Validation of Satellite Sea Surface Temperature Analyses in the Beaufort Sea Using UpTempO Buoys

Sandra L. Castro[1], Gary A. Wick[2], and Michael Steele[3]

[1]Department of Aerospace Engineering Sciences, University of Colorado at Boulder, Boulder, CO
[2]NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder, CO
[3]Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, WA

Abstract

Many different blended sea surface temperature (SST) analyses are currently available and exhibit significant differences in the high latitude regions. It is challenging for users to determine which of these products is most accurate and best suited for their applications. Nine different SST analyses and two single sensor satellite products are compared with independent observations from Upper Temperature of the polar Oceans (UpTempO) buoys deployed in the Beaufort Sea in 2012 and 2013 during the Marginal Ice Zone Processes Experiment (MIZOPEX). The relative skill of the different SST products is evaluated using a combination of Taylor diagrams and two different verification scores that weight different statistical measures. Skill thresholds based on satellite accuracy requirements are chosen to map products with similar performance into three discrete skill categories: excellent, good, and poor. Results are presented for three subsets of the buoys corresponding to different regimes: coastal waters, northerly waters, and extreme weather. The presence of strong thermal gradients and cloudiness posed problems for the SST products, while in more homogeneous regions the performance was improved and more similar among products. The impact of variations in the ice mask between the SST products was mostly inconsequential. While the relative performance of the analyses varied with regime, overall, the best performing analyses for this region and period included the NOAA Optimal Interpolation SST (OISST), the Canadian Meteorological Centre (CMC) SST, and the Group for High Resolution SST (GHRSST) Multi-Product Ensemble (GMPE).

1. Introduction

Accurate monitoring of environmental conditions in the Arctic warrants particular attention both for the unique measurement challenges and the potential for the high latitudes to serve as an early and strong indicator of potential climate change. Sea surface temperature (SST) is a fundamental variable critical to weather predictions, climate monitoring, and ship-based operations at high latitudes. An inherent challenge of high-latitude satellite SST production is persistent cloudiness, which hinders SST retrievals in the infrared portion of the spectrum, resulting in extended gaps in the satellite imagery. Microwave sensors, although able to "see" through clouds, have limitations close to land and ice. Multi-sensor, gridded, satellite-based SST analyses (Level 4 or simply L4 satellite products) offer tremendous potential for monitoring conditions throughout the Arctic domain over extended periods, since they combine information available from multiple satellites and types of sensors in an objective analysis to fill in the coverage gaps, but much remains unknown about the accuracy and representativeness of the SST analyses at these latitudes.

A large number of these gap-free SST products have been developed over recent years (Table 1). This abundance, however, presents the users with the challenge of choosing the analysis product that best suits their purpose. L4 SST products are available through the Group for High Resolution Sea Surface Temperature (GHRSST; Donlon et al., 2007), and are distributed in a common format for easy use. A thorough comparison of the GHRSST L4 SST analyses is described in Martin et al. (2012) and Dash et al. (2012). A persistent problem, however, is that while the different analyses perform fairly uniformly globally or in basin-wide regions, there are significant differences at high latitudes. As Dash et al. (2012) point out, mean analysis differences in excess of 2°C are frequently observed in the Arctic Ocean.

The research presented here was conducted in association with the Marginal Ice Zone Processes Experiment (MIZOPEX). This was a multi-institutional, multi-instrument Arctic observing campaign (http://ccar.colorado.edu/mizopex) led by the University of Colorado with the support of the National Aeronautics and Space Administration (NASA)and the National Oceanic and Atmospheric Administration (NOAA), that involved coordinating different types of unmanned autonomous vehicles (UAVs), with the simultaneous deployment of in situ instruments and satellite overpasses over the marginal ice zone (MIZ) to measure SST and sea ice during the 2012 – 2013 melt seasons. Our intercomparison of L4 SST products was motivated by the need of MIZOPEX management and flight planners to know which of the satellite SST products provided the most accurate information over the study area. This area is particularly challenging since persistent cloudiness during the melting season results in few infrared (IR) retrievals, and proximity to land and ice hinders microwave (MW) SST retrievals. The highest cloudiness in the Arctic (~80 – 90%) occurs from June to October (Przybylak, 2003), encompassing the duration of the MIZOPEX field campaign. The sparseness of IR SST retrievals made lower level (Level 2 and Level 3) SST products unfavorable for airborne mission planning, and even though MW SST retrievals are a valid option under these conditions, the Advanced Microwave Scanning Radiometer for EOS (AMSR-E) traditionally used for MW products was not operating at the time of this study. The one remaining option was to choose a satellite product from the plethora of L4 SST analyses. SST maps of the Beaufort Sea from several widely used analyses exhibited stark differences during the days leading to the field campaign. These discrepancies made many of the SST analyses unreliable for determining the locations where experiment activities should be conducted, but motivated devising a framework for choosing the more skillful L4 products for our particular application.

Validation of the SST analyses is especially challenging in the Arctic Ocean and in regions near sea ice due to the limited number of in situ SST observations. An unprecedented set of high quality buoys have been deployed in the Arctic Ocean by the Polar Science Center of the Applied Physics

Laboratory (APL) at the University of Washington every spring and summer since 2010. Though limited in number, the quality and high resolution of these Upper layer Temperature of the polar Oceans (UpTempO) buoys provides a unique opportunity to validate the different SST analyses in the region.

In this paper, we employ UpTempO buoys to perform a systematic inter-comparison of multiple SST analyses in the Beaufort Sea during the Arctic summers of 2012 and 2013. The quality of the individual L4 analyses, measured relative to the UptempO buoys, is demonstrated using a combination of performance metrics such as Taylor diagrams and skill scores. The main aims are to assess which of the products performs best in the Beaufort Sea, and to test a methodology for ranking the skill of the analyses. The UpTempO buoys uniquely facilitated this study by providing high-quality in situ observations independent from the analyses (at the time of this study, the UptempO SSTs were not being reported via the Global Telecommunications System (GTS)). Our focus in this study is on the seasonally open water of the Beaufort Sea, i.e., we avoid areas of sea ice cover. The UpTempO buoys are described in detail along with the SST analyses in section 2. The collocation approach and evaluation methodology are presented in section 3, followed by the results obtained in section 4, and conclusions in section 5.

2. Data Description

2.1. Level 4 SST Products:

SST L4 analyses are interpolated (gap-free), gridded SST products. These analyses assimilate both IR and MW satellite SSTs, as they are highly complementary and their error characteristics are independent of each other. The main passive MW instrument used for SST retrievals, AMSR-E, failed on October 5, 2011, and data from its successor, AMSR2, was first released in January 2014. During the gap between AMSR instruments, which coincidentally overlaps with our study period, some of the satellite SST data producers resorted to an alternative MW data source, WindSat, while others abstained from using any MW data at all, or temporarily halted production of their MW-based SST products. MW data is especially valuable in regions with persistent cloudiness where the “all-weather” coverage of MW sensors results in significant improvements in accuracy. As Brasnett (2008) points out, for some of these analyses, the MW and IR data contribute in equal measure to the analysis quality. It is understood then that the performance (accuracy) of the L4 products compared here, especially those that rely on MW data, was greatly compromised during the study period by the special circumstances of not having an AMSR instrument.

Most of the SST analyses used here are available in GHRSST NetCDF format, and can be downloaded from the GHRSST Long Term Stewardship and Reanalysis Facility (LTSRF) at the NOAA National Centers for Environmental Information (NCEI: www.nodc.noaa.gov/sog/GHRSST/accessdata.html). The following sections include brief introductions to the L4 products used in this inter-comparison, as they are extensively described elsewhere (see Table 1 for key references). Main features and contributing data sources for all the SST analyses are summarized in **Tables 1 and 2**, respectively. The different analyses can represent different SST quantities ranging from a simple daily average temperature to the “foundation” temperature representing the SST at a depth free from diurnal variability (e.g. Castro et al., 2014).

Table 1. Characteristics of the SST products considered in this analysis.

