Evaluation of Surface Conditions from Operational Forecasts Using In Situ Saildrone Observations in the Pacific Arctic

CHIDONG ZHANG,[a] AARON F. LEVINE,[b] MUYIN WANG,[a,b] CHELLE GENTEMANN,[c] CALVIN W. MORDY,[a,b] EDWARD D. COKELET,[a] PHILIP A. BROWNE,[d] QIONG YANG,[a,b] NOAH LAWRENCE-SLAVAS,[a] CHRISTIAN MEINIG,[a] GREGORY SMITH,[e] ANDY CHIODI,[a,b] DONGXIAO ZHANG,[a,b] PHYLLIS STABENO,[a] WANQIU WANG,[f] HONG-LI REN,[g] K. ANDREW PETERSON,[e] SILVIO N. FIGUEROA,[h] MICHAEL STEELE,[i] NEIL P. BARTON,[j] ANDREW HUANG,[k] AND HYUN-CHEOL SHIN[l]

[a] *NOAA/Pacific Marine Environmental Laboratory, Seattle, Washington*
[b] *University of Washington, Seattle, Washington*
[c] *Farallon Institute, Petaluma, California*
[d] *European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*
[e] *Environment and Climate Change Canada, Montreal, California*
[f] *NOAA/National Centers for Environmental Prediction, College Park, Maryland*
[g] *Chinese Academy of Meteorological Sciences, China Meteorological Administration, Beijing, China*
[h] *Center for Weather Forecasting and Climate Studies, National Institute for Space Research, São Paulo, Brazil*
[i] *Polar Science Center, Applied Physics Lab, University of Washington, Seattle, Washington*
[j] *Naval Research Laboratory, Monterey, California*
[k] *Science Applications International Corporation, Monterey, California*
[l] *Korea Meteorological Administration, Seoul, South Korea*

(Manuscript received 17 November 2020, in final form 27 February 2022)

ABSTRACT: Observations from uncrewed surface vehicles (saildrones) in the Bering, Chukchi, and Beaufort Seas during June–September 2019 were used to evaluate initial conditions and forecasts with lead times up to 10 days produced by eight operational numerical weather prediction centers. Prediction error behaviors in pressure and wind are found to be different from those in temperature and humidity. For example, errors in surface pressure were small in short-range (<6 days) forecasts, but they grew rapidly with increasing lead time beyond 6 days. Non-weighted multimodel means outperformed all individual models approaching a 10-day forecast lead time. In contrast, errors in surface air temperature and relative humidity could be large in initial conditions and remained large through 10-day forecasts without much growth, and non-weighted multimodel means did not outperform all individual models. These results following the tracks of the mobile platforms are consistent with those at a fixed location. Large errors in initial condition of sea surface temperature (SST) resulted in part from the unusual Arctic surface warming in 2019 not captured by data assimilation systems used for model initialization. These errors in SST led to large initial and prediction errors in surface air temperature. Our results suggest that improving predictions of surface conditions over the Arctic Ocean requires enhanced in situ observations and better data assimilation capability for more accurate initial conditions as well as better model physics. Numerical predictions of Arctic atmospheric conditions may continue to suffer from large errors if they do not fully capture the large SST anomalies related to Arctic warming.

KEYWORDS: Arctic; Atmosphere-ocean interaction; Forecast verification/skill

1. Introduction

Environmental predictions of the Arctic face new challenges as sea ice diminishes faster than anticipated (Yadav et al. 2020), exposing a vast area of the ocean that was previously covered by sea ice to the atmosphere in summer. This has resulted in unprecedented ocean surface warming (Steele et al. 2008), which led to impacts on the marine and terrestrial ecosystems (Lewis et al. 2020; Bhatt et al. 2017) and changes in air–sea gas exchange (DeGrandpre et al. 2020). In the long term, whether and when the summer Arctic will be ice-free are critical climate issues with tremendous global impacts (Wang and Overland 2009; Vihma 2014). In the shorter term, distributions of Arctic sea ice in the coming week, month, season, and year are vital information for coastal communities, management and conservation of marine resources, operation of shipping, fishing and energy industries in the Arctic, and weather forecasting of midlatitudes. Knowledge of interannual fluctuations in sea ice coverage is also needed to understand how consequential variability in surface energy fluxes may affect the Arctic atmosphere–ocean–sea ice system (Danielson et al. 2020).

Arctic prediction faces several challenges (Jung et al. 2016), rooted mainly in two sources. One is the lack of in situ observations for initial and boundary conditions and forecast validation (Smith et al. 2019). In the Arctic, it is extremely difficult to take operational in situ observations because of annual shifts in sea ice coverage. Increasing amounts of observations have been taken by research field experiments in the Arctic. They are,

Yang's current affiliation: The Climate Corporation, Seattle, Washington.

Corresponding author: Chidong Zhang, chidong.zhang@noaa.gov

DOI: 10.1175/MWR-D-20-0379.1

however, still spatially limited, and are rarely made available in a timely fashion to prediction centers. While satellite observing technology has rapidly progressed, many critical variables for prediction (e.g., sea ice thickness, conditions at the air–sea interface, upper ocean properties) are not reliably retrieved from spaceborne remote sensing with desirable accuracy or validated by an adequate suite of in situ observations (Castro et al. 2016; Sallila et al. 2019; Sedlar and Tjernstrom 2019; Naakka et al. 2019; Banzon et al. 2020). The second source of the challenges faced by Arctic prediction is inadequate model representations of key physical processes for the Arctic energy budget, notably ice physics (e.g., Hunke et al. 2011), ice–ocean–atmosphere interaction (e.g., Boutin et al. 2020), and cloud properties (e.g., Sedlar et al. 2020), which substantially affect solar and IR radiation. This second source is closely related to the first one because improving prediction models inevitably depends on new knowledge on the key processes that must be gained from observations.

Predictions of Arctic large-scale circulation patterns have been verified by comparing against the 500-hPa geopotential height from model (re)analysis products (Bauer et al. 2016; Jung and Matsueda 2016). For verification of Arctic predictions at the surface, the usage of global reanalysis products as surrogates to observations must be practiced with extreme caution owing to a lack of in situ observations to constrain data assimilation. Large biases exist in reanalysis products when compared to limited in situ observations (Francis 2002; Liu and Key 2016). Prediction errors in the 2-m air temperature compared with station observations are larger than compared with a reanalysis product (Bauer et al. 2016). Large discrepancies exist between reanalysis products and in situ observations at the surface and in the boundary layer in the Arctic (Beesley et al. 2000; Lüpkes et al. 2010; Lindsay et al. 2014). For 2-m temperature, different reanalysis products can differ by 10 K in their mean errors against observations at synoptic stations over land in the Arctic (Hersbach et al. 2020).

Key issues of advancing Arctic prediction systems include expanding in situ observations and wisely using existing observations, however sparse, to verify predictions (Jung et al. 2016). The World Meteorological Organization (WMO) Polar Prediction Project (Gordon et al. 2014), especially its key component Year of Polar Prediction (Goessling et al. 2016), mobilized international efforts to develop new capabilities of Arctic prediction, including observations, model development, prediction verification, and user engagement. In the Arctic, the impact of in situ observations on prediction is comparable to that of satellite data, even though the amount of in situ observations available to prediction systems is no more than 10% of satellite observations (Lawrence et al. 2019). This implies that the impact of in situ observations is much greater than satellite observations per data sample. Most in situ observations that have been used in model verification are from weather stations over land (Atkinson and Gajewski 2002; Bauer et al. 2016; Lawrence et al. 2019; Køltzow et al. 2019; Naakka et al. 2019) with fewer obtained from ships (Naakka et al. 2019; Lüpkes et al. 2010; Tjernström et al. 2021) and drifting ice camps (Lindsay 1998).

New opportunities of observations from uncrewed platforms have emerged to fill in the blind spots of conventional observing networks. A case in point is the Arctic Ocean where observations are extremely difficult to make because of the seasonal migration and erratic behavior of sea ice. Recently, a new type of uncrewed surface vehicles (USVs), saildrones, has been developed and deployed in the Arctic Ocean that can be used to validate numerical weather prediction (NWP) products. However, using USV observations to validate NWP products has never been done in the Arctic. Thus, its feasibility for this purpose needs to be tested and evaluated.

This article describes a study that uses in situ observations from saildrones to validate NWP products near the ocean surface in the Arctic for June–September 2019. Our first objective is to test the feasibility of using observations from saildrones to validate NWP products and to develop a suitable methodology that can be broadly applied to observations from other types of USVs. Two major difficulties are introduced by the mobility of observing platforms when their observations are used to validate NWP products. Even though comparisons between observations from a moving platform and gridded NWP products can be made at the same location at a given forecast time through interpolation, the model error growth between two forecast times does not carry the same meaning as at a fixed location because of the spatial variability over a distance covered by the moving platform during this time. We will discuss the extent to which this may affect the measured prediction error growth. In addition, observed standard deviations can be used as a benchmark to compare model errors in different variables. Standard deviations calculated using observations from a moving platform would, however, include variability in both time and space. We will discuss whether such observed standard deviations are still useful to comparisons of errors in different variables.

Our second objective is to interpret results from comparisons of saildrone observations and NWP products. Pertinent issues include the representativeness of the results from a limited time (summer 2019) and geographic domain (the Bering, Chukchi, and Beaufort Seas) of the observations, different behaviors among surface pressure, wind, temperature and humidity, connections between errors in surface air temperature and sea surface temperature, and the possible roles of uncertainties associated with initial conditions in forecast errors.

In this first attempt at validating NWP products using USV observations, we use a subset of multiyear saildrone deployments as well as deterministic model forecasts. Recent studies have suggested that the intrinsic limit of atmospheric predictability in the Arctic is approximately two to three weeks (Judt 2020). Realizing that no current model can reach even close to this predictability limit with its deterministic forecast, we confined our forecast validation to a lead time of 10 days. Once we ascertain the feasibility of using USV observations to validate NWP products, we plan to expand this study to include observations from saildrone deployments in other years and in other regions, and to use ensemble forecasts to cover the full range of potential atmospheric predictability.

The saildrone observations are introduced in section 2 along with descriptions of the NWP products compared against the observations and the methodology adopted. Results from such comparisons are presented in section 3, and their possible interpretations are provided in section 4. A summary and concluding remarks are given in section 5.

2. Observations, NWP products, and method

Saildrones are remotely piloted USVs, powered by wind and solar energy. They can be equipped with numerous sensors to measure physical, chemical, and biological variables (Cokelet et al. 2015; Meinig et al. 2015, 2019). They have been deployed in different climate zones of the world oceans, such as the Bering, Chukchi and Beaufort Seas, the Southern Ocean, the tropical Pacific and Atlantic, and the west coast of North America. They have served a variety of purposes, including surveys of fish and marine mammals (Mordy et al. 2017; Chu et al. 2019; De Robertis et al. 2019; Kuhn et al. 2020), measurement of surface fluxes of energy (Zhang et al. 2019) and carbon dioxide (Sutton et al. 2021), and evaluation of satellite data (Vazquez-Cuervo et al. 2019; Gentemann et al. 2020; Scott et al. 2020).

The observations used in this study are from a joint project conducted by NOAA Pacific Marine Environment Laboratory (PMEL) and the NASA Physical Oceanography program's Multisensor Improved Sea Surface Temperatures (MISST) team. Four saildrones were launched from Unalaska, Alaska, on 15 May. Their observations covered the Bering Sea, Chukchi Sea, and Beaufort Seas (Fig. 1). Three saildrones (sd-1034, 1036 and 1037) returned to Unalaska, Alaska, on 13 October. One saildrone (sd-1035) was recovered at Utqiaġvik, Alaska, on 4 October. The total number of days of observations was 147 (with days of no data toward the end), and the total distance traveled by all saildrones was over 65 000 km. The most northern point of the observations was 75.4°N (August 17). In this study, the saildrone observations from 1 June to 30 September were used.

The objectives of the saildrone deployment were to (i) explore the feasibility of using saildrone observations to evaluate NWP products, (ii) improve the calibration of satellite-based measurements of SST in polar waters, (iii) observe along four Distributed Biological Observatory lines (Grebmeier et al. 2019), (iv) perform cross-calibration tests for sensors of surface CO_2 partial pressure against instrumentation aboard the USCGC *Healy*, (v) survey ocean currents of Hanna Shoal, the Chukchi Shelf Current and the Alaskan Coastal Current (Li et al. 2019), and (vi) estimate surface energy fluxes of the Arctic Ocean and compare them with those based on other observations. This study covers only the first objective.

In this study, we used saildrone observations of surface barometric pressure (p) measured at 0.2 m above the sea surface (Vaisala PTB210 barometers), air temperature (T) and relative humidity (RH) at 2.3 m (Rototronic HC2-S3 probes), wind speed and direction at 5.2 m (Gill 1590-PK-020 anemometers), and SST at −0.5 m (Sea-Bird SBE37 MicroCAT). The sampling frequencies are 10 Hz for wind and 1 Hz for the other variables. The data used to compare with model predictions are hourly, averaged over six 1-min means at the start of every 10-min interval within the hour. Camera images were also available in the horizontal directions and downward from the mast. Detailed information on the sensors and data quality can be found in Meinig et al. (2015), Cokelet et al. (2015), Zhang et al. (2019), and Gentemann et al. (2020).

The accuracy of the saildrone measurement has been assessed by comparing observations of a saildrone and a moored buoy in the tropical eastern tropical Pacific within 12 km from each other for 315 h (Zhang et al. 2019). The buoy observations were treated as a standard. The root-mean-square (RMS) and mean differences between the two are listed in Table 1 (second column from the left).

