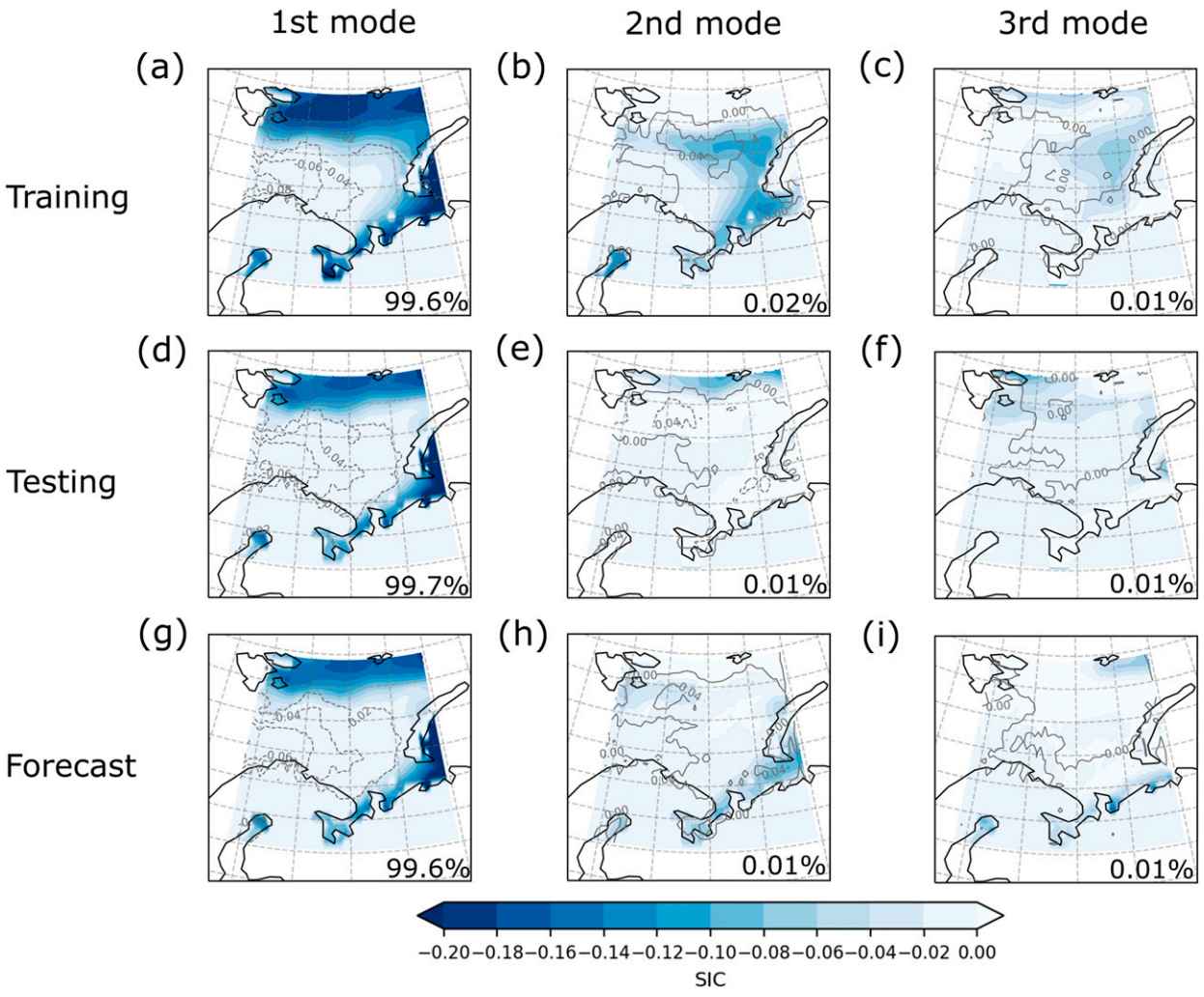

FIG. 15. Covariance map of SIC and OHC for the (a),(d),(g) first; (b),(e),(h) second; and (c),(f),(i) third SVD modes in (a)–(c) training, (d)–(f) testing, and (g)–(i) forecast data for the first week, with shades indicating the dimensionless SIC and contour lines indicating the dimensionless OHC. The SVD was performed on the covariance matrix of normalized SIC and OHC.