Level	Product	Data Producer	Spatial Resolution	SST Type	Ice Mask Source	Reference
L4	**CMC**	Canadian Meteorological Centre	0.20°	Foundation	CMC	Brasnett (2008)
	FNMOC	Naval Research Laboratory	9 km	Skin	FNMOC	Cummings and Smedstad (2013)
	GAMSSA	Australian Bureau of Meteorology/ BLUElink	0.25°	Foundation	NCEP	Beggs et al. (2011); Zhong and Beggs (2008)
	GMPE	UK Met Office	0.25°	Median Foundation	OSI-SAF	Martin et al. (2012)
	K10	NAVOCEANO	0.10°	Daily average at depth	N/A	
	MUR	NASA JPL	0.01°	Foundation	OSI-SAF	Chin et al. (1998)
	MWIR	REMSS	9 km	Foundation	OSI-SAF	Gentemann et al. (2006)
	OISST	NOAA/NCDC	0.25°	Daily average at depth	NCEP	Reynolds et al. (2007)
	OSTIA	UK Met Office	0.05° (~6 km)	Foundation	OSI-SAF	Donlon et al. (2012)
L3	**LAC**	NAVOCEANO	2 km	Skin		May et al. (1998)
	WindSat	REMSS	0.25°	Subskin		

Table 2. In situ and satellite data sources ingested into the different L4 SST products during 2012–2013. Note that while the Geosynchronous (IR Geo) and TMI sensors are included for completeness, they do not provide coverage in the Beaufort Sea, and thus are not discussed in the text.

Data type	In Situ			IR Polar			IR Geo		MW		
L4 product	Argo floats	Buoys GTS	Ships GTS	AVHRR NOAA	AVHRR MetOp	MODIS Aqua,Terra	SEVIRI MSG	GOES	TMI TRMM	WindSat	WindSat Ingest
CMC	✓	✓	✓	✓	✓				✓	✓	01/12
FNMOC	✓	✓	✓	✓	✓		✓	✓			
GAMSSA		✓	✓	✓	✓					✓	12/12
K10				✓	✓			✓		✓	01/13
MUR		✓		✓		✓				✓	10/11
MWIR						✓			✓	✓	10/11
OISST		✓	✓	✓	✓						
OSTIA		✓	✓	✓	✓		✓		✓		

The Canadian Meteorological Centre (CMC) SST analysis is tailored to the needs of the CMC numerical weather prediction (NWP) system (Brasnett, 2008). It merges the observations listed in Table 2 using optimal interpolation (OI, e.g., Gandin, 1965; Daley, 1991) to provide a daily foundation SST analysis.

Fleet Numerical Meteorology and Oceanography Center (FNMOC) SSTs: The US Office of Naval Research uses its multivariate OI analysis system, the Navy Coupled Ocean Data Assimilation version 3 (NCODA 3DVAR), run operationally at FNMOC, to produce global SST and sea ice concentration analyses for GHRSST. The analyses are executed using a 6-hour update cycle with the U.S. Navy ocean forecast model, the global Hybrid Coordinate Ocean Model (HYCOM), and are available within 6 hours of real-time. For the purpose of this intercomparison, only the 12:00 UTC –SST forecast will be used. The system assimilates satellite SSTs, in situ SSTs, temperature and salinity profiles, altimetric sea surface heights, and satellite sea ice observations. The analyses have a 12-km resolution at the equator and 9-km resolution at mid latitudes. The FNOMC GHRSST analyses are available through the US GODAE server at http://www.usgodae.org.

Global Australian Multi-Sensor SST Analysis (GAMSSA): the Australian Bureau of Meteorology produces this daily foundation SST analysis on a 1/4° grid, and it is used operationally as a boundary condition in their global NWP system and to initialize their seasonal forecast system. The GAMSSA is an extension of their 1/12° regional L4 product (RAMSSA: Beggs et al., 2011). The OI system ingests in situ SST and both IR and MW satellite SST data (Zhong and Beggs, 2008). Data are rejected for low NWP wind speed thresholds (6 m/s day, 2 m/s night) to reduce effects from diurnal warming on the analysis.

The Naval Oceanographic Office (NAVOCEANO) K10 SST analysis uses satellite data only, and combines the L2 SST products in a weighted average tuned to represent the SST at 1-m depth. This is one of the few L4s that does not use OI techniques. All the IR inputs are produced by NAVOCEANO using separate nonlinear regressions trained against quality-controlled GTS drifting buoys from the previous month. To preserve features, the weights decrease exponentially from the center of the averaging window and the elapsed time from the last observation (B. McKenzie, personal communication, 2011).

The Multi-scale Ultra-high Resolution (MUR) SST analysis is produced daily by the NASA Jet Propulsion Laboratory (JPL). In contrast to other L4s for which more traditional OI techniques are used, the MUR system uses a statistical interpolation method based on wavelet decomposition called Multi-Resolution Variational Analysis (e.g., Mallat 1989). This multiscale signal reconstruction technique is particularly suitable for dealing with the multiple spatial resolutions of the L2 products entering the analyses and the irregular swath patterns of the different satellites (Chin et al., 1998). The main contribution of this product is its fine spatial (horizontal) resolution and capability for resolving high-resolution SST features such as fronts.

The Remote Sensing Systems (REMSS) MW-IR (MWIR) SST product uses an OI analysis and satellite data only (http://www.remss.com/measurements/sea-surface-temperature/oisst-description). Inputs are diurnally corrected using an empirical diurnal warming model. This foundation SST product was originally designed for the National Hurricane Center to be used in conjunction with the Statistical Hurricane Intensity Prediction Scheme (SHIPS) model for hurricane intensity forecasting.

The NOAA OISST (Reynolds et al., 2007) is generated by NCEI (formerly the National Climatic Data Center). While both IR-only and IR-MW products are normally generated, the L4 analysis herein refers to the IR-only product due to the gap in AMSR data. In addition to the IR SSTs this product uses in situ data from ships and buoys as well as proxy SSTs, generated from sea ice concentrations, for the MIZ. The product was designed for applications that target high-resolution features such as fronts and hurricane forecasting, and to serve as boundary condition for atmospheric models. It represents a daily average SST (no diurnal warming correction is attempted), that is bias-adjusted using a spatially smoothed, 7-day in situ SST mean.

The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) analysis (Donlon et al., 2012) is produced daily by the UK Met Office and used operationally as a boundary condition in NWP and Numerical Ocean Forecast systems at the Met Office and the European Centre for Medium-range Forecasting (ECMWF). Using an OI analysis the OSTIA system normally assimilates MW data from AMSR, but abstained from ingesting WindSat SSTs during the gap period. Input data are filtered to remove daytime observations with winds < 6 m/s to eliminate possible instances of diurnal warming. While provided on a 0.05° (~6 km)-grid the OSTIA SSTs are effectively smoother by design (Donlon et al., 2012).

The GHRSST Multi-Product Ensemble (GMPE) system (Martin et al., 2012), developed and operated by the UK Met Office, consists of the daily ensemble median and standard deviation, on a homogenized 0.25°grid, of various GHRSST L4 operational analyses. Inputs contributing to the ensemble median include all of the above (with the exception of MUR at the time of this study), plus three other SST analyses not included in this study (Martin et al., 2012). Although the original purpose of the GMPE system was to advise GHRSST data users on the relative performance of the different L4 products by providing a near-real time global ensemble from a large number of L4 SST analyses, the system has proven useful for climate-related SST application. The GMPE data is available via the MyOcean project (http://myocean.eu.org).

Please note that while most analyses ingest AVHRR (NOAA and/or Metop) L2 SSTs (see Table 2), there are multiple AVHRR data providers; hence, the source of the AVHRR data can differ among L4 products. Additionally, different AVHRR satellites can be active at any given time. During the study period, NOAA-16, -18, -19, and Metop A and B were all being used, with NOAA-19 being the designated operational afternoon orbit and Metop (A in 2012, B in 2013) the morning orbit. For additional information about the specific AVHRR data sources and sensors being used, please consult the reference for the appropriate analysis or the metadata source field available in all GHRSST-compliant SST products. While Table 2 lists the different L2 products ingested during the study period, another important IR sensor, the Advanced Along-Track Scanning Radiometer (AATSR), was routinely ingested by some of the analyses (i.e., CMS, FNMOC, GAMSSA, and OSTIA), but ceased operations in April 2012.

2.2. Ice Masking

A main difference among the L4 SST analyses is their treatment of the SSTs near or under ice. Most of the L4s use independent sea ice concentration (SIC) analyses, generated from space-borne microwave sensors, to derive an ice mask based on some ice concentration threshold. That is, if the SIC in an analysis grid exceeds a minimum ice fraction, I_o, then the corresponding grid cell in the SST product is flagged as ice. While some SST producers opt for not reporting SSTs once $SIC \geq I_o$, others use ice information to compute proxy SSTs in the range $[I_o, 1]$, i.e., the SSTs are relaxed towards the freezing point temperature of seawater using empirical relationships between SIC and SST. At the time of this study, the only L4s that simulated proxy SSTs in $[I_o, 1]$ were the OISST, OSTIA and FNMOC. Detailed descriptions of the different methodologies used to simulate SSTs under ice can be found in the product references in Table 1. All other L4s set SST = -1.8°C in locations where $SIC \geq 0.5$.