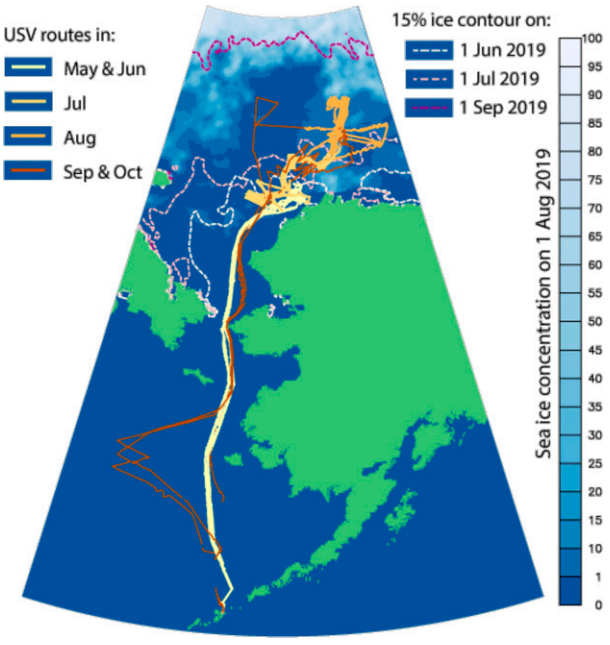

FIG. 1. Tracks of saildrones (solid lines) and contours of 15% sea ice concentration edges (dotted lines) for 1 Jun (white), 1 Jul (pink), and 1 Sep (magenta). Colors along the saildrone tracks indicate months. Background color is sea ice concentration on 1 Aug. Ice information is based on the AMSR-2 sea ice concentration product of the EUMETSAT Ocean and Sea Ice Satellite Application Facility (OSI SAF, www.osi-saf.org).

During the 2019 deployment, cross comparisons were made between the saildrones when they were sailing side by side (within 7 km) for seven hours at the beginning of the deployment on 17 May and for eight hours at the end on 6 October. The mean 95% confidence limits that resulted from these comparisons are also listed in Table 1 (third column from the left). There is no obvious reason to suspect that the measurement accuracy of the saildrones in the Arctic should be different from that in the tropics, except when water freezes on the sensors, which happened only a few times during the deployment. Data during these times were excluded from this study. The confidence limits derived from the Arctic comparison can be considered upper bounds of measurement uncertainties if they are larger than those based on the saildrone–buoy comparison. Both are much smaller than errors in the predictions as demonstrated in section 3.

Observations made by the saildrones during the 2019 Arctic deployment can be categorized into three scenarios based on camera images: open water without detectable sea ice (no sea ice in any image), which comprises the majority (96%) of the total observation samples; open water with sea ice detected at distances (sea ice in sideward images but not downward), about 2% of the total samples; and saildrone in contact with

TABLE 1. Measurement uncertainties.

Variable	RMS (mean) difference between a saildrone and buoy in the eastern tropical Pacific	95% confidence limit from Arctic cross comparisons, beginning and end of mission	Manufacturer's accuracy; stability
p (hPa)	—	±0.21 and ±0.36	±0.30; 0.10 yr^{-1}
RH (%)	2.3 (−1.2)	±1.34 and ±4.17	±0.8 at 23°C; 1.0 yr^{-1}
T (°C)	0.31 (−0.02)	±0.145 and ±0.089	±0.1 at 23°C; 0.1 yr^{-1}
Wind speed (m s^{-1})[a]	0.63 (−0.28)	±0.576 and ±0.544	±0.18 up to 45 m s^{-1}
Wind direction (°)[a]	16.0 (−3.9)	±6.11 and ±6.51	±2 up to 45 m s^{-1}
SST (°C)	0.047 (0.011)	±0.051 and ±0.022	±0.002; 0.0024 yr

[a] Stability information is not available from instrument manuals.

ice (ice in downward images), about 2% of the total samples. Most of the observations near sea ice were in June and August (Chiodi et al. 2021). The main results from this study are not affected by the observed ice scenarios when the full record of the data is used.

The NOAA daily Optimum Interpolation Sea Surface Temperature version 2.1, hereafter briefly as OISST v2.1 (Banzon et al. 2020), was used to provide climatology and anomalies in SST over the saildrone operation area and period.

Forecasts used in this study are from eight prediction centers (Table 2). Some of the forecasts are archived by The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE, Swinbank et al. 2016); others are provided directly by the model centers. The grid spacing of all data used in this study

TABLE 2. Prediction models.

Model ID	Model institute (country), model name	Grid spacing and No. of vertical levels of the atmospheric model	Atmosphere–ocean coupling	Assimilated saildrone observations	Reference(s)
CMA	Chinese Meteorological Administration (China), Global and Regional Assimilation and Prediction System	50 km, 31 levels	No	None	Shen et al. (2020)
CPTEC	Center for Weather Forecasting and Climate Studies (Brazil), Brazilian Atmospheric Model	104 km, 28 levels	No	None	Figueroa et al. (2016)
ECCC	Environment and Climate Change Canada (Canada)	39 km, 40 levels	Yes	SST, p, T	CMC (2019)
ECMWF	European Centre for Medium-Range Weather Forecasts (International), Integrated Forecasting System	16 km, 91 levels	Yes	SST,[a] p	ECMWF (2018); ECMWF (2019)
JMA	Japanese Meteorological Administration (Japan), Global Spectral Model	20 km, 60 levels	No	None	Yonehara et al. (2018)
KMA	Korean Meteorological Administration (Korea), United Model	32 km, 70 levels	No	None	Schellekens at al. (2011)
Navy-ESPC	Naval Research Laboratory, Monterey (United States), Global Prediction Systems–Earth System Prediction Capability	37 km, 60 levels	Yes	SST, p, T	Barton et al. (2020)
NCEP	National Centers for Environmental Prediction (United States), Global Forecast System, version 15	12 km before and 55 km after day 10, 28 levels	No	None	Yang and Tallapragada (2018)[b]

[a] Using the OSTIA analysis (Good et al. 2020).
[b] Also see https://www.emc.ncep.noaa.gov/emc/pages/numerical_forecast_systems/gfs.php.

is 0.5° × 0.5°, although the original grid spacing of each model is different (third column from the left in Table 2). Their initialization (i.e., 0 lead time) and forecasts at lead times of 1–10 days were used in this study. The models are initialized twice per day at 0000 and 1200 UTC every day, except for Navy-ESPC, which is initialized once per day at 1200 UTC, four days per week (Saturday, Sunday, Monday, and Tuesday). These discrepancies do not discernibly affect the results to be shown in section 3.

ECCC, ECMWF, and Navy-ESPC are atmosphere–ice–ocean coupled models; the others are uncoupled atmospheric-only models. CMA, CPTEC, and NCEP used SST from the NCEP analysis as their initial and lower boundary conditions. ECCC, JMA, and Navy-ESPC used SST from their own respective daily ocean analysis systems that assimilated observations from satellites, ships, and buoys. ECMWF and KMA used SST from the Operational Sea Surface Temperature and Sea Ice Analysis (Donlon et al. 2012). In uncoupled atmosphere-only model forecasts, SST fields persisted through the course of prediction. Even in uncoupled forecasts, it is useful to assess errors in SST fields as they may contribute to errors in other surface variables (e.g., temperature, humidity, winds). Only the three coupled models (ECCC, ECMWF, and Navy-ESPC) include dynamic sea ice. How sea ice prediction may affect errors in other predicted variables is not assessed in this study. Eight models are far fewer than commonly used sizes of operational ensemble forecasts (Leutbecher 2019). Model-averaged errors over the eight models were nevertheless calculated.

The saildrone observations applicable to numerical weather prediction were made available to prediction centers in real time via the Global Telecommunication System (GTS). Certain variables (e.g., p, T, SST) were assimilated into some of the forecast systems (5th column in Table 2). However, it is difficult to assess the impact the saildrone observations may have had on initial conditions of the models that included them in their data assimilation. Operational data assimilation receives observations from many sources (in situ and remote) and these observations are weighted differently when they are assimilated. Some of the received observations may be rejected for various reasons. Initial conditions may deviate substantially from observations even when observations are included in operational data assimilation. There are several methods that can quantitatively demonstrate the impact of the observations on forecast (e.g., Baker and Daley 2000; Errico 2007; Cardinali 2009; Zhu and Gelaro 2008; Lorenc and Marriot 2014), which are beyond the scope of this study. We, however, pay attention to discernible differences between forecasts by models that assimilated saildrone observations and those that did not.

To compare the saildrone observations with the model output, a 1-h average of the saildrone observations centered on a given synoptic hour (e.g., 1200 UTC) was made. Model output from that synoptic hour was linearly interpolated to a saildrone location at the same synoptic hour from the four model grid points closest to that location. The hourly average of the saildrone observations was then compared with the interpolated model forecast at the saildrone location. In this sense, at a given lead time, comparisons between saildrone observations and NWP products are at the same (saildrone) location at that time. This linear interpolation works reasonably well (section 3). We compare the six variables observed by the saildrones (p measured at 0.2 m, T and RH at 2 m, u and v at 5 m, and SST at −0.5 m) with the model output of sea level p, 2-m T, 2-m dewpoint, 10-m u and v, and sea surface skin temperature, respectively.

Model dewpoint or specific humidity was converted to RH using the Clausius–Claperon relationship. The model 10-m winds were converted to 5 m where saildrone wind observations are, following the neutral stability approach (Mears et al. 2001) with the roughness length of 1.52×10^{-4} m (Peixoto and Oort 1992). Preliminary calculations of skin SST based on the observations from one saildrone using the COARE flux algorithm (Fairall et al. 2003) indicate that the cool skin and warm layer rarely differ from observed SST at −0.5 m by more than 0.2°C. As will be shown (section 3), this is much smaller than errors in SST, predicted or prescribed in most models. Such conversion from −0.5-m temperature to skin temperature requires simultaneous radiation measurements, which are available from only two of the four saildrones. For consistency, we decided not to make this conversion. This does not compromise the main results from this study.

Once the model outputs were interpolated onto the locations of the four saildrones, their differences from the observations (models − observations) at each lead time (including day 0) were considered errors. At a given synoptic hour and lead time, errors were calculated for each model–saildrone pair. Further averages of 0000 and 1200 UTC were then made. To illustrate the error growth, daily errors were summarized into root-mean-square errors (RMSE) at each lead time, for the entire deployment period. For each forecast model, its RSME at a lead time τ was calculated as the following:

$$\mathrm{RMSE}(\tau) = \sqrt{\frac{\sum_{n}^{N}\sum_{s_n}^{S}\left[P(s_n,\tau) - O(s_n)\right]^2}{NS}}, \qquad (1)$$

where n is an index for the saildrones ($N = 4$), s_n is an index for the data point of each saildrone ($S = 120$), P represents the forecast, and O represents the saildrone observations. Each data point corresponds to a specific location along a saildrone track at a given time (day). Missing data in observations and model output are less than 2% of the total.

Mean standard deviations are used to measure relative amplitudes of errors in different variables. Errors of different variables cannot be directly compared to each other because of their different natural variability. One option is to normalize the errors using their respective standard deviations. But it is desirable to know the total values of the errors. The compromise would be to mark observed standard deviations on figures showing errors in section 3. For this purpose, a proxy of the mean standard deviation for a given variable at a given lead time was calculated using the saildrone observations. For instance, at the lead time of 5 days, observed standard deviations within a running window of 5 days through the entire deployment period was calculated for each saildrone and then averaged over the four saildrones. This mean standard deviation includes averaged variability in time within 5 days and

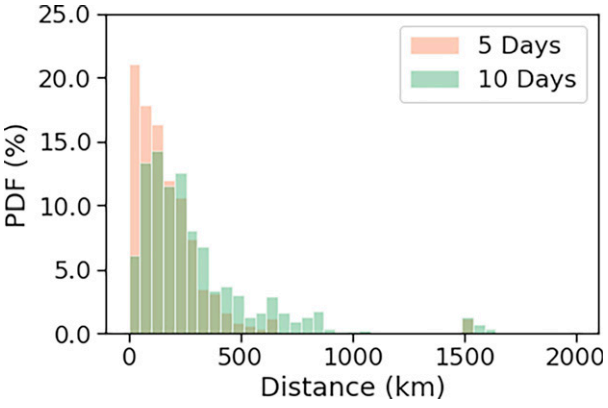

FIG. 2. Probability distributions of saildrone travel distance within 5 (orange) and 10 (light green) days. Dark green bars mark overlays of the two.

spatial variability over distances traveled by saildrones within 5 days. It was calculated the same way for all variables and used as a benchmark to compare forecast errors of different variables. This approach, however, was not applied to initial times. Observed standard deviations at initial times were estimated by extrapolation from those at lead times of 1–3 days.

In addition to the general issue of point-to-box comparisons that applies to all numerical model verification against in situ observations at fixed locations, there is an additional issue when observations are from moving platforms. In this study, predictions are always compared with observations at the same location along each of the saildrone tracks. Because of the saildrones' mobility, such validation locations vary with forecast lead time. This may introduce extra errors in time (see the appendix), especially when saildrones move from locations in a model grid box to those in another. Distances traveled by the saildrones within 5 and 10 days are summarized in Fig. 2. Even within 5 days, the saildrones could travel over 50 km, the size of the model grid boxes. The severity of this issue can be assessed by evaluating the dependence of prediction errors on the distances traveled by the saildrones over different lead times. We find such dependence of prediction errors on the saildrone travel distance insignificant (section 3). We concluded that the possible overestimate of error growth in time due to the mobility of the saildrones is negligible.