by the Canadian Meteorological Centre (see their Figs. 4–7) (Smith et al. 2016). We also inspect the errors based on the normalized sea ice concentration and find a competitive result (e.g., Fig. S16 in the supplemental material). However, we only include a brief assessment of the forecast skill between ConvLSTM and numerical weather forecast systems (NCEP and ECMWF ensemble forecast in S2S project) and an extended quantitative comparison between them is beyond our scope. For future work, it is recommended to evaluate ConvLSTM forecasts against numerical weather and seasonal forecast systems consistently using the same observation dataset and an extensive range of metrics.

Typically, operational weather forecast systems are ensemble forecast systems in order to sample several sources of uncertainty (Gneiting et al. 2007). It is also possible for the deep neural networks to model the uncertainty by either employing the deep learning based ensemble approach (e.g., an ensemble of deep neural networks) (Zaier et al. 2010; Wang et al. 2017), or implementing probabilistic deep neural networks (e.g., Vandal et al. 2018; McDermott and Wikle 2019, with Bayesian deep learning). For the deep learning based ensemble approach, the ensemble is generated through perturbing the structure of neural networks (e.g., number of hidden layers, filter size) and the size of input sequences, and therefore very difficult to control. The latter is more common in practice and this technique is generally known as Bayesian deep learning (BDL) (Blundell et al. 2015; Fortunato et al. 2017; Kendall and Gal 2017; Shridhar et al. 2019). With Bayesian deep learning, a deterministic NN can easily be transformed into a probabilistic NN, namely a Bayesian neural network (BNN), by replacing the weight with a distribution. Through sampling the distribution, an ensemble forecast can be generated to represent the uncertainty in the forecast to avoid overconfident forecasting. These techniques are able to capture both aleatoric and epistemic uncertainty, but they are very expensive (Kendall and Gal 2017).