The two most widely used SIC analyses are the ones produced by the NOAA National Centers for Environmental Prediction (NCEP)-Marine Modeling and Analysis Branch (MMAB) and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) - Ocean and Sea Ice Satellite Applications Facility (OSI-SAF). The newest (starting June 2012) NOAA/NCEP operational SIC analysis (Grumbine, 1996), distributed daily on 12.7-km polar stereographic projection hemispheric grids, uses the NASA Team 2 (NT2) sea ice retrieval algorithm (Markus and Cavalieri, 2000) and ice retrievals from the Special Sensor Microwave Imager (SSM/I) and Special Sensor Microwave Imager/Sounder (SSMIS). The SST analyses that rely on the NCEP SIC product for ice information are GAMSSA and the OISST. The OSI-SAF SIC product (Andersen et al., 2007) is based on the SSM/I sensors and is distributed on a 10-km polar stereographic projection grid every 24 h. The MUR, OSTIA, GMPE, and MWIR use the OSI-SAF SIC analysis for their ice mask (see Table 1). The FNMOC and CMC, being part of fully integrated NWP systems, construct their own SIC analyses. The FNMOC ice analysis system (Cummings and Smedstad, 2013) assimilates SSM/I and SSMIS ice retrievals, and use the NT2 to calculate SIC, using 6-h forecast windows. At CMC, the Global Ice Ocean Prediction System (GIOPS) assimilates passive MW satellite observations together with manual analysis from the Canadian Ice Service to provide a daily, global ice (and ocean) analysis (Buehner et al., 2013; Smith et al., 2015). The K10 is the only L4 that did not use an ice mask at the time of this study, relying on sea-ice extent climatologies instead. An ice mask for the K10 was not introduced until 2016.

2.3. Single Sensor SST Products

Even though the emphasis of this study is on the L4 products, we have also included direct comparisons against two different single-sensor SST products used as input in some of the analyses under consideration. These satellite SST products are lower processing-level data, i.e., there has been no intervention to fill in the gaps, but they have higher spatial resolution. They can be distributed in the swath of the sensor (Level 2 or L2), or can be gridded for distribution (Level 3 or L3), taking care of preserving the gaps during the gridding process. These are the NOAA Advanced Very High Resolution Radiometer (AVHRR) Local Area Coverage (LAC) SSTs produced by NAVOCEANO and the REMSS WindSat SSTs. The reason for including these products is two-fold: it helps interpret differences among the multiple L4 products and provides some information on potential limits to the accuracy of the analyses.

The LAC SST (obtainable from NAVOCEANO on request) is a 2-km L2 product. The native AVHRR LAC radiances have 1.1 km resolution at nadir, but retrieved signals from adjacent pixels are averaged into 2x2 pixel windows before being ingested in the NAVOCEANO SST algorithm. This effectively reduces the resolution of the L2 SST product to ~2 km. The L2 LAC SSTs were further mapped in house onto 2 km daily grids to facilitate comparison with the analyses. Hence, technically speaking, this product was transformed into a “collated level 3,” and hereafter it will be referred to as a L3, despite the fact that the resolution of the grid did not allow for further aggregation (oversampling) of the data and it is distributed in L2 format. The AVHRR LAC SSTs from NOAA-19 were selected for this study. The WindSat SSTs (available from www.remss.com) is a L3 product, mapped onto a global 0.25°x0.25° regular grid. WindSat SSTs were largely ignored before the AMSR gap because it was a demonstration mission sponsored by the U.S. Navy and did not always satisfy timeliness requirements for an operational mission. Including this product offers a good opportunity to assess the impact of assimilating WindSat SSTs on the quality of the analyses. This is an important issue for future satellite SST production since, unlike the diversity of IR spaceborne sensors, there is no redundancy of MW sensors.

2.4. UpTempO Buoys

The UpTempO buoys discussed here all consisted of a drifting surface float and a 60-80 m long string of thermistors. A schematic is provided in **Figure 1**. The buoys reported hourly via Iridium satellite until individual sensors or the entire buoy failed (often this happened in fall or winter during ice rafting and ridging). A full description of these buoys will be provided in a forthcoming manuscript (Steele et al., 2016). Here we provide a brief description of the particular buoys used in this study. While all of these buoys had a string of thermistors, here we generally only use the uppermost sensor in order to determine SST. Deeper thermistor data are in fact used initially, but only to show the frequent presence of an isothermal surface layer. Three different manufacturers supplied the drifting buoys used in this study:

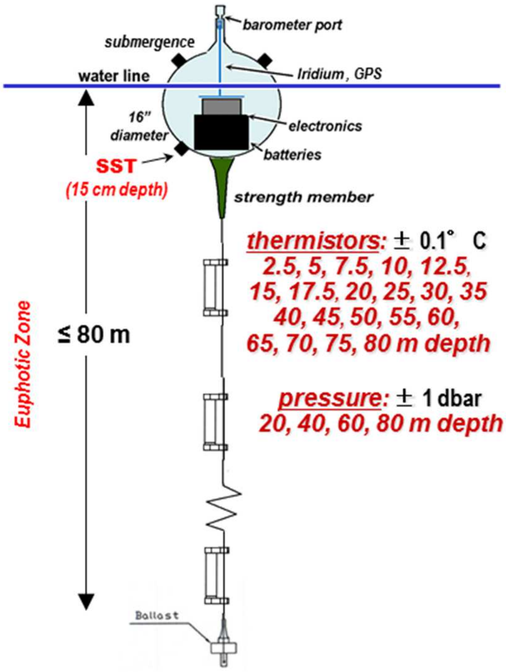

Figure 1. UpTempO buoy schematic showing key components and typical depths for temperature sensors.

(1) Louis 2012-03 and Louis 2012-04 were made by MetOcean Data Systems in Bedford, Nova Scotia, Canada. These had a surface hull with electronics, batteries, Iridium satellite antenna, sea level pressure sensor, surface atmospheric temperature sensor on a 1-meter mast, but no SST sensor. Below the hull was a 60 m long sensor string composed of 12 thermistors and 2 ocean pressure sensors. The uppermost thermistor (the only one used in this study) was at 2.5-m nominal depth (i.e., when the sensor string was vertical). This sensor was a SBE 39 from Seabird Electronics, with manufacturer-stated accuracy of ±0.002°C.

(2) Louis 2012-05, Healy 2012-07, and Healy 2013-16, -17, and -18 were made by Marlin-Yug Ltd in Sevastopol, Ukraine. The surface float had the same features as the MetOcean buoys, but without the atmospheric temperature mast and with a thermistor in the bottom half of the hull to provide SST with ±0.1°C accuracy at 15-cm nominal depth, which is what we used in this study. These buoys also had a sensor string with 16 thermistors and one ocean pressure sensor.

(3) Ukpik 2013-04 was made by Pacific Gyre Inc. in Oceanside, California, USA. The surface float had similar features to those made by Marlin-Yug, including an SST sensor with ±0.1°C accuracy at 15-cm nominal depth (used in this study), but also including an anemometer. The sensor string had 12 thermistors, 3 ocean pressure sensors, and one Seabird SBE 37-IM for recording ocean temperature and salinity.

A map of the track of the UpTempO buoys deployed in the Beaufort Sea in 2012 and 2013 is shown in **Figure 2**. The deployment location is indicated by a triangle. Time series of SST-at-depth from the thermistors in the top 20 m for all the buoys under consideration are presented in **Figure 3**. Note that the UpTempO temperatures shown in this figure are all fairly uniform throughout the upper 10 m. The absence of significant thermal gradients near the surface facilitates the comparisons with the satellite analyses as they characterize different types of SST (Table1), mostly foundation temperatures. The well-mixed, isothermal surface layer, however, suggests that there were no differences among the L4s attributable to the SST type. For each buoy, the observations closest to the surface are utilized in the comparisons.