3. Prediction validation

In this section, we present the results from comparing model forecasts against saildrone observations. We also discuss issues in calculations of forecast errors related to the mobility of saildrones as mentioned in the previous section. We defer possible interpretations of the error behaviors found in this section to section 4.

Time series of the observed and predicted p, T, and SST are given in Fig. 3 as examples of two contrasting prediction error behaviors. Each curve in the figure is an average over four saildrone locations at a given time. Model initial conditions for p match the observations very well with relatively small intermodel spread (Fig. 3a). Differences between the observations and predictions increase but stay smaller than synoptic and intraseasonal variations up to approximately day 7 (Figs. 3c). By day 10, the various models appear to have very little or no deterministic predictive skill (Fig. 3d). In contrast, for T,

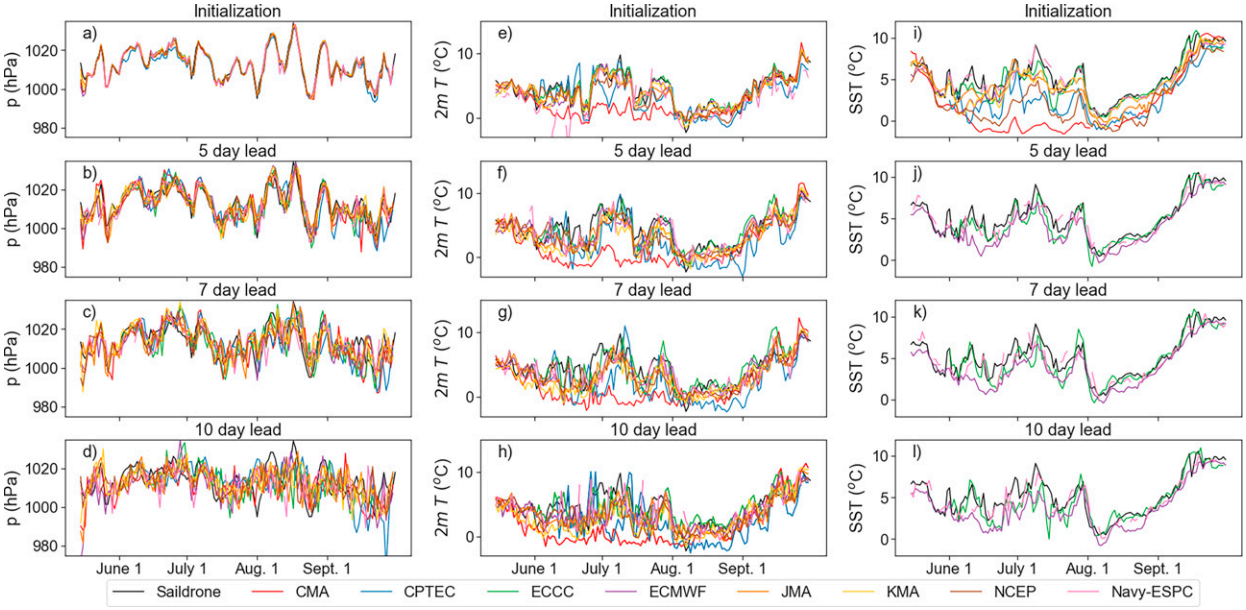

FIG. 3. Examples of time series in (a)–(d) p, (e)–(h) T, and (i)–(l) SST from saildrone observations (black curves) and models (colored curves) for lead times of 0 (initial conditions), 5, 7, and 10 days, shown from top to bottom, all averaged over the four saildrone tracks at a given time. For SST, all model initial conditions are included in (i), while only predictions by the coupled models are included at the forecast lead times in (j)–(l).

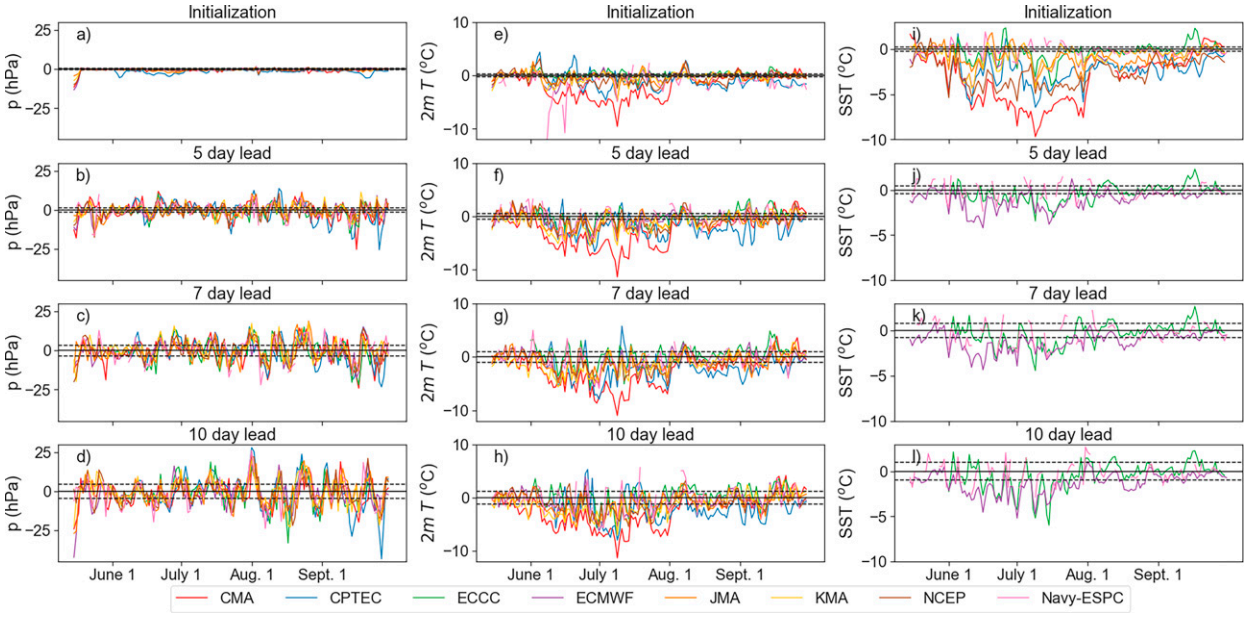

FIG. 4. Examples of time series of prediction errors (prediction − observation) in (a)–(d) p, (e)–(h) T, and (i)–(l) SST for lead times of 0 (initial conditions), 5, 7, and 10 days, shown from top to bottom, all averaged over the four saildrone tracks at a given time. For SST, all model initial conditions are included in (i), while only predictions by the coupled models are included at the forecast lead times in (j)–(l). Horizontal dotted lines indicate one standard deviation of the saildrone observations (see section 2 for information on their calculations).

model initial conditions show a considerable intermodel spread, with deviations from the observations for some models very large during June and July and smaller the rest of the period (Fig. 3e). These differences between the observations and forecasts remain at all forecast lead times (Figs. 3f–h). Interestingly, errors in SST share similarities to those in T, with large differences between the observations and model initial conditions in June and July (Fig. 3i). Systems (CMA, CPTEC) with the largest errors in initial SST also show the largest T prediction errors at the lead time of 10 days (Fig. 3h). Coupled models (ECCC, ECMWF, Navy ESPC) show relatively large forecast errors in SST (Figs. 3j–l) in the same months of large forecast errors in T, suggesting these forecast errors are related. Note that the coupled models are among the systems with the smallest T error at day 10 (Fig. 3h).

Prediction errors (model minus observation) perceived from Fig. 3 are quantified in Fig. 4 where their amplitudes (averaged over four saildrone locations at a given time) are compared against observed standard deviations (see section 2 for their calculations). As suggested by Fig. 3, errors in p grow in all months (Fig. 4, left column), whereas errors in T and SST remain obviously larger in June and July than the rest of the period from the initial time to day 10 (Fig. 4, center and right columns). On synoptic time scales, forecast errors in p from all models appear to fluctuate together; this is not observed for errors in T. RMS differences (RMSD) in forecast errors (normalized by observed standard deviations within a 10-day running window so the results for p and T can be compared, see section 2) between ECMWF as a reference and each of the other models were calculated at the lead time of 10 days over the entire period. The RMSD ranges from 1.14 to 1.60 for p and from 0.85 to 2.16 for T. This illustrates that intermodel differences in forecast errors are smaller for p than T. Errors in SST at the initial time and forecast lead times are mostly negative (Figs. 4i–l), meaning models underestimated SST. Possible reasons for this are discussed in section 4d.

The general growth of prediction errors for all variables included in this study is demonstrated in terms of root-mean-squared errors (RSMEs) for the entire deployment period (Fig. 5) without the information of their possible dependence on the months during the period but also without averaging errors over the four saildrones [Eq. (1) in section 2]. To make errors in different variables comparable to each other, standard deviations calculated using the saildrone observations over the spans of the forecast lead times (see section 2 for details) are used as benchmarks to measure the relative amplitudes of the errors. Several points can be made from Fig. 5. First, there is discernible gain in prediction skill through multimodel means (gray line with triangles) for p, u, and v close to 10-day lead time but not for T and RH. Second, the amplitudes of prediction errors in p are indistinguishable from its observed standard deviations during the early time of the forecast (<6 days) but grow rapidly (amplitude increasing by more than three to five times within 10 days) and become obviously greater than its observed standard deviations during the later time of the forecast for all models (Fig. 5a). The situation is similar although to a lesser degree for u and v (Figs. 5c and 5e). In contrast, errors in T are greater than its observed standard deviations from the initial time to the end of the forecast period; models (CMA, CPTEC) with larger initial errors in T suffer from larger prediction errors all the time (Fig. 5b). Errors in RH are consistently large from the initial time to the end of the forecast

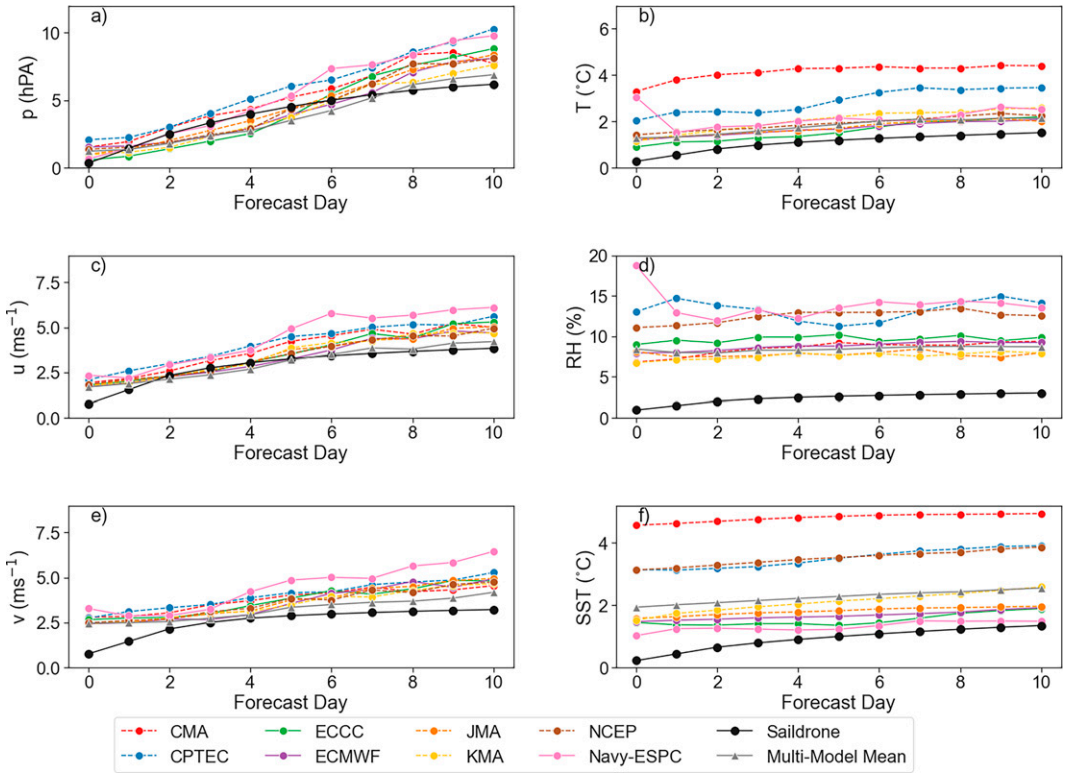

FIG. 5. Prediction errors (RMSEs) as functions of forecast lead time (day) for (a) p, (b) T, (c) u, (d) RH, (e) v, and (f) SST. Standard deviations calculated using the saildrone observations (see section 2) are denoted by black curves. Solid color lines with dots are for the coupled models.

period for all models (Fig. 5d). Errors in T and RH grow much more slowly (less than doubling within 10 days) than p, u, and v. The connections between initial and forecast errors shown in Fig. 6 illustrate that models suffering from larger forecast errors have larger initial errors for T and RH (the right column), to a much lesser degree for u and v, and not for p at all (left column). Third, two models (CMA, CPTEC) have the largest errors in both T and SST, and five models (ECCC, ECMWF, JMA, KMA, Navy-ESPC) have the smallest errors in both T and SST (Figs. 5b and 5f). An exception is NCEP, which has relatively small errors in T but large errors in SST. Initial errors are larger than prediction errors in T and RH from Navy-ESPC (Fig. 5b and d) are likely a result of initialization shock (Mulholland et al. 2015) in these forecasts as they were initialized from a separate ocean and atmosphere analysis. This is prone to occur when initial conditions at the sea surface were prepared through data assimilation that is not fully coupled. The precise reasons for this large initial error in Navy-ESPC diagnosed here are currently under investigation.