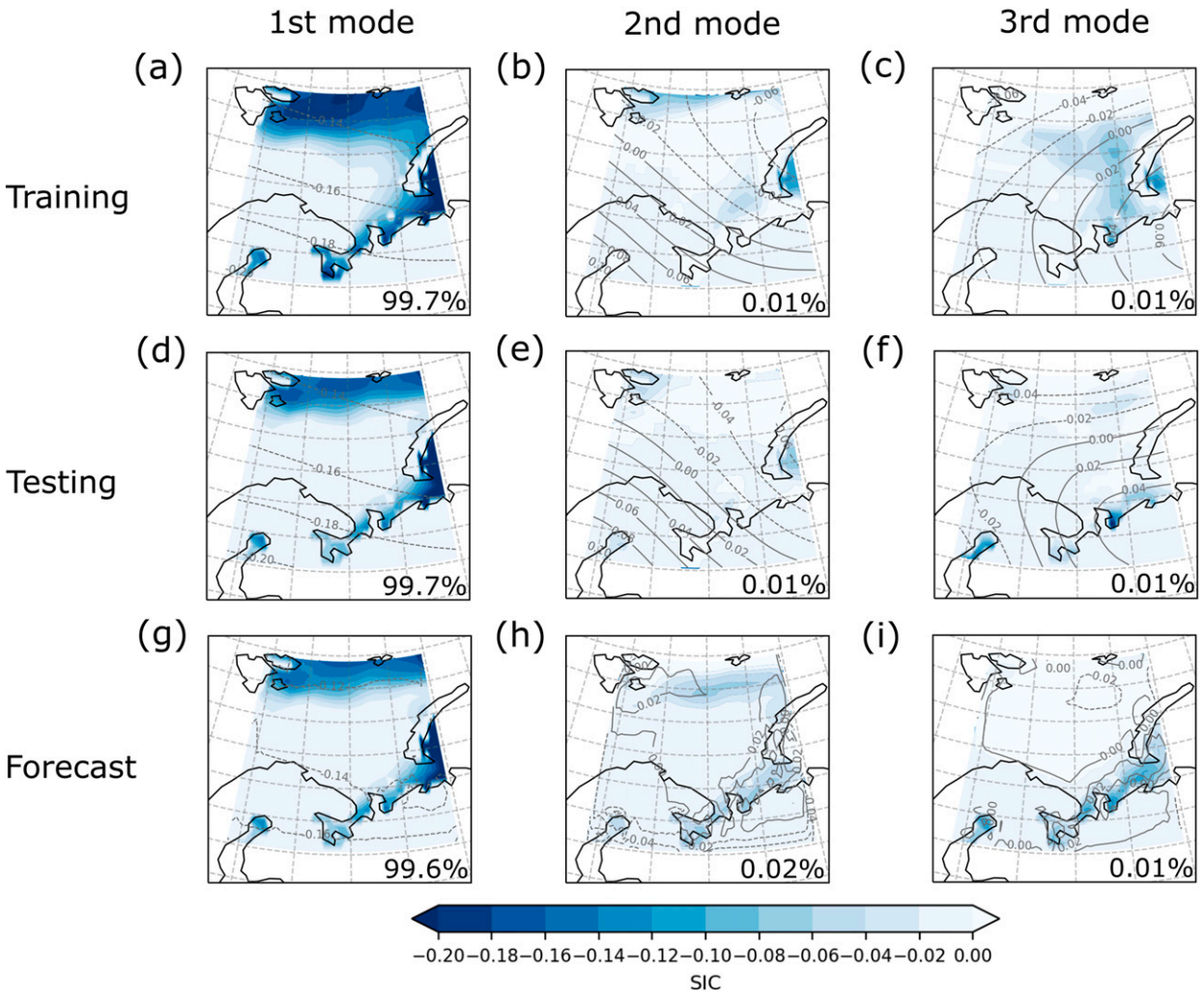

FIG. 16. Covariance map of SIC and Z500 for the (a),(d),(g) first; (b),(e),(h) second; and (c),(f),(i) third SVD modes in (a)–(c) training, (d)–(f) testing, and (g)–(i) forecast data for the first week, with shades indicating the dimensionless SIC and contour lines indicating the dimensionless Z500. The SVD was performed on the covariance matrix of normalized SIC and Z500.

This study is primarily a proof of concept of the capabilities of ConvLSTM, which goes beyond the extensive use of LSTMs in modern meteorological forecast literature. For this study, incorporating an ensemble or Bayesian network was beyond scope, but we will study it in our following work.

Due to our limited access to the computational resources, the experiments with the ConvLSTM were carried out with a relatively simple network structure and a small dataset. Given the nature of deep neural networks, there is still a lot of room for improvement by increasing the complexity of the neural network (e.g., add more LSTM layers), including more input fields which are physically consistent with sea ice variability (e.g., separate terms in the surface energy budget), enlarging the input area to reduce the boundary effect, training separate models for the forecasts with different lead time to avoid the accumulation of forecast errors and account for the teleconnections between remote areas. Theoretically, a more complicated deep neural network serves as a better representation of nonlinearity between variables, thus the nonlinear relationship between all the meteorological fields in a chaotic climate system. Our study has shown that this is far from straightforward as the inclusion of noisy atmospheric fields deteriorated the forecast skill.

Nevertheless, it should be noted that, compared to numerical weather forecast systems, the deep-learning-based methods still have drawbacks. For instance, these methods always have a very limited and specialized forecast target and they are not flexible after expensive training and tuning processes. Moreover, such data driven approaches normally require a large training set and therefore the left validation data may not be enough to allow for any statistical significance tests. However, in terms of the fast development of deep learning techniques and the corresponding evolution of hardware, these deep learning approaches could potentially enhance weather forecasts in the near future.