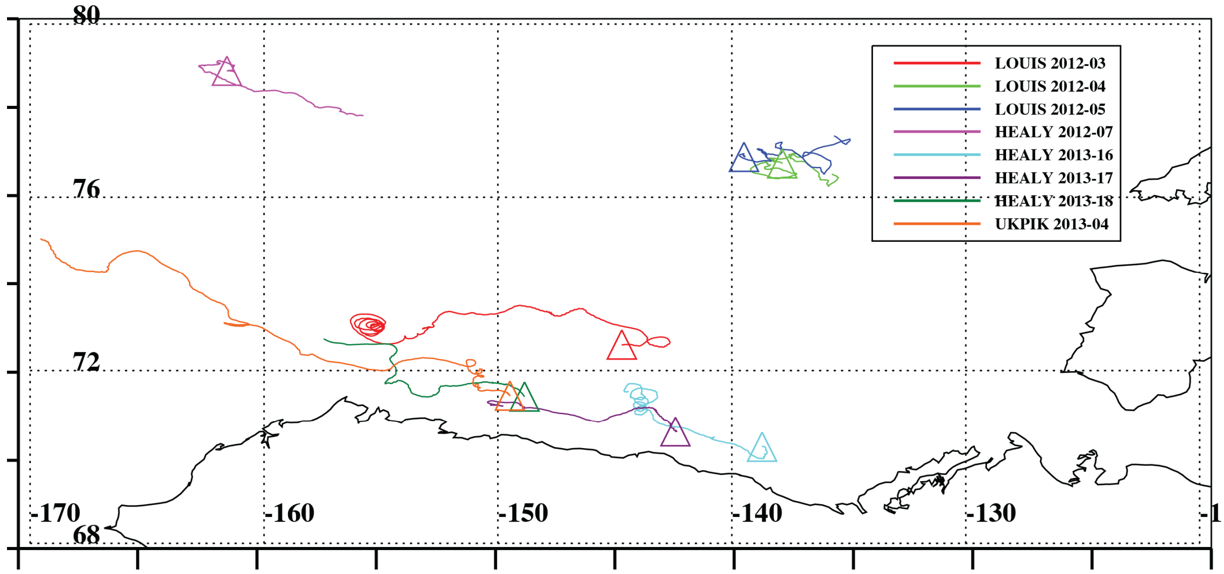

Figure 2. UpTempO buoy tracks deployed in the Beaufort Sea during the Arctic Summers of 2012 and 2013. The triangles indicate the initial deployment location.

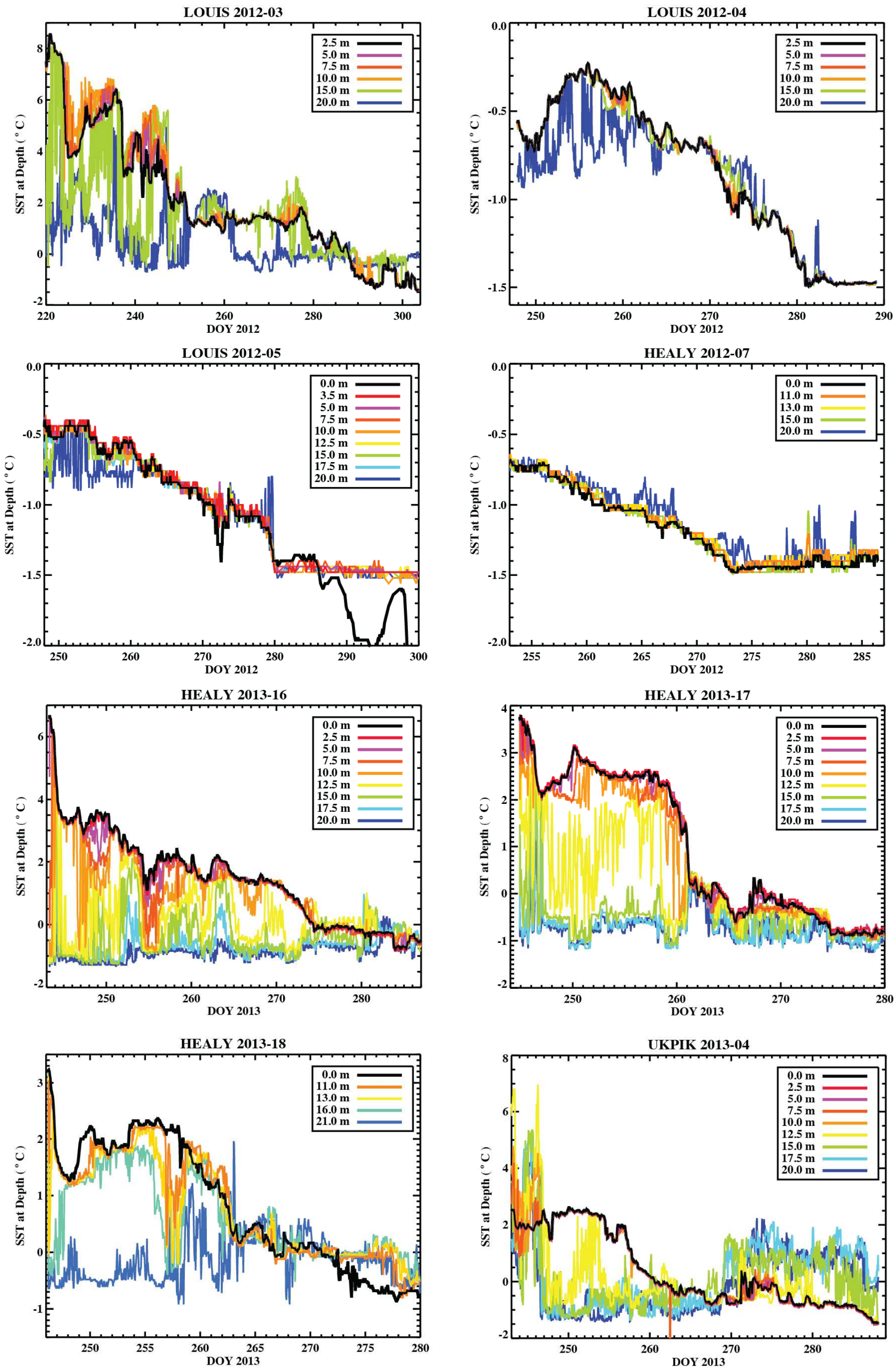

Figure 3. Time series of temperatures from the UpTempO thermistors in the top 20 m for all the buoys used in this study. The color coding corresponds to the sensor depth as indicated in the legends.

3. Theoretical Background

3.1 Taylor Diagrams

Taylor diagrams (Taylor, 2001) are the selected method for comparing the different SST analyses, since they provide a means of summarizing the relative accuracies of several competing products graphically, in such a way that they can convey information much more readily and concisely than the analogous tabular presentation. In its basic form, a Taylor diagram is a two-dimensional scatter plot in which discrete points give an indication of how closely the various L4 SSTs resemble the UpTempO observations in terms of their correlation (ρ), centered root-mean-square (RMS') error, and standard deviations (σ), all at once. These three statistics are related by the equation: $RMS' = 1/N \sum_{n=1}^{N} [(\sigma_{sat} - \bar{\sigma}_{sat}) - (\sigma_{obs} - \bar{\sigma}_{obs})]^2$. The diagram is built by plotting a triangle in a rectangular coordinate system with one point at the origin, and the other two representing the buoy observations (located along the abscissa), and the corresponding satellite SST matchups, respectively. The radial distances to the satellite and the buoy SSTs are proportional to their respective standard deviations (σ_{sat} and σ_{obs}), whereas the distance from the satellite product to the buoy observations (leg opposite the origin) is proportional to the *RMS'*. The correlation coefficient is given by the cosine of the azimuthal angle between the radii for the buoy observations and the satellite retrievals. It is important to know that the means of the SST products are subtracted out before computing σ_{sat} and σ_{obs}, so the Taylor diagram does not provide information about the overall biases; just about the centered errors.

We aim to compare the performance of L3 and L4 products relative to the diagnostic buoy data set via Taylor diagrams. There is, however, a mismatch in the number of collocated satellite – buoy pairs between the L4 and L3 comparisons due to the inherent gaps in the L3 products. Because of the different sampling sizes, the statistics must be standardized before constructing the Taylor diagrams.

One way to do this is to normalize the RMS' and all the standard deviations by the standard deviation of the matched observations, i.e., $\widehat{RMS}' = RMS'/\sigma_{obs}$, $\hat{\sigma}_{sat} = \sigma_{sat}/\sigma_{obs}$, and $\hat{\sigma}_{obs} = 1$.

The merits of the different SST products can be inferred visually just by looking at their position in the diagram. The closer the point representing a satellite product is to the buoy observations, the better the agreement between the two (satellite products lying near the observations have relatively high correlation and low RMS). In normalized diagrams, these are the points closest to the arc for $\hat{\sigma}_{obs}$. Differences between their respective variances, however, will tend to increase the RMS' pushing the point farther away from the observations. Point spread from the $\hat{\sigma}_{obs}$-arc along the radial distance from the origin will reflect differences in amplitude between the satellite product and the observations. The correlation coefficient, on the other hand, remains unaffected by differences in variance but is sensitive to the relative phasing between the estimates and their associated observations. Hence, differences in phase will be reflected in azimuthal angle spread (the angle between the point representing an SST product and the x-axis). As a result, it is possible to have satellite products whose SST patterns are uniformly too weak or too strong, and still have high correlation with the buoy measurements or, alternatively, have SST patterns that do not necessarily align with the observations despite having the right amplitude variability. In other words, the magnitude and the direction of the scattering of the points representing the SST products in the Taylor diagram helps determine whether the overall errors in the SST products are attributable to differences in variance or to poor pattern correlation.