Before we attempt to further explain the results shown in Figs. 3–6, we need to address the issues related to the moving platform as mentioned in the previous sections: Does linear interpolation from model grid points to a saildrone track work as well as to a stationary observational site? How much would spatial variability along a saildrone track contaminate the analysis of prediction error growth? To address the first issue, we examine the spatial variability in forecast errors of p and T along the saildrone tracks [$\delta_s E(\delta s, t_1)$ in the appendix]. This is measured by differences between forecast errors at pairs of saildrone locations as a function of their distance, normalized by their standard deviations for direct comparisons between the variables. There is no discernable increase of forecast errors with the distance (not shown). This suggests that the mobility of the observing platform (saildrones) did not contaminate the forecast error calculations in any obvious way. We thus are confident that the distance traveled by saildrones over any two given forecast times within 10 days does not contribute to the behaviors of prediction errors (the appendix).

We evaluated additional diagnostics to verify that the results from our forecast validation against observations from moving platforms are consistent with those against observations at fixed locations. For this we selected a surface weather station (ENHE, 65.5°N, 2.3°E, elevation 1 m) on an offshore oil production platform in the Heidrun Oil Field off the west coast of Norway, an oceanic environment similar to that for the saildrones deployment. Being on the Atlantic side of the Arctic, this station provides additional information of possible geographic dependence of our results, which will be discussed in section 4. Validation of the forecasts against observations from this station (without SST) over the same period of the saildrone deployment led to results very similar to those against the saildrone observations. For example, errors in p

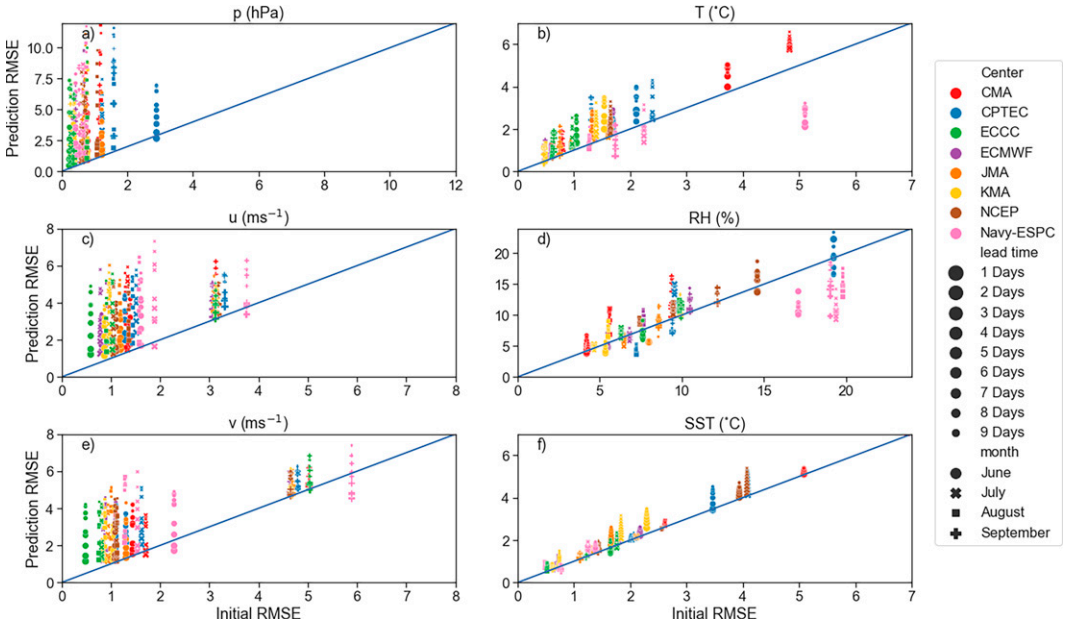

FIG. 6. Scatter diagrams of monthly RMSE [using Eq. (1) in section 2 for each month, e.g., $S = 30$] at lead times of 1–10 days vs initial times for (a) p, (b) T, (c) u, (d) RH, (e) v, and (f) SST. Legends for models, lead times and months are given at the right. Blue lines mark the 1:1 relationship.

are indistinguishable from its observed standard deviations (calculated using the station observations the same way as using the saildrone observations , see section 2) during the short-range forecast (<6 days) but become distinguishably greater than the observed standard deviations at the longer lead time (cf. Figs. 5a and 7a). Errors in T and RH at the station can be consistently higher than their observed standard deviations from the initial time to the end of the forecast (Figs. 7b and 7d), as seen along the saildrone tracks (Figs. 5b and 5d), although they are smaller and less spread among the models than along the saildrone tracks. Among all variables, errors in RH do not grow much at the station (Fig. 7d) as seen along the saildrone tracks (Fig. 5d).

Yamagami et al. (2019) validated forecasts of Arctic cyclones against a global reanalysis product using sea level pressure. They found that forecast errors in central pressure persistently increase with lead time, as seen in our results (Fig. 5a). The error growth in our results is roughly from 2 hPa in a 1-day forecast to 7–10 hPa in a 10-day forecast, which is within the error growth roughly from 3 hPa in a 1-day forecast to 12–15 hPa in a 6-day forecast found by Yamagami et al. (2019). Tjernström et al. (2021), using observations from an icebreaker between Sweden and the North Pole, also found relatively small initial errors in p and wind speed followed by their obvious error growth in forecasts in contrast to relatively large initial errors in T and surface specific humidity without obvious error growth in forecast. These similarities between the error behaviors at the fixed locations (the weather station and model grids) and along tracks of moving platforms (ship, saildrones) lend us confidence that observations from saildrones can be used for validation of NWP products as observations from fixed locations.

4. Discussion

Results presented in the previous section raise several questions that are addressed in this section.

a. Are the results presented in section 3 specific to the limited geographic domain (the Bering, Chukchi, and Beaufort Seas) and time (summer of 2019) of the saildrone deployment?

The similarities in the error behaviors found by using the saildrone observations on the Pacific side of the Arctic and based on the weather station observations on the Atlantic side (Figs. 5 and 7) suggest that our results are applicable to a broad region of the Arctic Ocean. This conclusion is further supported by results from Køltzow et al. (2019) who used surface observations from land-based weather stations on the Atlantic side of the Arctic (surrounding the Norwegian Sea) in winter to validate forecasts by four numerical models. They found that errors in initial conditions are relatively smaller for p than T, whereas error growth is greater for p than T, similar to what is shown in Fig. 7 for boreal summer. These similarities between the results from Køltzow et al. (2019), Tjernström et al. (2021), and ours suggest that these error behaviors are independent of the season and locations within the Arctic. This is, however, not the case for regions outside the Arctic. To demonstrate this, we took a different approach. We compared ECMWF forecasts of p and SST (similar to T) against its own analysis as proxies of observations in five different regions: Northern Hemisphere (NH), Southern Hemisphere (SH), the tropics (TR), Mediterranean, and the Bering, Chukchi, and Beaufort Seas (B/C/B Seas). The strong regional dependence of the error growth is obvious in Fig. 8. Error growth is slowest

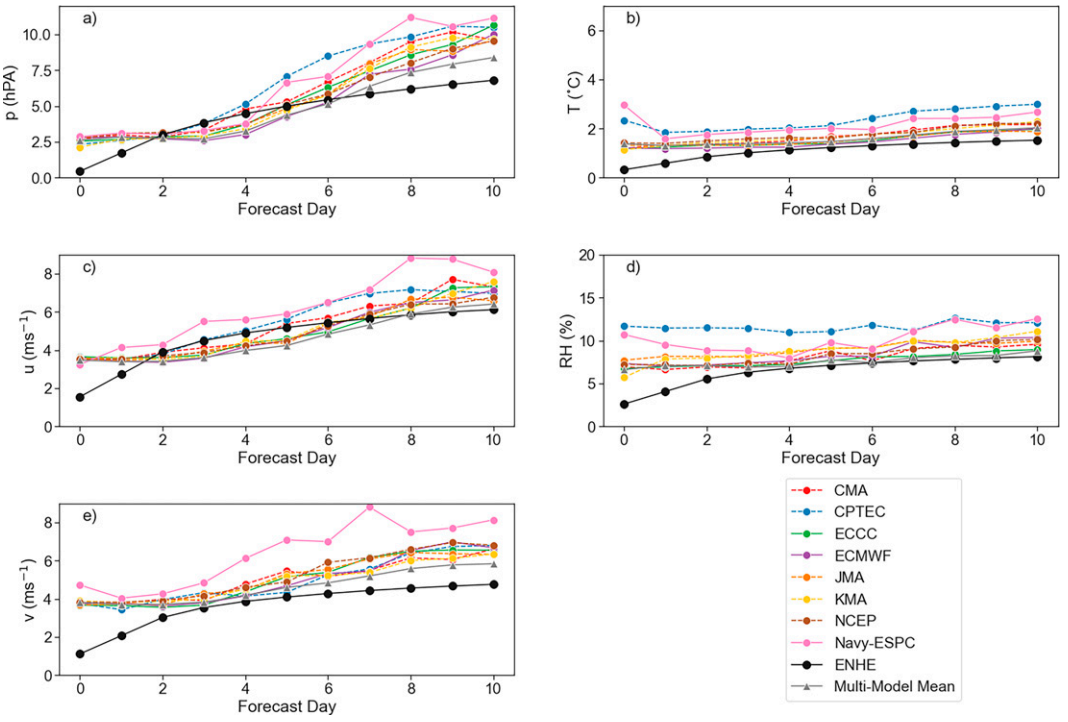

FIG. 7. As in Fig. 5, but using observations from a weather station (ENHE; 65.5°N, 2.3°E) without SST.

in the tropics (orange lines) and relatively rapid in the Arctic (red lines). This may not be surprising given that the atmosphere is more predictable in the tropics than in the Arctic (Judt 2020).

b. *Why do the errors behave differently for different variables?*

The errors behave differently in several ways. (i) Errors of p and wind are relatively small in the short-range (<6 days) forecasts but become much greater in the longer forecasts (Figs. 5 and 7). For T and RH, errors can be large at the initial time and remain so throughout the forecast period (Figs. 5 and 7; Køltzow et al. 2019). (ii) Models with larger initial errors in T and RH suffer from larger forecast errors, which is not the case for p and wind (Fig. 6). (iii) Multimodel means improve the prediction skill for p, u, and v but not T and RH (Fig. 5).

Forecast errors depend mainly on three factors: errors in initial conditions, model deficiencies, and the predictability limit of the natural system. Arctic in situ observations available in real time for NWP centers to prepare their forecast initial conditions are extremely sparse. It is thus not surprising to see large initial errors in T and RH. The relatively small initial errors in p and wind suggest that their initial conditions depend less on the local conditions because the spatial scale of their variability is large or they are constrained over a large scale both horizontally and vertically. This is indeed the case, as measured by differences between simultaneous observations from different saildrones as a function of their distance when normalized by their standard deviation through the deployment period (Fig. 9). The smaller scatter of p and larger scatter of T as functions of distances between saildrones suggest that the spatial scales (or correlation length scales) of p are much larger than those of T. Over the ocean, T is more influenced by SST than large-scale dynamics, whereas p is a product of mass in the atmospheric column, which is closely related to tropospheric temperature. Satellite observations of radiance or retrieved tropospheric temperature are assimilated at most NWP centers, which helps with the accuracy of p in initial conditions. Even though surface winds (u, v) cannot always be well observed by satellites because of cloudiness, they are geostrophically constrained by p and surface drag, albeit with the complications of momentum entrainment across the top of the atmospheric boundary layer (Stevens et al. 2002). An accurate initial condition in p would help with the accuracy of initial conditions in winds over the ocean at high latitudes. Near-surface air temperature is difficult to derive from satellite radiances due to the sensitivity of low-peaking channels to other factors, such as land/sea surface temperature and temperatures higher in the air column. Many operational data assimilation systems are not able to use the full information content from such channels, and coupled data assimilation is being developed to improve their usage (Frolov et al. 2020). As a case in point, global gridded surface energy flux products may use satellite retrievals for surface wind and humidity, but not for surface air temperature (Bentamy et al. 2003; Yu and Weller 2007; Tomita et al. 2019).