5. Conclusions

A ConvLSTM is a useful tool to incorporate both the spatial and temporal information within meteorological fields in a model. In this work, we demonstrate that this deep neural network approach can effectively be used to predict sea ice characteristics in the Barents Sea at weekly to submonthly time scales. Different combinations of meteorological fields were tested as input variables to train the neural networks and make forecasts. Sea ice forecasts were evaluated against climatology, persistence, and the baseline statistical model. They were also compared to operational subseasonal to seasonal forecast systems. It is found that in most cases, a ConvLSTM can outperform persistence and climatology for the extended-range sea ice forecast up to lead week 5. However, the choices of input meteorological fields should be made in a smart way based on the physical consistency between sea ice variation and the variability of predictors. Being dependent on the predictability of chosen meteorological fields, an interpretation of forecasts with the ConvLSTM in terms of observed linear relationships indicates that the ConvLSTM is able to preserve the physical consistency between various predictors and predictands. In addition, it should be noted that forecasts with ConvLSTM shows state-dependency and the start dates have impact on the forecast quality, just like forecasts using numerical models.

Both lead-time-dependent constrained and retrospective forecasts were performed in this study and their results are slightly different due to the criterion in the loss function. Moreover, sensitivity tests were conducted based on the lead-time-dependent constrained forecasts and we notice that energy budget related fields can add skill to the sea ice forecasts in the Barents Sea at weekly to submonthly time scales. Krikken and Hazeleger (2015) also found that OHC and SFlux have a strong impact on the predictability of the Arctic sea ice. This indicates the important role of the energy budget components in the variability of Arctic sea ice and poses a request for accurate and reliable quantification of the energy budget and a deep understanding of the energy balance in the climate system (Liu et al. 2020a,b). Moreover, the fields representing the midtroposphere also have an impact on the predictability of sea ice in the Barents Sea. In contrast, the atmospheric circulation fields close to the surface seem to deteriorate the performance of the deep neural network, thus reducing the forecast skill. Although such deep neural networks are ''black boxes,'' they can potentially help us gain more knowledge about the nonlinear relationship between multiple meteorological fields.

So far, our experiments with ConvLSTM are limited to simple network structures and small training datasets. It is possible to improve the performance of ConvLSTM by increasing the complexity of the network and including large datasets for training. Fortunately, such attempts will hardly influence the time to obtain a forecast, but the training cost will increase. Compared to the relatively expensive operational numerical forecast systems based on numerical weather models, this method with the ConvLSTM could be suitable for forecasts where time to produce the forecast is limited. In addition, the forecasts made by the ConvLSTM can also potentially be assimilated by the operational weather forecast systems to improve their robustness and reliability. Given its advantages over the conventional way of weather forecast, as well as the fast development of deep learning techniques and the corresponding evolution of hardware, this method is promising to serve as an additional fast and cost-efficient operational sea ice forecast component of a forecast system in the future.

Acknowledgments. The authors gratefully acknowledge the support by the Netherlands eScience Center and Wageningen University. This study is supported by Blue Action project (European Union's Horizon 2020 research and innovation program, Grant 727852). We thank SURFsara (Netherlands) for providing us their super computing infrastructure for our project. We also acknowledge the editor Dr. Josh Hacker, reviewer Dr. Steffen Tietsche, and another anonymous reviewer for their help to improve the manuscript.