3.2 Skill Scores

Even though Taylor diagrams are very useful in identifying common performance patterns at a glance, we also desire a verification score that helps us quantify the relative skill of the different L4s at high latitudes. Under such a scoring system, satellite products with reduced RMS' should be rewarded, since this indicates close agreement with the observations. The issue remains as to what to prioritize next: pattern similarity or correct amplitude variability. Since there is no universal skill score that satisfies all criteria simultaneously, we will consider two different skill scores that emphasize slightly different aspects of the product performance.

A basic verification score, proposed by Taylor (2001) to evaluate the skill of several precipitation models, is given by:

$$TS = \frac{4\,(1+\rho)^4}{(\hat{\sigma}_{sat} + 1/\hat{\sigma}_{sat})^2 (1+\rho_{max})^4} \tag{1}$$

where ρ_{max} is the maximum potentially realizable correlation given the uncertainty associated with unforced variability. This skill score is defined to vary from one (most skillful) to zero (absence of any skill) and fulfills the attributes stated above, i.e., $TS \to 1$ when $\hat{\sigma}_{sat} \to \hat{\sigma}_{obs} = 1$ and $\rho \to \rho_{max}$, and $TS \to 0$ as $\hat{\sigma}_{sat} \to 0$ or $\hat{\sigma}_{sat} \to \infty$ and $\rho \to 0$. Note also that when $\hat{\sigma}_{sat} \to 0, TS \propto \hat{\sigma}^2$, and when $\hat{\sigma}_{sat} \to \infty$, $TS \propto 1/\hat{\sigma}^2$; i.e, for products with small variance, the skill is proportional to the variance, but when the variance is large, the skill is inversely proportional to the variance; thus, the skill always decreases when the RMS increases.

A potential deficiency of the Taylor verification score and Taylor diagram may arise from the fact that centered moments are being used to correct for non-zero mean SST biases. Dash et al. (2012) have found differences >2°C among the L4 SST products at high latitudes. Consequently, if large unconditional biases exist in some of the satellite-derived SST products, the TS score can substantially overestimate their skill. Following Murphy (1988), we propose a complementary skill score that uses the un-centered second moments, to account for potential biases in the satellite products being evaluated. For a sample of N pairs of matched SST estimates and observations, the “uncorrected” MSE can be expressed as:

$$MSE(SST_{sat}, SST_{obs}) \equiv \frac{1}{N}\sum_{n=1}^{N}(SST_{sat} - SST_{obs})^2. \tag{2}$$

A skill score (SS) that takes into account the tradeoffs between possible biases and the variance can formulated as the loss function:

$$SS(SST_{sat}, SST_{clim}, SST_{obs}) = 1 - [MSE(SST_{sat}, SST_{obs})/MSE(SST_{clim}, SST_{obs})]. \quad (3)$$

Because the climatological reference can be defined, at least hypothetically, based on a sample of observations from the experimental period (Murray, 1988), we use the mean of the collocated UpTempO temperatures as our climatological reference, i.e., $SST_{clim} = \overline{SST}_{obs}$. From (2) it follows that $MSE(SST_{clim}, SST_{obs}) = \sigma^2_{obs}$. Adding and subtracting $\overline{SST}_{sat}$ and $\overline{SST}_{obs}$ within the parentheses on the RHS of (3) and expanding the binomial formula, yields

$$MSE(SST_{sat}, SST_{obs}) = RMS'^2 + (\overline{SST}_{sat} - \overline{SST}_{obs})^2. \quad (4)$$

Substituting $MSE(SST_{clim}, SST_{obs})$ and (4) into (3), it follows that

$$SS(SST_{sat}, \overline{SST}_{obs}, SST_{obs}) = 1 - \widehat{RMS'}^2 - (\overline{SST}_{L4} - \overline{SST}_{obs})^2/\sigma^2_{obs}, \quad (5)$$

where the second term on the RHS of (5) is the square of the normalized centered RMS used in constructing the Taylor diagrams; the last term, the square of the mean error normalized by the variance of the observations, is a non-dimensional measure of the overall bias in the SST estimates. This term vanishes only when the satellite estimates are unbiased. Thus, $SS = 1$ (perfect score) when $\widehat{RMS}'^2 = 0$ and $\overline{SST}_{sat} = \overline{SST}_{obs}$. Since the latter two terms in (5) are nonnegative and are preceded by a negative sign, the skill of the SST products tends to decrease as both of these terms increase. Furthermore, because there is no upper bound on the growth of the quadratic terms in (5), negative skill scores can, and will, occur when the analyses have poor or no skill (*SS* is defined in the interval $(-\infty, 1]$). This decomposition implies that $\widehat{RMS}'^2$ is in itself a sort of idealized skill score, attainable when biases are eliminated. In summary, the Taylor diagram and the TS score can be viewed as measures of potential skill in the absence of any bias (systematic errors of mean bias and standard deviation are intrinsically removed in their computations), while the *SS* can be viewed as a measure of actual skill since it incorporates biases (it allows more spread relative to the observations).

3.3 Thresholds for Product Performance Classification

We will categorize the high-latitude performance of the SST products by ranking the scores from Equations (1) and (5), into three discrete categories: excellent, good, and poorly skilled. To do so, we need to establish some basic classification rules for mapping the skill scores into these discrete categories. This is commonly done via a threshold choice method. Even though there are many methods for determining decision thresholds (Hernández-Orallo et al., 2012), we opted for the simplest of all, which is a score-fixed threshold classifier; that is, we use two predefined score thresholds, such that the satellite products are assigned into one of the three categories mentioned above if their score is within the limits established by the corresponding thresholds. Because in reality there is no perfect separation between categories, the threshold, once fixed, can dramatically impact the performance ranking of the products being compared.

A threshold, T, is usually determined by estimating the cost incurred in misclassifying a product, and setting the threshold to the value that minimizes the expected loss over different conditions ($L(t)$). Hernández-Orallo et al. (2012) showed that when dealing with fixed-score thresholds, the accuracy is the performance metric that minimizes the expected misclassification losses. In this study, we fix the score thresholds based on the accuracy requirements of operational, real time satellite SST products , i.e.,

$$T = L(t) = 1 - \sigma^2. \quad (6)$$

where the standard deviation, σ, is the adopted GHRSST convention for SST product accuracy. GHRSST has established that for global open ocean products, the user accuracy requirement is $\sigma < 0.4$K, with a tighter user requirement ($\sigma < 0.3$K) for coastal and high-resolution (<2 km) products (Donlon et al., 2007). Satellite SST L2 products, however, display higher levels of error in the Arctic Ocean with Metop-A and AVHRR-GAC SSTs showing standard deviations between 0.4 and 0.5°C and MODIS and AMSR-E between 0.5°C and 0.8°C (Hoyer et al., 2012). In a different study, Martin et al. (2012) found that all the

L4 products considered here have σ < 0.7K globally, with the GMPE, CMC, GAMSSA, K10, and OSTIA having σ < 0.5K.

Given the current achievable accuracies of satellite SST products, we assume that σ between 0.5K – 0.7K is a reasonable, if somewhat conservative, absolute accuracy for high-latitude L4 products. From Equation (6) it follows that the thresholds for σ = 0.5K and 0.7K are T = 0.75 and T = 0.51 (labeled T_{75} and T_{51}), respectively. Note that the equation (6) is independent of score and thus T_{75} and T_{51} will be applied to the two score metrics tested here. In summary, L4 products with skill scores in the (0.75, 1.0] interval will be classified as having excellent high-latitude performance; products with scores in the [0.51, 0.75) interval will be grouped in the good performance category, and products with scores < 0.51 will be labeled as having poor performance under our set of operating conditions.

4. Methodology

4.1 Data Grouping

For analysis purposes, the data are divided into three groupings. Among the buoys deployed during the summer of 2012, Louis 2012-03 is singled out, since this buoy overlapped with a rare weather event and strong SST gradients, and thus, it is assumed that it was operating under “extreme conditions.” Louis 2012-04, Louis 2012-05, and Healy 2012-07, on the other hand, were deployed later in the summer season (September 4, 5, 10 respectively) in cold waters farther offshore (North of 76N, a more quiescent SST environment), and propagated in a cyclonic direction (opposite to Louis 2012-01) toward the interior of the Arctic Basin. The latter buoys are grouped together under the “cold northerly waters” category. The 2013 buoys remained closer to coast, and drifted westward toward the Chukchi Sea, displaying similar conditions to those sampled by Louis 2012-03 after the extreme weather event dissipated; hence, the latter portion of Louis 2012-03 and the 2013 buoys are grouped together under the “coastal waters” category.