The very similar amplitudes and timing of errors in p across all models (e.g., at 7- and 10-day forecasts in Fig. 4), not seen for T and RH, are intriguing. It suggests that all models missed the evolution of p in a similar way. An example is shown in Fig. 10, using the evolution of ECMWF analysis as proxies to observations (red) and deterministic forecasts

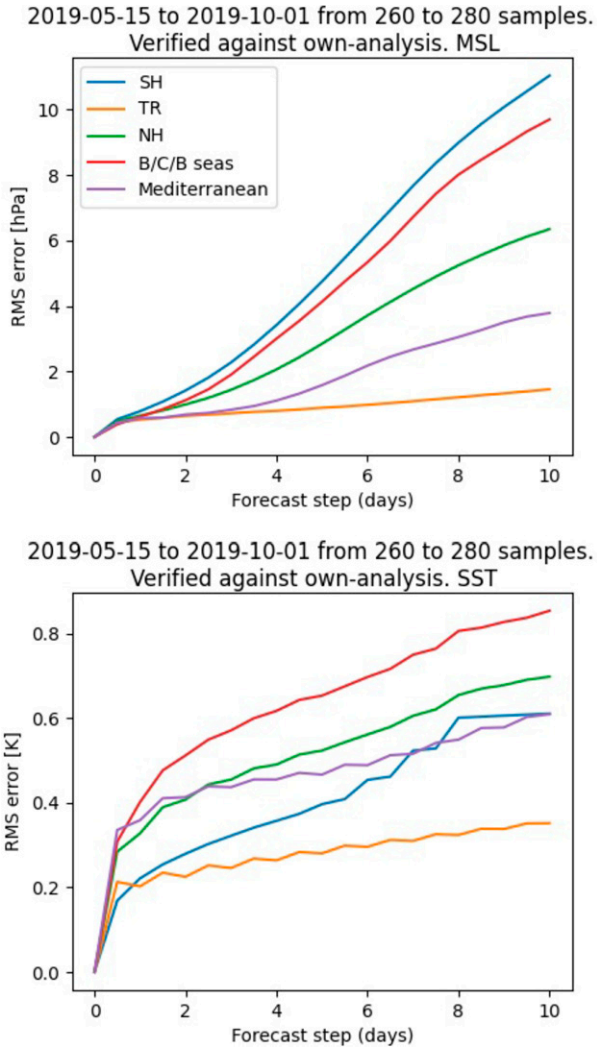

FIG. 8. RMSEs of (top) p and (bottom) SST in the Northern Hemisphere (NH, green, north of 20°N), Southern Hemisphere (SH, blue, south of 20°S), the tropics (TR, orange, 20°S–20°N), Mediterranean (purple, 31°–46°N, 6°W–36°E), and the Bering, Chukchi, and Beaufort Seas (B/C/B Seas, red, 50°–80°N, 170°E–150°W) calculated using ECMWF forecasts and analysis.

initialized at 0000 UTC 8 August 2019 (blue). On 9 and 11 August (Figs. 10b and 10c), the low pressure center developed offshore Kamchatka Peninsula was well captured by the 1–4-day forecasts. But the 7-day forecast misplaced a new low center on 15 August (Fig. 10h). This new low center stayed northeast of Kamchatka Peninsula through 17 August (Fig. 10j) and disappeared on 18 August, whereas the forecasted low center moved northeastward and maintained itself through 18 August (Fig. 10k). The misplacement of the low centers in the forecast (Figs. 10h–j) makes forecast errors in p inversely proportional to the observed p: large negative errors are found where a low center exists in the forecast but not in observations, and large positive errors are found where there is a low center in observations but not in forecasts (Fig. 11a). A comprehensive forecast skill assessment for p should also include other metrics in addition to pointwise comparison, such as spatial coherence, which is, however, unfeasible to accomplish using observations from moving platforms. The forecast models may suffer from common deficiencies that prevent their prediction skills from approaching a longer predictability limit.

c. *Are T, SST, and their errors related?*

While surface air temperature T is closely related to SST in both observations and most forecasts (Fig. 12) as a result of air–sea interactions in coupled model or SST boundary conditions in uncoupled models, there are discernable differences in this T–SST connection between the observations and forecasts. In the observations, T is lower than SST more often than otherwise, especially for low SST (Fig. 12i). This is likely a result of a particular wind regime sampled during the saildrone cruises. In the Chukchi Sea, the saildrones were following the ice retreat to the north and winds in July–September are normally toward the southwest off the ice. This cold air over the warmer sea surface was reproduced but exaggerated by some models (e.g., Fig. 12h). Other models were more likely to produce warmer than colder air over the sea surface at low SST (Figs. 12a,b,g).

The near air–sea equilibrium (e.g., T and SST vary in tandem) would lead to a possible connection between errors in T and SST. This has been indicated in Figs. 5b and 5f. Prediction errors in T and SST are connected to different degrees for different models (Fig. 13). For example, such a connection is the strongest for CMA, an uncoupled model (Fig. 13a), and weakest for CPTEC, an uncoupled model (Fig. 13b), and Navy-ESPC, a coupled model (Fig. 13g). How errors in SST and T are related may be affected by parameterization schemes for the atmospheric boundary layer as well as air–sea coupling in the models.

d. *What role do initial conditions play in prediction errors?*

The role of initial conditions in forecast errors in T and RH is best illustrated in Figs. 5 and 6. In this particular case, the amplitude of forecast errors in T and RH are directly related to the magnitude of their initial errors. Models with relatively large initial errors in T and RH have large errors throughout the entire 10-day forecast period (CMA and CPTEC for both T and RH, and NCEP and Navy-ESPC for RH). For these models, more accurate initial conditions may help reduce forecast errors in T and RH, at least at the short-range lead times. This is much less so for forecast errors in p, u, and v.

The quality of initial conditions depends on three factors: observations, data assimilation algorithms, and forecast models. Observations may suffer from their sparsity and errors in measurement. Data assimilation can be compromised by assumptions made in its algorithm. Even with highly accurate observations and excellent data assimilation algorithms, the accuracy of initial conditions can only be as good as what can be made by a forecast model used in the data assimilation. It would be interesting to find out how the three factors contribute to the differences between initial errors in p, u, v and in T, RH. Two uncoupled models (JMA and KMA) suffer

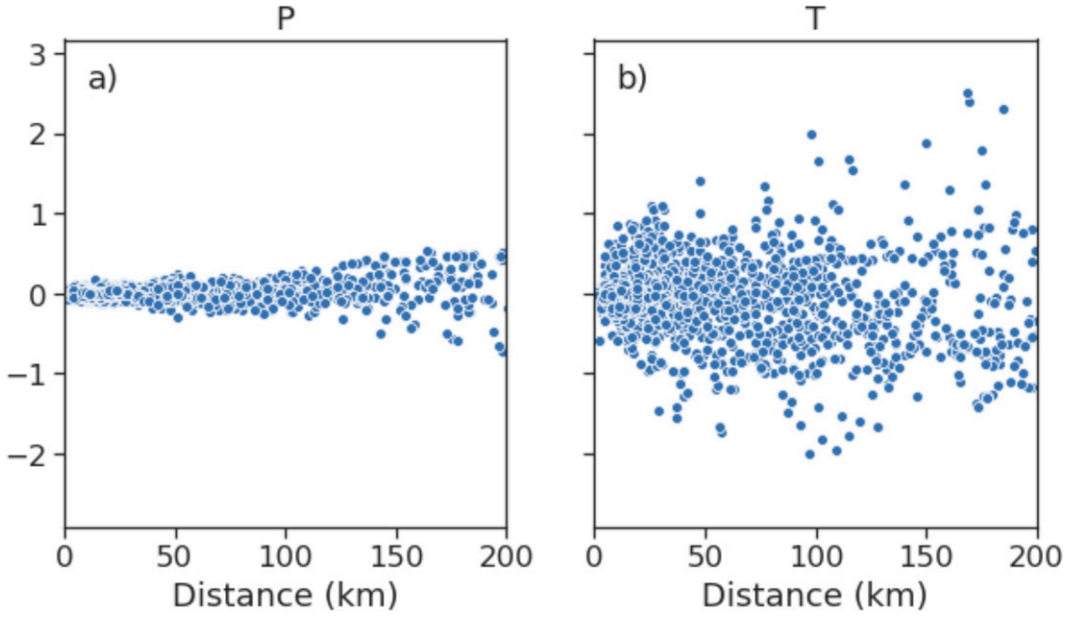

FIG. 9. Scatter diagram of differences between simultaneous observations from each pair of saildrones and their distances for (a) p and (b) T, normalized by their respective standard deviations observed through the entire deployment.

relatively small initial and prediction errors in T and RH as the coupled models (Figs. 5b and 5d). These two uncoupled models are initialized using their own data assimilation systems (section 2). The other two coupled models (CMA, CPTEC) that use initial conditions from an analysis prepared by a different model (NCEP) both suffer relatively large initial and prediction errors in T, RH (CPTEC only), and SST (Figs. 5b,d,f). The NCEP model is an interesting hybrid case. It is uncoupled and suffers from relatively large initial and prediction errors in SST and RH (Figs. 5f and 5d), but not in T (Fig. 5b).

Satellite retrievals of SST are available, even though in the Arctic they suffer from serious issues (e.g., Castro et al. 2016; Banzon et al. 2020), whereas satellite retrieval of T is not. This and the apparent connection between errors in T and SST in initial conditions and in prediction (Fig. 13) for most models point to the importance of initial conditions in SST (which serve as lower boundary conditions for uncoupled models). The evolution of errors in SST along the saildrone tracks is very similar to that in T (Fig. 4, right column): For certain models, they are large in late June through July and smaller in the rest of the deployment period from the initial time to forecast lead time of 10 days. This is so as errors of coupled model predictions (Fig. 4) and uncoupled model lower boundary conditions (not shown). This suggests that initial and forecast errors cannot be completely attributed to model biases.

The large initial errors in SST during June and July may be related to the unusual warming in 2019 in the Bering, Chukchi, and Beaufort Seas (SST anomalies ≥ 5°C) that was not captured by model initialization. This led to larger SST errors, and hence larger T errors, over warmer water (Figs. 11b and 11c). In June and July, the saildrone tracks ran through areas of very large positive SST anomalies (>4°C), whereas in August and September, they were more spread and some of them were in areas of relatively small SST anomalies (Fig. 14a, thin red lines). The SST anomalies averaged over the saildrone tracks (Fig. 14a, thick red line) were 4.42°, 5.55°, 2.64°, and 3.55°C for June, July, August, and September, respectively. Models that did not fully capture the unusual surface warming could suffer large initial errors in June and July along the saildrone tracks.

Given the limited observations of SST in the Arctic, some models manage to prepare their SST initial conditions much better than others (Figs. 3i, 5f, and 11c). This speaks to the importance of data assimilation. Data assimilation matters in several ways: whether data assimilation is fully or partially coupled or not coupled at all, and whether data assimilation is done using the same model for prediction. Another critical issue is how many observations are made available in real time to operational data assimilation. To illustrate this, we compare saildrone observations of SST to the OISST v2.1 product (Banzon et al. 2020) along the saildrone tracks. In situ observations of SST, including those from our saildrones in 2019, were used to correct satellite retrievals in OISST v2.1. This made the OISST v2.1 SST match the saildrone observations (blue lines in Fig. 14b) very well in general, with their differences mostly less than 2°C (teal lines in Fig. 14b). The monthly RMSE along individual saildrone tracks range from 0.60° to 1.64°C, with track averages of 1.00°, 0,96°, 0.80°, and 0.37°C for June, July, August, and September, respectively. These differences are much smaller than those between the saildrone observations and initial conditions of the models, which are 1°–4.6°C (Fig. 5f).

Because all in situ observations used in OISST v2.1 were not available in real time through GTS, OISST v2.1 is not a real-time product that can be used by NWP centers to prepare

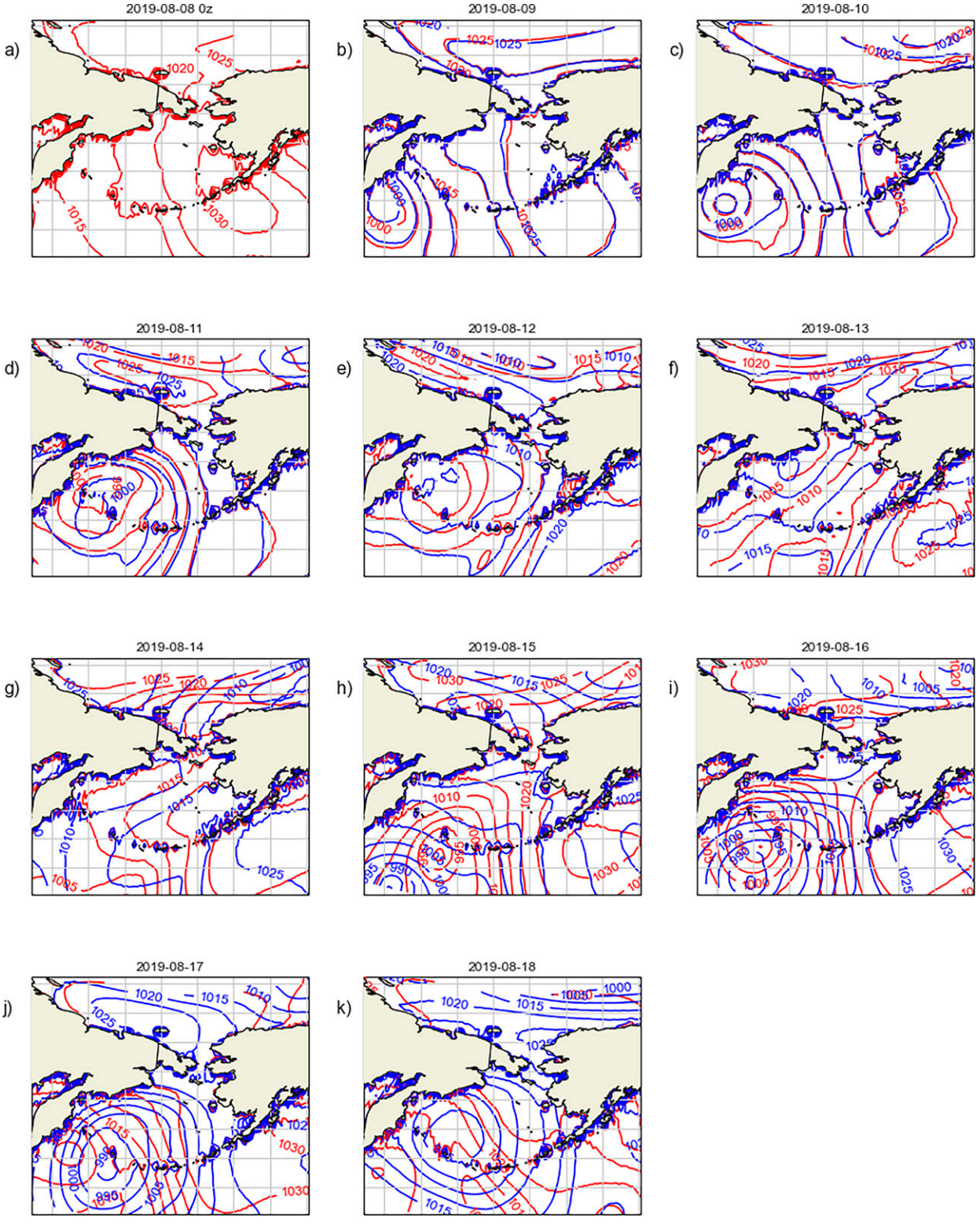

FIG. 10. Evolution of p from the ECMWF analysis as a proxy of observations (red contours) at 0000 UTC each day from 8 to 18 Aug 2019 and its forecast by ECMWF initialized at 0000 UTC 8 Aug (blue contours).