APPENDIX

Numerical Methods—ConvLSTM

The whole numerical processes of training the ConvLSTM and making forecasts with ConvLSTM are illustrated in detail in this part. We train ConvLSTM with the entire time series of the training set (1979–2008), which is a matrix with a dimension of $X \times 24 \times 56 \times 1440$ (X indicates the number of input fields). The whole time series are fed into the ConvLSTM time step by time step. Each time step contains $X \times 24 \times 56$ points and the convolution takes place in these input layers [e.g., $\mathbf{W}_{xi} * x_t$ in Eq. (1)]. During the convolution, the spatial structure (24×56) is preserved by using paddings. After the convolutional processes, the output (dimension $Y \times 24 \times 56$, with Y indicating number of channels) is fed into the LSTM layer and it will pass through the input gate (i_t) in LSTM [see Eq. (1)]. The cell state (c_t) will be updated if this input gate is activated. At the same time, the forget gate (f_t) will decide if the previous cell state (c_{t-1}) need to be forgotten or not. Finally, the output gate (o_t) will determine whether the hidden state (h_t) shall include contributions from the current cell state(c_t). It will provide a forecast for the next time step, and the hidden state (h_t) and cell state (c_t) will be passed to the next time step for LSTM related computations. The forecast will be evaluated using the chosen loss function (MSE in this case, see section 2d) and it will provide a training error for this time step.

This process will be repeated until every time step in the time series of the train set has been fed into the ConvLSTM. Then the aggregation of the training error from each time step will be used to perform the back-propagation and one epoch of training is complete by now. The whole training procedure will be repeated until the required number of epoch is reached.

The convolutions take place though the implementation of filters inside convolutional cells before the start of LSTM processes. Therefore, the convolutions are included in the recurrence. By introducing convolutional layers, communications between adjacent cells are enabled and flow of spatial

information is allowed between time steps. In other words, the cross correlations between neighboring nodes are learnt by the neural network, both spatially and temporally. Multiple LSTM layers can be stacked to improve the complexity of the network.

The forecast (from 2013 to 2016 with testing set $X \times 24 \times 56 \times 192$) procedure is the same as above, except for the back-propagation processes.

REFERENCES

Årthun, M., T. Eldevik, L. Smedsrud, Ø. Skagseth, and R. Ingvaldsen, 2012: Quantifying the influence of Atlantic heat on Barents sea ice variability and retreat. *J. Climate*, **25**, 4736–4743, https://doi.org/10.1175/JCLI-D-11-00466.1.

Balmaseda, M. A., K. Mogensen, and A. T. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quart. J. Roy. Meteor. Soc.*, **139**, 1132–1161, https://doi.org/10.1002/qj.2063.

Berrisford, P., D. Dee, K. Fielding, M. Fuentes, P. Kallberg, S. Kobayashi, and S. Uppala, 2009: The Era-Interim archive: Version 1.0. ERA Rep. Series 1, 16 pp., https://www.ecmwf.int/en/elibrary/8173-era-interim-archive.

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

——, C. Bitz, and M. Holland, 2011b: Influence of initial conditions and climate forcing on predicting Arctic sea ice. *Geophys. Res. Lett.*, **38**, L18503, https://doi.org/10.1029/2011GL048807.

Blundell, C., J. Cornebise, K. Kavukcuoglu, and D. Wierstra, 2015: Weight uncertainty in neural networks. https://arxiv.org/abs/1505.05424.

Bonan, D. B., M. Bushuk, and M. Winton, 2019: A spring barrier for regional predictions of summer Arctic sea ice. *Geophys. Res. Lett.*, **46**, 5937–5947, https://doi.org/10.1029/2019GL082947.

Bretherton, C. S., C. Smith, and J. M. Wallace, 1992: An intercomparison of methods for finding coupled patterns in climate data. *J. Climate*, **5**, 541–560, https://doi.org/10.1175/1520-0442(1992)005<0541:AIOMFF>2.0.CO;2.

Bushuk, M., R. Msadek, M. Winton, G. A. Vecchi, R. Gudgel, A. Rosati, and X. Yang, 2017: Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.*, **44**, 4953–4964, https://doi.org/10.1002/2017GL073155.

Chattopadhyay, A., P. Hassanzadeh, and D. Subramanian, 2020: Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: Reservoir computing, artificial neural network, and long short-term memory network. *Nonlinear Processes Geophys.*, **27**, 373–389, https://doi.org/10.5194/npg-27-373-202.