4.2 Collocation Criteria

The first step is to construct a matched set of the buoy observations and corresponding satellite SST products. Collocation of the buoy data with the gridded analyses is straightforward. Buoy observations are simply compared with the values for the grid cells containing the buoy position on that day. Multiple buoy observations on a given day were all matched independently with the daily analysis. Given the gaps in the L3 products, additional steps were taken. Matchups with the AVHRR LAC data were constructed for satellite observations agreeing within 10 km of the buoy position. Since the WindSat data are provided in separate ascending and descending grids, available observations for the grid cell containing the buoy for the orbital segment closest in time to the buoy measurement were used. This implies a maximum temporal separation of approximately 12 hours.

4.3 Implementation

Collocated satellite-buoy SST pairs were segregated using different UpTempO buoy combinations as mentioned above, and normalized Taylor diagrams were generated for each of the classifications. To facilitate labeling in the diagrams, each of the SST products was given an abbreviated name consisting of the first two letters of the product's name (with the exception of WindSat, abbreviated WS, and K10 and LAC which stayed the same). The standard deviation of the UpTempO SSTs is represented in the normalized diagrams by the purple dashed arc depicting $\hat{\sigma}_{obs} = 1$ and corresponding purple dot, labeled "observed," at unit distance from the origin along the x-axis. Additional isolines for σ are drawn at 1°C-intervals as continuous arcs in black. Isolines for the $\widehat{RMS}'$ are depicted every 0.25°C, as concentric circles in green, centered on the observations; radial lines in blue, labeled according to the cosine of the angle made with the abscissa, correspond to 0.1 increments in correlation. Since σ_{obs}, σ_{sat} and R*MS'* cannot be retrieved from normalized Taylor diagrams, each diagram has an associated table that includes the sample standard deviations of the collocated SST products and corresponding observations, as well as the mean bias, the standard deviation (STDEV) and the RMSE of the paired satellite – buoy differences. Numerical values of the skill scores (Equations 5 and 9) for each of the SST products are also included in these tables.

5. Results

5.1 Extreme Weather Event – Strong Gradients: Louis 2012-03

The first buoy deployed during the MIZOPEX summers, Louis 2012-03, experienced unique conditions. A temporal SST animation (not shown) indicates that it drifted along a temperature gradient, not far from the coast, for several weeks following its deployment. The spiral at the end of the track in Figure 2, also indicates that the buoy became trapped in a strong anti-cyclonic circulation that eventually coincided with its demise. The high spatial variability near the thermal gradient is problematic for satellite SST product comparisons, since buoys make point measurements throughout the day, whereas satellite SST estimates are representative of spatial averages over much larger areas. In the case of the L4, the estimates are expected to be smoother and coarser than non-analyzed SSTs. The grid resolutions specified in Table 1 define the smallest possible SST features that can be resolved by the different L4 products. Furthermore, oceanic mesoscale fronts promote convection, and hence are associated with clouds, which in turn, hamper the IR SST retrievals. Martin et al. (2012) have shown that differences among L4 products tend to be accentuated near coastal and strong gradient regions.

To further complicate matters, the Louis 2012-03 deployment coincided with the appearance of a rare and very powerful summer storm over the Arctic Ocean. The storm, dubbed the Great Arctic Cyclone of 2012 (GAC-2012, Simmonds and Rudeva, 2012), was first identified over northern Siberia on 2 August 2012. It then crossed into the Arctic basin, intensified off the coast of Alaska on 6 August 2012, and then tracked into the center of the Arctic basin before slowly dissipating over the next several days (14 August 2012, day of year 226). Louis 2012-03 was deployed on August 7 off the coast of Alaska (72.6N, 144.64W) at a time when the storm reached its greatest size and depth. As of summer 2016, this was the most extreme summer storm on record since satellite observations of polar orbiters began in 1979, and it is believed that the churning action of the cyclone contributed significantly to the rapid ice melt observed during August 2012 (Simmonds and Rudeva, 2012; Zhang et al., 2013).

The thermal structure of the ocean surface sampled by Louis 2012-03 (Figure 3) is characterized by a rapid cooling of the ocean surface from ~8.5°C to 1°C during the first month of the deployment, followed by a more gradual cooling for the remainder of its life. Note that the drop in SST during the storm is about 4.5°C, possibly due to enhanced mixing associated with the storm-generated winds. Time series of the satellite SST products matched to Louis 2012-01 measurements (**Figure 4**) show extreme differences among the L4 products for the early part of the deployment (Figure 4a), but after about day of year (DOY) 260 the various time series quickly converge, indicating renewed agreement among the L4 products. Corresponding comparisons with the L3 products (Figure 4b) show good overall agreement with Louis 2012-03, although for temperatures below 2-3 °C, the WindSat SSTs appear to have a warm bias relative to the buoy. Differences in L4 products are further emphasized when comparing their corresponding images for a randomly chosen day (DOY 226, **Figure 5**) within the period with the greatest discrepancies. Clearly, significant differences exist for the Beaufort Sea in terms of SST amplitude, variability, and ice mask coverage, suggesting that some of these products are ill-equipped to deal with the harsh conditions encountered by Louis 2012-03 during the first month of its deployment. This clearly motivates determining which, if any, of the analyses accurately represent the actual conditions.

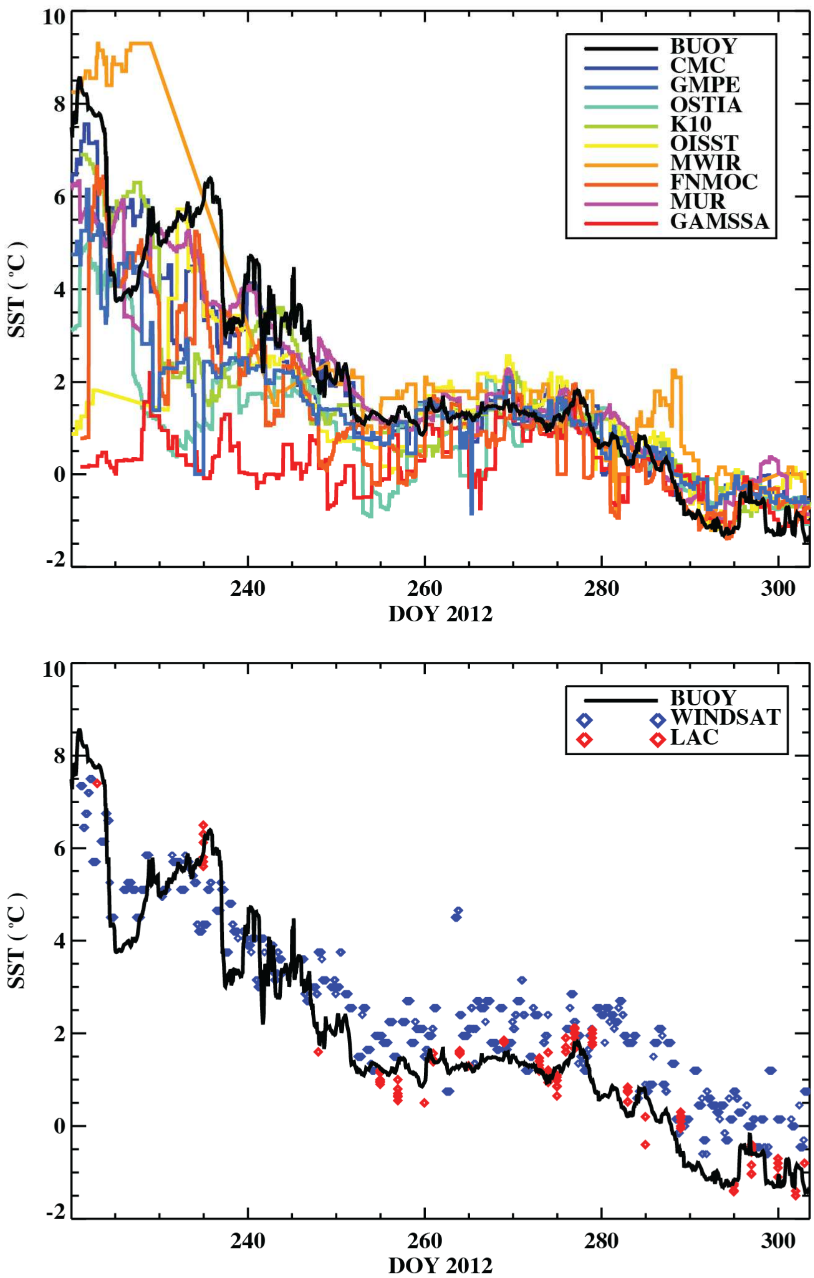

Figure 4. Time series of the SST products compared to the observations from the Louis 2012-03 buoy. The upper panel compares the various L4 products while the lower panel displays the available single sensor retrievals.