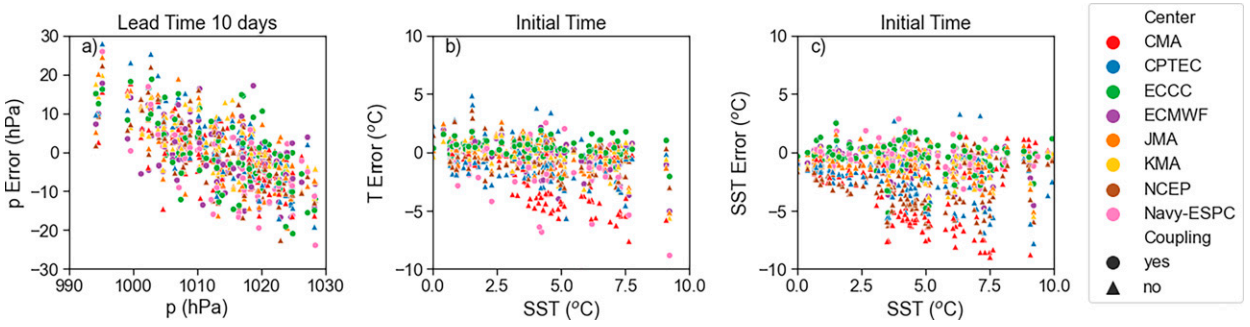

FIG. 11. Scatter diagram of (a) prediction errors in p vs observed p at the lead time of 10 days, (b) initial errors in T vs observed SST, and (c) initial errors in SST vs observed SST. Triangles mark the uncoupled models, dots mark the coupled models.

their initial conditions. Mean errors in SST in model initial conditions and OISST v2.1 were calculated along each saildrone track, then averaged over all tracks, and averaged over the coupled and uncoupled models, respectively. They are apparently the largest for the uncoupled models (Fig. 11c; red line in Fig. 14c) and relatively smaller for the coupled models (purple) and OISST v2.1 (green). For the entire saildrone deployment period, the mean errors (RMSE) are 2.08°C (1.96°C) for the uncoupled models, 0.43°C (1.3°C) for the coupled models, and 0.33°C (0.83°C) for OISST v2.1. If OISST v2.1 was a real-time product and available to NWP, then it is possible that errors in SST would be substantially reduced in the initial conditions. This could possibly further reduce prediction errors in SST of the coupled models and in T of all models.

5. Summary and conclusions

This study has used in situ observations from saildrones to evaluate numerical weather prediction (NWP) products in the Arctic during June–September 2019. While NWP validations in the Arctic have been done previously (Francis 2002; Atkinson and Gajewski 2002; Bauer et al. 2016; Jung and Matsueda 2016; Lawrence et al. 2019; Køltzow et al. 2019), ours focused on the air–sea interface in the Bering, Chukchi, and Beaufort Seas, which, to the best of our knowledge, had rarely been done before mainly because few, if any, observations of all key surface variables (T, RH, p, u, v, SST) were available there. This is also the first attempt of using observations from remotely controlled uncrewed surface vehicles (USVs) to evaluate NWP products in the Arctic. This study discussed complications of using observations from mobile platforms in NWP validation, and demonstrated that such complications do not hinder the utility of observations from USVs in such validation, at least up to a forecast time of 10 days. Results from this study demonstrate that in a data-sparse or data-void region, limited observations from USVs can lead to useful information for NWP that is otherwise unavailable.

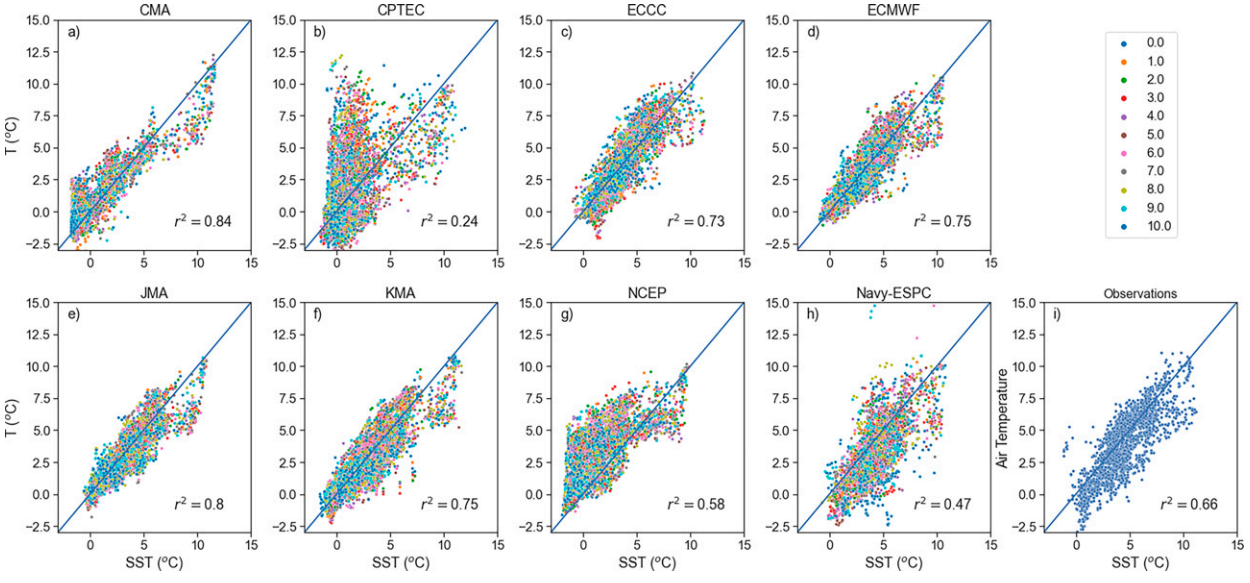

FIG. 12. Scatter diagram of T vs SST from the models and observations and their correlations squared (r^2). For model output, the colors denote forecast lead times.

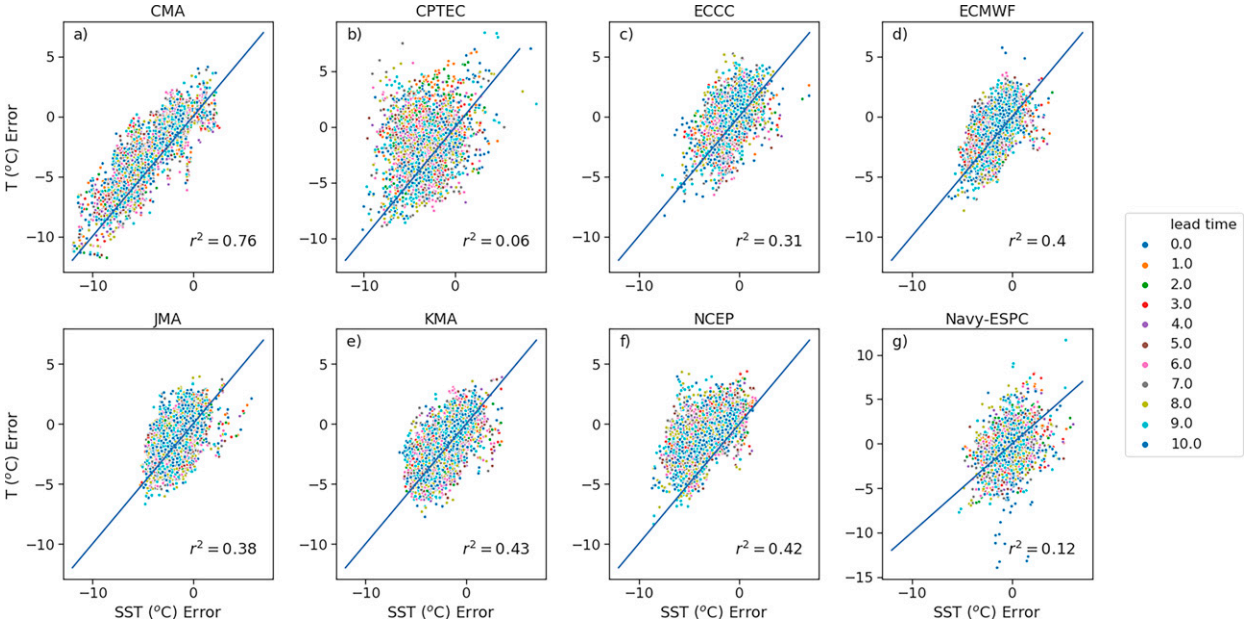

FIG. 13. Scatter diagrams of model errors in SST and T. Colors denote forecast lead times.

The main results from this study and their implications are as follows:

1) Prediction errors behave differently for p, u, v and T, RH when measured against their respective observed standard deviations. For p, u, v, errors are small (indistinguishable from their observed standard deviations) at initial and short prediction lead times (<6 days) but they grow and become much larger than their observed standard deviations at longer lead times. All models diagnosed in this study perform similarly in this regard. Non-weighted multimodel means outperform all individual models, especially at longer prediction lead time (>6 days). In contrast, for T and RH, sizes of forecast errors are determined by initial errors that are larger than their observed standard deviations. Non-weighted multimodel means do not outperform all individual models. There are several possible interpretations for the different error behaviors. They include dominant large-scale variability and constraints for p, u, v; small-scale variability of T and RH; and differences in their initialization.
2) The prediction error behaviors found in the Bering, Chukchi, and Beaufort Seas in summer can also be found in other parts of the Arctic in other seasons (Køltzow et al. 2019; Tjernström et al. 2021), but they are in sharp contrast to error behaviors in regions outside the Arctic, especially the tropics. It is known that the predictability limit varies with region (Judt 2020). Physical processes pertaining to the performance of model parameterization schemes may also depend on region.
3) The relatively larger errors in T and RH at the initial time and their persistence through 10-day forecasts strongly suggest that improving their initial conditions may reduce their prediction errors. The close connection between T and SST and between their errors at the initial time indicates that improving initial conditions of SST would tremendously benefit prediction of T and perhaps also RH even for short-range forecasts. Factors that would contribute to improved initial conditions include more in situ observations, their real-time availability to prediction centers, and better data assimilation capabilities. Satellite observations are and will continue to be the main sources of observations of SST over the Arctic Ocean. More in situ observations of SST (Castro et al. 2016; Banzon et al. 2020) are needed to help improve satellite retrievals of SST and further benefit NWP initial conditions.

This study can be expanded in several ways, given the confidence we have gained from it that saildrone observations are useful in the validation of NWP products. Saildrone observations in the Arctic Seas in the past and future can be used to confirm the results from this study. The potential benefits of an expanded campaign with a greater number of saildrones or other types of USVs covering a larger area can be beneficial. Comparisons between USV observations and NWP products can be extended beyond 10 days using ensemble forecasts. Similar comparisons can be made between USV observations and global reanalysis products. USV observations from other remote oceanic regions with sparse or no in situ observations at the sea surface can also be used in NWP validation.

In addition to improving initial conditions and improving model parameterizations, the extent to which predictability might be different for p, u, v and T, RH is yet to be determined. It is also necessary to determine how much room there is for improvement of Arctic prediction within the predictability limit of the dynamic system (Lorenz 1982; Krishnamurthy 2019; Shen et al. 2021). Based on the recent study of Judt (2020), the

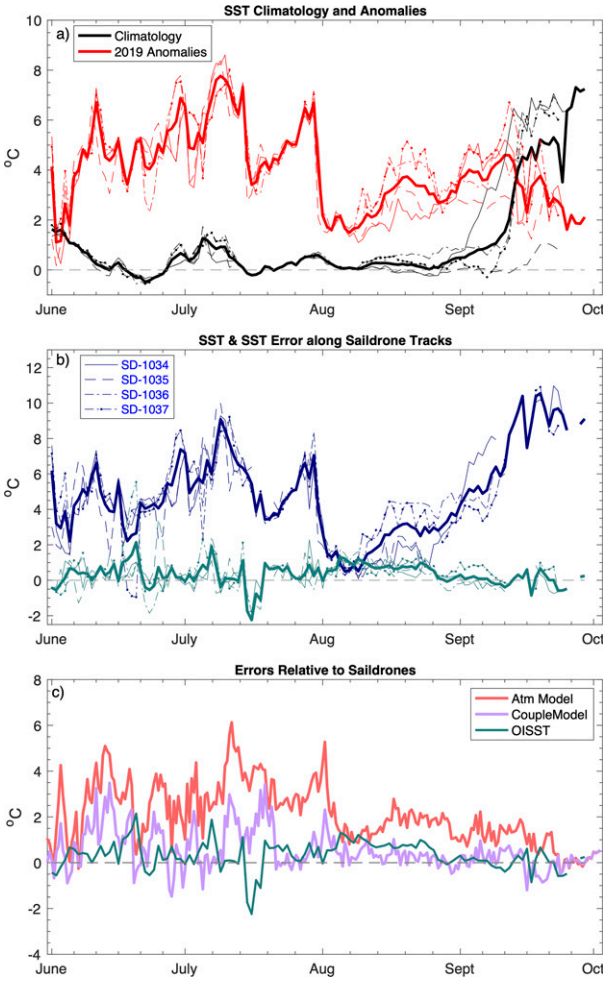

FIG. 14. (a) Daily SST climatology (1982–2010) (black) and anomalies (red) along each saildrone track (thin lines) and their average (thick) based on OISST v2.1. (b) Time series of SST observed by each saildrone (blue) and difference between OISST v2.1 and saildrone observations (teal) along each saildrone track (thin) and their averages (thick). (c) Deviations of OISST v2.1 (green), mean initial SST error for coupled models (purple curve), and uncoupled models (red) from saildrone observations averaged over the four tracks at a given time.

predictability of the Arctic is different from other regions of the world. Predictability near the surface might be different from that aloft (Judt 2018).