Cruz-García, R., V. Guemas, M. Chevallier, and F. Massonnet, 2019: An assessment of regional sea ice predictability in the Arctic Ocean. *Climate Dyn.*, **53**, 427–440, https://doi.org/10.1007/s00382-018-4592-6.

Dee, D. P., and Coauthors, 2011: The Era-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Ferry, N., and Coauthors, 2012: Nemo: The modeling engine of global ocean reanalyses. *Mercator Ocean Quarterly Newsletter*, No. 46, 46–59.

Fortunato, M., C. Blundell, and O. Vinyals, 2017: Bayesian recurrent neural networks. https://arxiv.org/abs/1704.02798.

Frankignoul, C., N. Chouaib, and Z. Liu, 2011: Estimating the observed atmospheric response to SST anomalies: Maximum covariance analysis, generalized equilibrium feedback assessment, and maximum response estimation. *J. Climate*, **24**, 2523–2539, https://doi.org/10.1175/2010JCLI3696.1.

Fukushima, K., 1980: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, **36**, 193–202, https://doi.org/10.1007/BF00344251.

Gascard, J.-C., K. Riemann-Campe, R. Gerdes, H. Schyberg, R. Randriamampianina, M. Karcher, J. Zhang, and M. Rafizadeh, 2017: Future sea ice conditions and weather forecasts in the arctic: Implications for arctic shipping. *Ambio*, **46**, 355–367, https://doi.org/10.1007/s13280-017-0951-5.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Goessling, H. F., S. Tietsche, J. J. Day, E. Hawkins, and T. Jung, 2016: Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.*, **43**, 1642–1650, https://doi.org/10.1002/2015GL067232.

Gope, S., S. Sarkar, P. Mitra, and S. Ghosh, 2016: Early prediction of extreme rainfall events: A deep learning approach. *Industrial Conf. on Data Mining*, New York, NY, Springer, 154–167.

Guemas, V., and Coauthors, 2016: A review on arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quart. J. Roy. Meteor. Soc.*, **142**, 546–561, https://doi.org/10.1002/qj.2401.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, https://doi.org/10.1038/s41586-019-1559-7.

Hebert, D. A., R. A. Allard, E. J. Metzger, P. G. Posey, R. H. Preller, A. J. Wallcraft, M. W. Phelps, and O. M. Smedstad, 2015: Short-term sea ice forecasting: An assessment of ice concentration and ice drift forecasts using the U.S. Navy's Arctic cap nowcast/forecast system. *J. Geophys. Res. Oceans*, **120**, 8327–8345, https://doi.org/10.1002/2015JC011283.

Hochreiter, S., and J. Schmidhuber, 1997: Long short-term memory. *Neural Comput.*, **9**, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

Hohenegger, C., and C. Schar, 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1794, https://doi.org/10.1175/BAMS-88-11-1783.

Howell, C., M. Richard, J. Barnes, and T. King, 2015: Short-term operational sea ice forecasting for Arctic shipping. *ASME 2015 34th Int. Conf. on Ocean, Offshore and Arctic Engineering*, Mombetsu, Hokkaido, Japan, American Society of Mechanical Engineers Digital Collection, 197–199.

Kendall, A., and Y. Gal, 2017: What uncertainties do we need in Bayesian deep learning for computer vision? *31st Conf. on Neural Informing Processing System*, Long Beach, CA, NIPS, 5580–5590.

Kim, S., S. Hong, M. Joh, and S.-k. Song, 2017: DeepRain: ConvLSTM network for precipitation prediction using multichannel radar data. https://arxiv.org/abs/1711.02316.

——, H. Kim, J. Lee, S. Yoon, S. E. Kahou, K. Kashinath, and M. Prabhat, 2019: Deep-hurricane-tracker: Tracking and forecasting extreme climate events. *2019 IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa, HI, IEEE, 1761–1769.