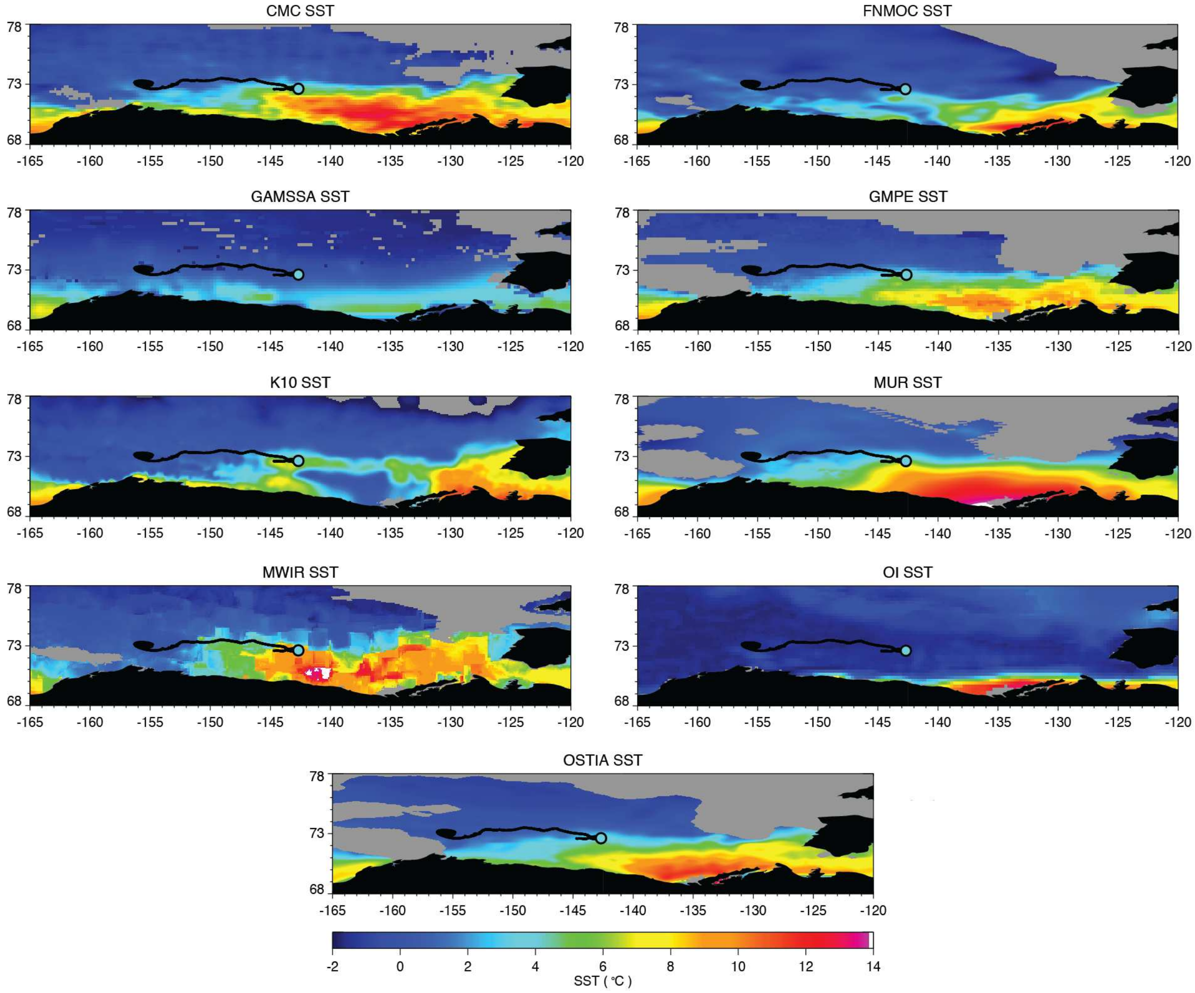

Figure 5. Comparison of the selected L4 SST analyses in the Beaufort Sea on 13 August, 2012 (DOY 226). The color scale has been fixed to facilitate comparisons between products, and the trajectory of Louis 2012-03 has been plotted over the images with the circle showing the position of the buoy for that particular day and the color indicating the corresponding buoy temperature. The gray areas indicate that the respective ice mask has been applied, if available. Note that no ice mask is shown in the OISST since the ice and water masks were mistakenly inverted during this period and the buoy location was inaccurately flagged as ice covered. The anomalously low OISST temperature at the buoy location is a result of the improper masking.

Because of the two different thermal regimes before and after DOY 260, and because a climate event of the proportion of GAC-2012 might have generated oceanic variability that could have prevented the analyzed products from agreeing with the buoy observations, we subsampled the Louis 2012-03 measurements into two datasets for the period prior/post DOY 260. The normalized Taylor diagrams showing the performance of the satellite SST products relative to the Louis 2012-03 measurements at 2.5-m depth for these two periods are presented in **Figure 6** and associated dimensional statistics are summarized in **Tables 3 and 4**, respectively. The σ_{obs} used in the normalization of Figure 6 are given in the Tables. These diagrams do suggest two very different regimes, as reflected by the re-arrangements of the dots representing the satellite products, with a wide scatter in the radial direction for the period overlapping the storm (Figure 6a), and a much closer clustering for the remaining of the melting season (Figure 6b).

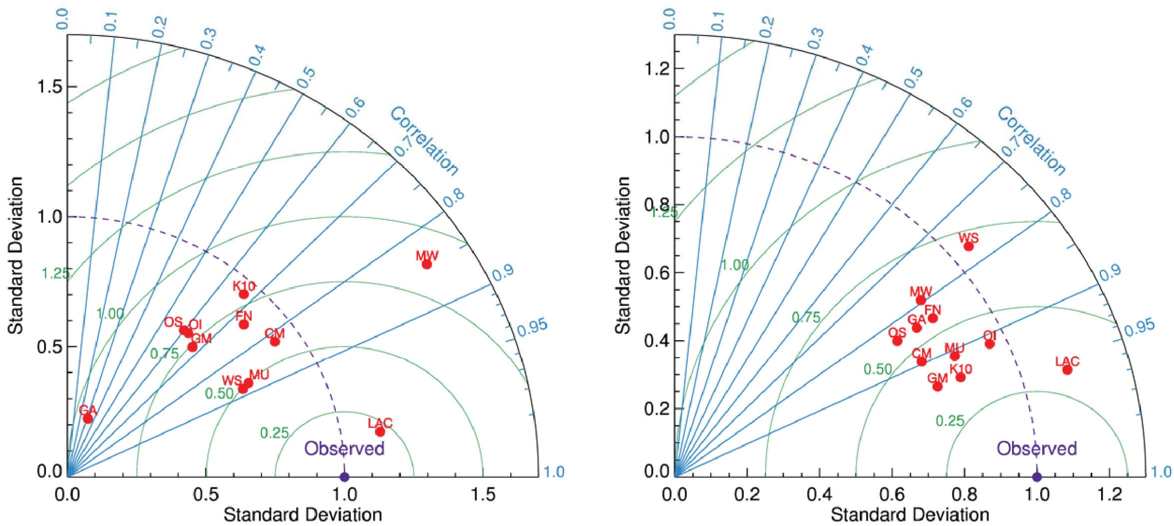

Figure 6. Normalized Taylor diagrams showing differences between matched SST from the UpTempO buoy Louis 2012-03 2.5 m thermistor and eleven satellite SST products considering matchups before DOY 260 (left) and after DOY 260 (right).

Table 3. Louis 2012-03 statistics for the period before DOY 260. All quantities are as defined in the text.