The Arctic warming in 2019 took place under unusual conditions. North of 60°N for October 2018–August 2019 the annual land surface air temperature was the second warmest since 1900 (Overland et al. 2019). The early onset of melting in the spring of 2019 led to one of the lowest sea ice extents on record in the Bering and Chukchi Seas, accompanied by the second highest SSTs on record in August 2019 (Meier et al. 2019; Timmermans and Ladd 2019). These unusual conditions may make Arctic predictions more difficult than before. Given the recent accelerated warming in the Arctic, numerical predictions of the Arctic may continue to suffer from large errors if their initial conditions do not capture the large SST anomalies. Additional in situ observations have been proven beneficial to Arctic prediction (Inoue 2020). This calls for more in situ observations in the Arctic to be made available in real time and more efforts to fully incorporate in situ observations in preparation of initial conditions of numerical predictions.

Acknowledgments. The authors thank Heather Tabisola of PMEL/University of Washington for her assistance to the saildrone mission and Alexandra Darden of National Ice Center for providing ice information during the saildrone deployment. Collaborative information exchanges with Craig M. Lee and Luc Rainville of the Applied Physical Laboratory, University of Washington, Falko Judt of the National Center for Atmospheric Research, and Masahiro Kazumori of Japanese Meteorological Administration are also appreciated. Comments from Daniel Hodyss and four anonymous reviewers helped to substantially improve the manuscript. The effort of TIGGE to archive the forecasts used in this study is acknowledged. This work has been supported by NOAA OMAO and PMEL, and by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA15OAR4320063. MS and CG were funded by the MISST3 Project, NASA Grant 80NSSC20K0768. MS was also funded by NASA Grants NNX16AK43G and 80NSSC20K0134, and NSF Grants OPP-1751363 and PLR 1603266. NPB was sponsored by OPNAV N2N6E and the Office of Naval Research's Earth System Prediction Capability, PE0603207N. This is PMEL Contribution Number 5122, JISAO Contribution Number 2020-1105, and EcoFOCI Contribution Number EcoFOCI-0954.

Data availability statement. Saildrone data are available at https://ferret.pmel.noaa.gov/pmel/erddap/search/index.html?page=1&itemsPerPage=1000&searchFor=±2019±Arctic±Saildrone. Model output is from TIGGE at https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=cf/ for CMA, ECMWF, KMA, JMA, NCEP, from http://ftp.cptec.inpe.br/pesquisa/bamc/saildrone_Arctic2019/ for CPTEC, and from https://www.pmel.noaa.gov/cwi/saildrone_Arctic2019 for ECCC. The Navy-ESPC output is freely available for public use. Please contact Dr. Neil Barton (neil.barton@nrlmry.navy.mil) for data dissemination information. OISST is from https://www.ncdc.noaa.gov/oisst/optimum-interpolation-sea-surface-temperature-oisst-v21. The observations from station ENHE are from University of Wyoming (http://weather.uwyo.edu/upperair/sounding.html)

APPENDIX

Using Observations from Mobile Platforms to Measure Prediction Error Increment in Time

When observations used to validate NWP products are from a mobile platform, biases can be introduced to the measured prediction error increment over time because of the spatial variability over a distance traveled by the platform during that time. This is explained in the following.

Let the forecast error be

$$E(s, t) = P(s, t) - O(s, t), \tag{A1}$$

where s represents location, t is time, P is the forecast, and O is the observations. The error increment at a fixed location s_2 over a time period $\delta t = t_2 - t_1$ is

$$\delta_t E(s_2, \delta t) = E(s_2, t_2) - E(s_2, t_1). \tag{A2}$$

For a mobile platform, from time t_1 to t_2, its location would change from s_1 to s_2. So the error increment from time t_1 to t_2 measured at location s_2 is

$$\begin{aligned}\delta E(s_2, \delta t) &= E(s_2, t_2) - E(s_1, t_1)\\ &= [E(s_2, t_2) - E(s_2, t_1)] + [E(s_2, t_1) - E(s_1, t_1)]\\ &= \delta_t E(s_2, \delta t) + \delta_s E(\delta s, t_1),\end{aligned} \tag{A3}$$

where $\delta_t E(s_2, \delta t)$ is the intended measure of the error increment in time [Eq. (A2)], and $\delta_s E(\delta s, t_1) = E(s_2, t_1) - E(s_1, t_1)$ measures the spatial variability at t_1 over the distance $\delta s = s_2 - s_1$ traveled by the mobile platform during δt. The practically measured error increment $\delta E(s_2, \delta t)$ can be taken as a close approximation of the intended error increment in time $\delta_t E(s_2, \delta t)$ only if $\delta_s E(\delta s, t_1)$ is negligibly small.

REFERENCES

Atkinson, D. E., and K. Gajewski, 2002: High-resolution estimation of summer surface air temperature in the Canadian Arctic Archipelago. *J. Climate*, **15**, 3601–3614, https://doi.org/10.1175/1520-0442(2002)015<3601:HREOSS>2.0.CO;2.

Baker, N. L., and R. Daley, 2000: Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Quart. J. Roy. Meteor. Soc.*, **126**, 1431–1454, https://doi.org/10.1002/qj.49712656511.

Banzon, V., T. M. Smith, M. Steele, B. Huang, and H.-M. Zhang, 2020: Improved estimation of proxy sea surface temperature in the Arctic. *J. Atmos. Oceanic Technol.*, **37**, 341–349, https://doi.org/10.1175/JTECH-D-19-0177.1.

Barton, N., and Coauthors, 2020: The Navy's Earth System Prediction Capability: A new global coupled atmosphere-ocean-sea ice prediction system designed for daily to subseasonal forecasting. *Earth Space Sci.*, **7**, e2020EA001199, https://doi.org/10.1029/2020EA001199.

Bauer, P., L. Magnusson, J. N. Thépaut, and T. M. Hamill, 2016: Aspects of ECMWF model performance in polar areas. *Quart. J. Roy. Meteor. Soc.*, **142**, 583–596, https://doi.org/10.1002/qj.2449.

Beesley, J. A., C. S. Bretherton, C. Jakob, E. L. Andreas, J. M. Intrieri, and T. A. Uttal, 2000: A comparison of cloud and boundary layer variables in the ECMWF forecast model with observations at Surface Heat Budget of the Arctic Ocean (SHEBA) ice camp. *J. Geophys. Res.*, **105**, 12 337–12 349, https://doi.org/10.1029/2000JD900079.

Bentamy, A., K. B. Katsaros, A. M. Mestas-Nuñez, W. M. Drennan, E. B. Forde, and H. Roquet, 2003: Satellite estimates of wind speed and latent heat flux over the global oceans. *J. Climate*, **16**, 637–656, https://doi.org/10.1175/1520-0442(2003)016<0637:SEOWSA>2.0.CO;2.

Bhatt, U. S., and Coauthors, 2017: Changing seasonality of panarctic tundra vegetation in relationship to climatic variables. *Environ. Res. Lett.*, **12**, 055003, https://doi.org/10.1088/1748-9326/aa6b0b.

Boutin, G., C. Lique, F. Ardhuin, C. Rousset, C. Talandier, M. Accensi, and F. Girard-Ardhuin, 2020: Towards a coupled model to investigate wave–sea ice interactions in the Arctic marginal ice zone. *Cryosphere*, **14**, 709–735, https://doi.org/10.5194/tc-14-709-2020.

Cardinali, C., 2009: Monitoring the observation impact on the short-range forecast. *Quart. J. Roy. Meteor. Soc.*, **135**, 239–250, https://doi.org/10.1002/qj.366.

Castro, S. L., G. A. Wick, and M. Steele, 2016: Validation of satellite sea surface temperature analyses in the Beaufort Sea using UpTempO buoys. *Remote Sens. Environ.*, **187**, 458–475, https://doi.org/10.1016/j.rse.2016.10.035.

Chiodi, A. M., and Coauthors, 2021: Exploring the Pacific Arctic seasonal ice zone with saildrone USVs. *Front. Mar. Sci.*, **8**, 640690, https://doi.org/10.3389/fmars.2021.640697.

Chu, D., S. Parker-Stetter, L. C. Hufnagle, R. Thomas, J. Getsiv-Clemons, S. Gauthier, and C. Stanley, 2019: 2018 Unmanned Surface Vehicle (Saildrone) acoustic survey off the west coasts of the United States and Canada. *OCEANS 2019 MTS/IEEE SEATTLE*, Seattle, WA, Institute of Electrical and Electronics Engineers, https://doi.org/10.23919/OCEANS40490.2019.8962778.

CMC, 2019: Global Ensemble Prediction System (GEPS) update from version 5.0.0 to version 6.0.0. Environment and Climate Change Canada, Tech Doc., 72 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_geps-600_20190703_e.pdf.

Cokelet, E. D., C. Meinig, N. Lawrence-Slavas, P. J. Stabeno, C. W. Mordy, H. M. Tabisola, R. Jenkins, and J. N. Cross, 2015: The use of Saildrones to examine spring conditions in the Bering Sea. *OCEANS 2015-MTS/IEEE Washington*, Washington, DC, Institute of Electrical and Electronics Engineers, https://doi.org/10.23919/OCEANS.2015.7404357.

Danielson, S. L., and Coauthors, 2020: Manifestation and consequences of warming and altered heat fluxes over the Bering and Chukchi Sea continental shelves. *Deep-Sea Res. II*, **177**, 104781, https://doi.org/10.1016/j.dsr2.2020.104781.

De Robertis, A., and Coauthors, 2019: Long-term measurements of fish backscatter from Saildrone unmanned surface vehicles and comparison with observations from a noise-reduced research vessel. *ICES J. Mar. Sci.*, **76**, 2459–2470, https://doi.org/10.1093/icesjms/fsz124.

DeGrandpre, M., W. Evans, M. L. Timmermans, R. Krisheld, B. Williams, and M. Steele, 2020: Changes in the Arctic Ocean carbon cycle with diminishing ice cover. *Geophys. Res. Lett.*, **47**, e2020GL088051, https://doi.org/10.1029/2020GL088051.

Donlon, C. J., M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Wimmer, 2012: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.*, **116**, 140–158, https://doi.org/10.1016/j.rse.2010.10.017.

ECMWF, 2018: IFS documentation CY45R1—Part I: Observations. ECMWF, https://www.ecmwf.int/node/18711.

——, 2019: IFS documentation CY46R1—Part I: Observations. ECMWF, https://www.ecmwf.int/node/19305.

Errico, R. M., 2007: Interpretations of an adjoint-derived observational impact measure. *Tellus*, **59A**, 273–276, https://doi.org/10.1111/j.1600-0870.2006.00217.x.

Fairall, C. W., E. F. Bradley, J. E. Hare, A. A. Grachev, and J. B. Edson, 2003: Bulk parameterization of air–sea

fluxes: Updates and verification for the COARE algorithm. *J. Climate*, **16**, 571–591, https://doi.org/10.1175/1520-0442(2003)016<0571:BPOASF>2.0.CO;2.

Figueroa, S. N., and Coauthors, 2016: The Brazilian Global Atmospheric Model (BAM): Performance for tropical rainfall forecasting and sensitivity to convective scheme and horizontal resolution. *Wea. Forecasting*, **31**, 1547–1572, https://doi.org/10.1175/WAF-D-16-0062.1.

Francis, J. A., 2002: Validation of reanalysis upper-level winds in the Arctic with independent rawinsonde data. *Geophys. Res. Lett.*, **29**, 1315, https://doi.org/10.1029/2001GL014578.

Frolov, S., W. Campbell, B. Ruston, C. H. Bishop, D. Kuhl, M. Flatau, and J. McLay, 2020: Assimilation of low-peaking satellite observations using the coupled interface framework. *Mon. Wea. Rev.*, **148**, 637–665, https://doi.org/10.1175/MWR-D-19-0029.1.

Gentemann, C. L., and Coauthors, 2020: Saildrone: Adaptively sampling the marine environment. *Bull. Amer. Meteor. Soc.*, **101**, E744–E762, https://doi.org/10.1175/BAMS-D-19-0015.1.

Goessling, H. F., and Coauthors, 2016: Paving the way for the year of polar prediction. *Bull. Amer. Meteor. Soc.*, **97**, ES85–ES88, https://doi.org/10.1175/BAMS-D-15-00270.1.