Kim, Y. J., H.-C. Kim, D. Han, S. Lee, and J. Im, 2020: Prediction of monthly arctic sea ice concentrations using satellite and reanalysis data based on convolutional neural networks. *Cryosphere*, **14**, 1083–1104, https://doi.org/10.5194/tc-14-1083-2020.

Knüsel, B., M. Zumwald, C. Baumberger, G. H. Hadorn, E. M. Fischer, D. N. Bresch, and R. Knutti, 2019: Applying big data beyond small problems in climate research. *Nat. Climate Change*, **9**, 196–202, https://doi.org/10.1038/s41558-019-0404-1.

Krikken, F., and W. Hazeleger, 2015: Arctic energy budget in relation to sea ice variability on monthly-to-annual time scales. *J. Climate*, **28**, 6335–6350, https://doi.org/10.1175/JCLI-D-15-0002.1.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *Nature*, **521**, 436–444.

Leppäranta, M., V. P. Meleshko, P. Uotila, and T. Pavlova, 2020: Sea ice modelling. *Sea Ice in the Arctic: Past, Present, and Future*, O. M. Johannessen et al., Eds., Springer, 315–387.

Liu, Y., J. Attema, and W. Hazeleger, 2020a: Atmosphere–ocean interactions and their footprint on heat transport variability in the Northern Hemisphere. *J. Climate*, **33**, 3691–3710, https://doi.org/10.1175/JCLI-D-19-0570.1.

——, ——, B. Moat, and W. Hazeleger, 2020b: Synthesis and evaluation of historical meridional heat transport from mid-latitudes towards the Arctic. *Earth Syst. Dyn.*, **11**, 77–96, https://doi.org/10.5194/esd-11-77-2020.

Madec, G., 2008: NEMO reference manual, ocean dynamics component: NEMO-OPA. Notes du Pole de modélisation de l'Institut Pierre-Simon Laplace (IPSL), France, Note 27, https://doi.org/10.5281/zenodo.3878122.

McDermott, P. L., and C. K. Wikle, 2019: Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy*, **21**, 184, https://doi.org/10.3390/e21020184.

Metzger, E. J., and Coauthors, 2014: U.S. Navy operational global ocean and Arctic ice prediction systems. *Oceanography*, **27**, 32–43, https://doi.org/10.5670/oceanog.2014.66.

Mogensen, K., M. A. Balmaseda, and A. Weaver, 2012: The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for System 4. ECMWF Tech. Memo 669, 61 pp., https://www.ecmwf.int/en/elibrary/11174-nemovar-ocean-data-assimilation-system-implemented-ecmwf-ocean-analysis-system-4 .

Mohammadi-Aragh, M., H. Goessling, M. Losch, N. Hutter, and T. Jung, 2018: Predictability of Arctic sea ice on weather time scales. *Sci. Rep.*, **8**, 6514, https://doi.org/10.1038/s41598-018-24660-0.

Onarheim, I. H., T. Eldevik, M. Årthun, R. B. Ingvaldsen, and L. H. Smedsrud, 2015: Skillful prediction of Barents Sea ice cover. *Geophys. Res. Lett.*, **42**, 5364–5371, https://doi.org/10.1002/2015GL064359.

Pacanowski, R. C., K. Dixon, and A. Rosati, 1991: The GFDL modular ocean model users guide. GFDL Ocean Group Tech. Rep. 2, 142 pp.

Peng, G., W. N. Meier, D. Scott, and M. Savoie, 2013: A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring. *Earth Syst. Sci. Data*, **5**, 311–318, https://doi.org/10.5194/essd-5-311-2013.

Perovich, D. K., and J. A. Richter-Menge, 2009: Loss of sea ice in the Arctic. *Ann. Rev. Mar. Sci.*, **1**, 417–441, https://doi.org/10.1146/annurev.marine.010908.163805.