Good, S., and Coauthors, 2020: The current configuration of the OSTIA system for operational production of foundation sea surface temperature and ice concentration analyses. *Remote Sens.*, **12**, 720, https://doi.org/10.3390/rs12040720.

Gordon, N., T. Jung, and S. Klebe, 2014: The polar prediction project. *WMO Bull.*, **63**, 42–44.

Grebmeier, J. M., S. E. Moore, L. W. Cooper, and K. E. Frey, 2019: The distributed biological observatory: A change detection array in the Pacific Arctic—An introduction. *Deep-Sea Res. II*, **162**, 1–7, https://doi.org/10.1016/j.dsr2.2019.05.005.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hunke, E. C., D. Notz, A. K. Turner, and M. Vancoppenolle, 2011: The multiphase physics of sea ice: A review for model developers. *Cryosphere*, **5**, 989–1009, https://doi.org/10.5194/tc-5-989-2011.

Inoue, J., 2020: Review of forecast skills for weather and sea ice in supporting Arctic navigation. *Polar Sci.*, **27**, 100523, https://doi.org/10.1016/j.polar.2020.100523.

Judt, F., 2018: Insights into atmospheric predictability through global convection-permitting model simulations. *J. Atmos. Sci.*, **75**, 1477–1497, https://doi.org/10.1175/JAS-D-17-0343.1.

——, 2020: Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *J. Atmos. Sci.*, **77**, 257–276, https://doi.org/10.1175/JAS-D-19-0116.1.

Jung, T., and M. Matsueda, 2016: Verification of global numerical weather forecasting systems in polar regions using TIGGE data. *Quart. J. Roy. Meteor. Soc.*, **142**, 574–582, https://doi.org/10.1002/qj.2437.

——, and Coauthors, 2016: Advancing polar prediction capabilities on daily to seasonal time scales. *Bull. Amer. Meteor. Soc.*, **97**, 1631–1647, https://doi.org/10.1175/BAMS-D-14-00246.1.

Køltzow, M., B. Casati, E. Bazile, T. Haiden, and T. Valkonen, 2019: An NWP model intercomparison of surface weather parameters in the European Arctic during the year of polar prediction special observing period Northern Hemisphere 1. *Wea. Forecasting*, **34**, 959–983, https://doi.org/10.1175/WAF-D-19-0003.1.

Krishnamurthy, V., 2019: Predictability of weather and climate. *Earth Space Sci.*, **6**, 1043–1056, https://doi.org/10.1029/2019EA000586.

Kuhn, C. E., and Coauthors, 2020: Test of unmanned surface vehicles to conduct remote focal follow studies of a marine predator. *Mar. Ecol. Prog. Ser.*, **635**, 1–7, https://doi.org/10.3354/meps13224.

Lawrence, H., N. Bormann, I. Sandu, J. Day, J. Farnan, and P. Bauer, 2019: Use and impact of Arctic observations in the ECMWF numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **145**, 3432–3454, https://doi.org/10.1002/qj.3628.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quart. J. Roy. Meteor. Soc.*, **145**, 107–128, https://doi.org/10.1002/qj.3387.

Lewis, K. M., G. L. van Dijken, and K. R. Arrigo, 2020: Changes in phytoplankton concentration now drive increased Arctic Ocean primary production. *Science*, **369**, 198–202, https://doi.org/10.1126/science.aay8380.

Li, M., and Coauthors, 2019: Circulation of the Chukchi Sea shelfbreak and slope from moored timeseries. *Prog. Oceanogr.*, **172**, 14–33, https://doi.org/10.1016/j.pocean.2019.01.002.

Lindsay, R., 1998: Temporal variability of the energy balance of thick Arctic pack ice. *J. Climate*, **11**, 313–333, https://doi.org/10.1175/1520-0442(1998)011<0313:TVOTEB>2.0.CO;2.

——, M. Wensnahan, A. Schweiger, and J. Zhang, 2014: Evaluation of seven different atmospheric reanalysis products in the Arctic. *J. Climate*, **27**, 2588–2606, https://doi.org/10.1175/JCLI-D-13-00014.1.

Liu, Y., and J. R. Key, 2016: Assessment of Arctic cloud cover anomalies in atmospheric reanalysis products using satellite data. *J. Climate*, **29**, 6065–6083, https://doi.org/10.1175/JCLI-D-15-0861.1.

Lorenc, A. C., and R. T. Marriot, 2014: Forecast sensitivity to observations in the Met Office Global numerical weather prediction system. *Quart. J. Roy. Meteor. Soc.*, **140**, 209–224, https://doi.org/10.1002/qj.2122.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Lüpkes, C., T. Vihma, E. Jakobson, G. König-Langlo, and A. Tetzlaff, 2010: Meteorological observations from ship cruises during summer to the central Arctic: A comparison with reanalysis data. *Geophys. Res. Lett.*, **37**, L09810, https://doi.org/10.1029/2010GL042724.

Mears, C. A., K. Deborah, and F. J. Wentz, 2001: Comparison of special sensor microwave imager and buoy measured wind speeds from 1987 to 1997. *J. Geophys. Res.*, **106**, 11 719–11 729, https://doi.org/10.1029/1999JC000097.

Meier, W. N., and Coauthors, 2019: Sea ice. NOAA Arctic Report Card, NOAA, https://doi.org/10.25923/y2wd-fn85.

Meinig, C., N. Lawrence-Slavas, R. Jenkins, and H. M. Tabisola, 2015: The use of Saildrones to examine spring conditions in the Bering Sea: Vehicle specification and mission performance. *OCEANS 2015-MTS/IEEE Washington*, Washington, DC, Institute of Electrical and Electronics Engineers, https://doi.org/10.23919/OCEANS.2015.7404348.

——, and Coauthors, 2019: Public–private partnerships to advance regional ocean observing capabilities: A Saildrone and NOAA-PMEL case study and future considerations to expand to global scale observing. *Front. Mar. Sci.*, **6**, 448, https://doi.org/10.3389/fmars.2019.00448.

Mordy, C. W., and Coauthors, 2017: Advances in ecosystem research: Saildrone surveys of oceanography, fish, and marine

mammals in the Bering Sea. *Oceanography*, **30**, 113–115, https://doi.org/10.5670/oceanog.2017.230.

Mulholland, D. P., P. Laloyaux, K. Haines, and M. A. Balmaseda, 2015: Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. *Mon. Wea. Rev.*, **143**, 4631–4644, https://doi.org/10.1175/MWR-D-15-0076.1.

Naakka, T., T. Nygård, M. Tjernstrom, T. Vihma, R. Pirazzini, and I. M. Brooks, 2019: The impact of radiosounding observations on numerical weather prediction analyses in the Arctic. *Geophys. Res. Lett.*, **46**, 8527–8535, https://doi.org/10.1029/2019GL083332.

Overland, J. E., and Coauthors, 2019: Surface air temperature. NOAA Arctic Report Card 2020, NOAA, 7 pp., https://doi.org/10.25923/gcw8-2z06.

Peixoto, J. P., and A. H. Oort, 1992: *Physics of Climate.* American Institute of Physics, 520 pp.

Sallila, H., S. L. Farrell, J. McCurry, and E. Rinne, 2019: Assessment of contemporary satellite sea ice thickness products for Arctic sea ice. *Cryosphere*, **13**, 1187–1213, https://doi.org/10.5194/tc-13-1187-2019.

Schellekens, J., A. H. Weerts, R. J. Moore, C. E. Pierce, and S. Hildon, 2011: The use of MOGREPS ensemble rainfall forecasts in operational flood forecasting systems across England and Wales. *Adv. Geosci.*, **29**, 77–84, https://doi.org/10.5194/adgeo-29-77-2011.

Scott, J. P., S. Crooke, I. Cetinić, C. E. Del Castillo, and C. L. Gentemann, 2020: Correcting non-photochemical quenching of Saildrone chlorophyll-a fluorescence for evaluation of satellite ocean color retrievals. *Opt. Express*, **28**, 4274–4285, https://doi.org/10.1364/OE.382029.

Sedlar, J., and M. Tjernstrom, 2019: A climatological process-based evaluation of AIRS tropospheric thermodynamics over the high-latitude Arctic. *J. Appl. Meteor. Climatol.*, **58**, 1867–1886, https://doi.org/10.1175/JAMC-D-18-0306.1.

——, and Coauthors, 2020: Confronting Arctic troposphere, clouds, and surface energy budget representations in regional climate models with observations. *J. Geophys. Res. Atmos.*, **125**, e2019JD031783, https://doi.org/10.1029/2019JD031783.

Shen, B.-W., R. A. Pielke Sr., X. Zeng, J.-J. Baik, S. Faghih-Naini, J. Cui, and R. Atlas, 2021: Is weather chaotic? Coexistence of chaos and order within a generalized Lorenz model. *Bull. Amer. Meteor. Soc.*, **102**, E148–E158, https://doi.org/10.1175/BAMS-D-19-0165.1.

Shen, X. S., and Coauthors, 2020: Research and operational development of numerical weather prediction in China. *J. Meteor. Res.*, **34**, 675–698, https://doi.org/10.1007/s13351-020-9847-6.

Smith, G. C., and Coauthors, 2019: Polar ocean observations: A critical gap in the observing system and its effect on environmental predictions from hours to a season. *Front. Mar. Sci.*, **6**, 429, https://doi.org/10.3389/fmars.2019.00429.

Steele, M., W. Ermold, and J. Zhang, 2008: Arctic Ocean surface warming trends over the past 100 years. *Geophys. Res. Lett.*, **35**, L02614, https://doi.org/10.1029/2007GL031651.

Stevens, B., J. Duan, J. C. McWilliams, M. Münnich, and J. D. Neelin, 2002: Entrainment, Rayleigh friction, and boundary layer winds over the tropical Pacific. *J. Climate*, **15**, 30–44, https://doi.org/10.1175/1520-0442(2002)015<0030:ERFABL>2.0.CO;2.

Sutton, A. J., N. L. Williams, and B. Tilbrook, 2021: Constraining Southern Ocean CO_2 flux uncertainty using uncrewed surface vehicle observations. *Geophys. Res. Lett.*, **48**, e2020GL091748, https://doi.org/10.1029/2020GL091748.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Timmermans, M.-L., and C. Ladd, 2019: Sea surface temperature. NOAA Arctic Report Card, https://arctic.noaa.gov/Report-Card/Report-Card-2019/ArtMID/7916/ArticleID/840/Sea-Surface-Temperature.

Tjernström, M., G. Svensson, L. Magnusson, I. M. Brooks, J. Prytherch, J. Vüllers, and G. Young, 2021: Central Arctic weather forecasting: Confronting the ECMWF IFS with observations from the Arctic Ocean 2018 expedition. *Quart. J. Roy. Meteor. Soc.*, **147**, 1278–1299, https://doi.org/10.1002/qj.3971.

Tomita, H., T. Hihara, S. I. Kako, M. Kubota, and K. Kutsuwada, 2019: An introduction to J-OFURO3, a third-generation Japanese ocean flux data set using remote-sensing observations. *J. Oceanogr.*, **75**, 171–194, https://doi.org/10.1007/s10872-018-0493-x.

Vazquez-Cuervo, J., J. Gomez-Valdes, M. Bouali, L. E. Miranda, T. Van der Stocken, W. Tang, and C. Gentemann, 2019: Using saildrones to validate satellite-derived sea surface salinity and sea surface temperature along the California/Baja Coast. *Remote Sens.*, **11**, 1964, https://doi.org/10.3390/rs11171964.

Vihma, T., 2014: Effects of Arctic sea ice decline on weather and climate: A review. *Surv. Geophys.*, **35**, 1175–1214, https://doi.org/10.1007/s10712-014-9284-0.

Wang, M., and J. E. Overland, 2009: A sea ice free summer arctic within 30 years? *Geophys. Res. Lett.*, **36**, L07502, https://doi.org/10.1029/2009GL037820.

Yadav, J., A. Kumar, and R. Mohan, 2020: Dramatic decline of Arctic sea ice linked to global warming. *Nat. Hazards*, **103**, 2617–2621, https://doi.org/10.1007/s11069-020-04064-y.

Yamagami, A., M. Matsueda, and H. L. Tanaka, 2019: Skill of medium-range reforecast for summertime extraordinary Arctic cyclones in 1986–2016. *Pol. Sci.*, **20**, 107–116, https://doi.org/10.1016/j.polar.2019.02.003.

Yang, F., and V. Tallapragada, 2018: Evaluation of retrospective and real-time NGGPS FV3GFS experiments for Q3FY18 beta implementation. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 5B.3, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345231.html.

Yonehara, H., and Coauthors, 2018: Upgrade to JMA's operational NWP high-resolution global model. *CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling*, RSMC Tech Rev. 23, Tokyo, Japan, RSMC, 6 pp., https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/techrev/text23-1.pdf.

Yu, L., and R. A. Weller, 2007: Objectively analyzed air–sea heat fluxes for the global ice-free oceans (1981–2005). *Bull. Amer. Meteor. Soc.*, **88**, 527–540, https://doi.org/10.1175/BAMS-88-4-527.

Zhang, D., and Coauthors, 2019: Comparing air-sea flux measurements from a new unmanned surface vehicle and proven platforms during the SPURS-2 field campaign. *Oceanography*, **32**, 122–133, https://doi.org/10.5670/oceanog.2019.220.

Zhu, Y., and R. Gelaro, 2008: Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. *Mon. Wea. Rev.*, **136**, 335–351, https://doi.org/10.1175/MWR3525.1.