Petrou, Z. I., and Y. Tian, 2019: Prediction of sea ice motion with convolutional long short-term memory networks. *IEEE Trans. Geosci. Remote Sens.*, **57**, 6865–6876, https://doi.org/10.1109/TGRS.2019.2909057.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Reichstein, M., and Coauthors, 2019: Deep learning and process understanding for data-driven earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Rousset, C., and Coauthors, 2015: The Louvain-la-Neuve sea ice model LIM3. 6: Global and regional capabilities. *Geosci. Model Dev.*, **8**, 2991–3005, https://doi.org/10.5194/gmd-8-2991-2015.

Saha, S., and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Salman, A. G., B. Kanigoro, and Y. Heryadi, 2015: Weather forecasting using deep learning techniques. *2015 Int. Conf. on Advanced Computer Science and Information Systems (ICACSIS)*, Depok, Indonesia, IEEE, 281–285.

Shridhar, K., F. Laumann, and M. Liwicki, 2019: A comprehensive guide to Bayesian convolutional neural network with variational inference. https://arxiv.org/abs/1901.02731.

Smith, G. C., F. Roy, and B. Brasnett, 2013: Evaluation of an operational ice–ocean analysis and forecasting system for the Gulf of St Lawrence. *Quart. J. Roy. Meteor. Soc.*, **139**, 419–433, https://doi.org/10.1002/qj.1982.

——, and Coauthors, 2016: Sea ice forecast verification in the Canadian global ice ocean prediction system. *Quart. J. Roy. Meteor. Soc.*, **142**, 659–671, https://doi.org/10.1002/qj.2555.

Stephenson, S. R., and R. Pincus, 2018: Challenges of sea-ice prediction for Arctic marine policy and planning. *J. Borderl. Stud.*, **33**, 255–272, https://doi.org/10.1080/08865655.2017.1294494.

Van Woert, M. L., C.-Z. Zou, W. N. Meier, P. D. Hovey, R. H. Preller, and P. G. Posey, 2004: Forecast verification of the polar ice prediction system (PIPS) sea ice concentration fields. *J. Atmos. Oceanic Technol.*, **21**, 944–957, https://doi.org/10.1175/1520-0426(2004)021<0944:FVOTPI>2.0.CO;2.

Vandal, T., E. Kodra, J. Dy, S. Ganguly, R. Nemani, and A. R. Ganguly, 2018: Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning. *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, New York, NY, ACM, 2377–2386.

Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Walsh, J. E., J. S. Stewart, and F. Fetterer, 2019: Benchmark seasonal prediction skill estimates based on regional indices. *Cryosphere*, **13**, 1073–1088, https://doi.org/10.5194/tc-13-1073-2019.

Wang, H., G. Li, G. Wang, J. Peng, H. Jiang, and Y. Liu, 2017: Deep learning based ensemble approach for probabilistic wind power forecasting. *Appl. Energy*, **188**, 56–70, https://doi.org/10.1016/j.apenergy.2016.11.111.

Wang, L., X. Yuan, and C. Li, 2019: Subseasonal forecast of Arctic sea ice concentration via statistical approaches. *Climate Dyn.*, **52**, 4953–4971, https://doi.org/10.1007/s00382-018-4426-6.

Xingjian, S., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Info. Process. Syst.*, **2015-January**, 802–810.

Yuan, X., D. Chen, C. Li, L. Wang, and W. Wang, 2016: Arctic sea ice seasonal prediction by a linear Markova model. *J. Climate*, **29**, 8151–8173, https://doi.org/10.1175/JCLI-D-15-0858.1.

Zaier, I., C. Shu, T. Ouarda, O. Seidou, and F. Chebana, 2010: Estimation of ice thickness on lakes using artificial neural network ensembles. *J. Hydrol.*, **383**, 330–340, https://doi.org/10.1016/j.jhydrol.2010.01.006.

Zampieri, L., H. F. Goessling, and T. Jung, 2018: Bright prospects for arctic sea ice prediction on subseasonal time scales. *Geophys. Res. Lett.*, **45**, 9731–9738, https://doi.org/10.1029/2018GL079394.