SST Product	No. Pts	σ_{obs} (K)	σ_{sat} (K)	Bias (K)	STDEV (K)	RMS (K)	ρ	SS	TS
CMC	935	1.98	1.80	-0.57	1.14	1.28	0.82	0.58	0.71
FNMOC	934	2.03	1.76	-1.58	1.40	2.11	0.74	-0.08	0.56
GAMSSA	934	2.03	0.48	-3.64	1.93	4.12	0.32	-3.12	0.00
GMPE	935	1.98	1.33	-1.51	1.47	2.11	0.67	-0.13	0.29
K10	934	2.03	1.92	-1.10	1.60	1.94	0.67	0.08	0.51
MUR	935	1.98	1.48	-0.41	0.99	1.07	0.88	0.71	0.59
MWIR	479	2.35	3.60	1.34	2.04	2.44	0.85	-0.08	0.40
OISST	767	2.01	1.42	-1.40	1.59	2.12	0.62	-0.11	0.29
OSTIA	935	1.98	1.39	-2.12	1.60	2.65	0.60	-0.80	0.27
LAC	92	2.02	2.31	-0.30	0.44	0.53	0.99	0.93	0.97
WindSat	715	2.14	1.53	0.06	1.06	1.07	0.88	0.75	0.55

Table 4. Louis 2012-03 statistics for the period after DOY 260. All quantities are as defined in the text.

SST Product	No. Pts	σ_{obs} (K)	σ_{sat} (K)	Bias (K)	STDEV (K)	RMS (K)	ρ	SS	TS
CMC	1045	1.05	0.80	0.22	0.49	0.53	0.90	0.74	0.65
FNMOC	1042	1.02	0.87	-0.18	0.56	0.59	0.83	0.67	0.68
GAMSSA	1045	1.02	0.82	-0.26	0.57	0.62	0.83	0.63	0.62
GMPE	1045	1.05	0.81	0.21	0.40	0.45	0.94	0.82	0.73
K10	1045	1.02	0.86	0.12	0.37	0.39	0.94	0.86	0.84
MUR	1045	1.05	0.89	0.28	0.44	0.52	0.91	0.75	0.79
MWIR	949	1.02	0.87	0.56	0.63	0.84	0.79	0.33	0.62
OISST	1042	1.05	1.00	0.37	0.43	0.57	0.91	0.71	0.88
OSTIA	1045	1.05	0.77	0.07	0.58	0.59	0.84	0.69	0.53
LAC	375	0.99	1.12	0.09	0.32	0.33	0.96	0.89	0.93
WindSat	850	0.99	1.05	0.98	0.70	1.21	0.77	-0.47	0.64

During the first period (Figure 6a), the satellite SST product in closest agreement with the observations by far is the LAC (L3) SSTs with $\widehat{RMS}'_{LAC} = 0.22$ $\hat{\sigma}_{LAC} = 1.14$, and ρ_{LAC} = 0.99. The second and third closest in terms of low RMSE and high correlation are the L3 WindSat and the L4 MUR. Both of these products lie on the 0.5 $\widehat{RMS}'$-arc and the 0.88 ρ-radius, but are farther from the observed variance ($\hat{\sigma}_{WS} = 0.74$ and $\hat{\sigma}_{MUR} = 0.78$). It is expected that, when available, the lower-level processing products should outperform the L4 products, particularly in the presence of strong gradients, as the OI analysis system is in itself a spatial and temporal smoothing filter, damping some of the natural

SST variability. The CMC is closer to the arc for $\hat{\sigma}_{obs} = 1$ ($\hat{\sigma}_{CMC} = 0.91$), but has $\widehat{RMS}' > 0.5$. Although MUR has better overall agreement with the buoy than CMC (smaller RMSE), the fact that $\hat{\sigma}_{CMC}$ is closer to the $\hat{\sigma}_{obs}$-arc, suggests that, at least in terms of the analyzed SST amplitude variability matching the observed, the CMC does better than MUR. Which of these two products is appraised over the other depends on the features valued by the different scoring systems. Of the remaining products, K10 and FNMOC lie closer to the arc for $\hat{\sigma}_{obs} = 1$, whereas GMPE, OSTIA, and the OI align with the arc for $\hat{\sigma}_{obs} = 0.7$. The positioning of the OI, in particular, follows after careful screening of the data for the period corresponding to August 10 – August 16, 2012, when the ice and water masks for the Arctic region were inverted in the OISST product. Failing to remove data from this period results in further degradation of the OISST statistics. The products whose $\hat{\sigma}$ are farthest apart from $\hat{\sigma}_{obs} = 1$ are GAMSSA and the MWIR. These products display a wide range of SST amplitudes, with GAMSSA being much smoother ($\hat{\sigma}_{GAMSSA} = 0.24$) and the MWIR being much noisier ($\hat{\sigma}_{MWIR} = 1.53$) than the other L4s. Note also that their time series exhibit the largest departures from the buoy during the early portion of Figure 4. Since the spread in the radial direction gives an indication of the degree to which temporal and spatial SST variability is affecting the SST amplitudes, the uncertainty attributable to variability appears to be significant for these two products. Our result supports the findings of Reynolds and Chelton (2010) which found that, since the MWIR attempts to resolve very small SST features based on the 1-km MODIS (MODerate Resolution Imaging Spectroradiometer) SSTs, when the high-resolution IR data is missing or the coverage is reduced due to persistent cloud cover over multiple days, as was likely the case during the storm/proximity to a thermal front, insufficient high resolution IR data results in small-scale noise. It is noteworthy that the MWIR is highly correlated with the observations (ρ_{MWIR} = 0.85) despite having too much variability.

For this period along the strong SST gradient, the skill scores for the L4 comparisons are generally quite low, confirming once again how ill-equipped the L4s are for extreme conditions. Sorting the *TS* scores (Table 3) in descending order of skill (decreasing score magnitude), results in: $\mathrm{TS}_{LAC} \ggg \mathrm{TS}_{CMC} > \mathrm{TS}_{MUR} > \mathrm{TS}_{FNMOC} > \mathrm{TS}_{WS} > \mathrm{TS}_{K10} \geqslant \mathrm{TS}_{MWIR} > \mathrm{TS}_{GMPE} > \mathrm{TS}_{OI} > \mathrm{TS}_{OSTIA} > \mathrm{TS}_{GAMSSA}$. Similarly, the sorted list for the *SS* scores (Table 3) indicates: $\mathrm{SS}_{LAC} > \mathrm{SS}_{WS} \ggg \mathrm{SS}_{MUR} > \mathrm{SS}_{CMC} \geqslant \mathrm{SS}_{K10} > \mathrm{SS}_{FNMOC} > \mathrm{SS}_{MWIR} > \mathrm{SS}_{OI} > \mathrm{SS}_{GMPE} > \mathrm{SS}_{OSTIA} > \mathrm{SS}_{GAMSSA}$. The symbols $\ggg$and $\geqslant$ indicate the relative positions of the fixed thresholds for T_{75} and T_{51}, delimiting the discrete skill categories of excellent, good or poor product performance. Only the L3 products can be classified as excellent during this challenging period. The high-resolution LAC data, when available, clearly can capture the gradient, but the blending and reduced resolution of the analyses tends to miss or smear the gradient. MUR, having the highest resolution of the L4 products considered, does among the best in this case. It can be seen that CMC and MUR, as well as OISST and GMPE, trade positions in the two scoring systems, since the TS rewards products with $\hat{\sigma}$ closer to $\hat{\sigma}_{obs}$, whereas SS rewards products with small $\widehat{RMS}'$ and small bias. The dimensional statistics in Table 3 indicate significant absolute biases (> 1.0°C) with respect to the LOUIS 2012-03 for all satellite products (also evident in Figure 3), except WindSat (0.06°C), LAC (-0.30°C), MUR (-0.41°C) and CMC (-0.57°C). The K10 and the FNMOC, which fared well in terms of the TS classifier, are severely penalized due to their large biases, and downgraded to poor skill in the SS scale.

During this portion of the Louis 2012-03 deployment it is difficult to assert whether differences in skill were solely sampling errors associated with the proximity to the strong temperature gradients or also contained retrieval errors, perhaps associated with the extreme storm conditions. Since Louis 2012-03 followed a front for much of its trajectory, it is plausible to ascribe a fraction of the high RMSE values in Table 3 to sampling variability. In and around the front, the RMS would have been particularly sensitive to variations in the analyzed SST amplitudes. Note that in this comparison, we did not attempt to interpolate the different satellite products to a common grid; instead we opted for working with the products in their native spatial resolution. It is also possible that the IR SST retrievals were limited by the presence of clouds associated with the storm and strong SST fronts, as suggested by the small number of AVHRR LAC matchups obtained before DOY 260 (Figure 4b). The WindSat – buoy matchups, on the other hand, were not affected by clouds, giving an apparent advantage to CMC and MUR (two of just three L4s that ingested WindSat SSTs at this time; see Table 2). The MWIR, despite ingesting WindSat, did not generate many SST retrievals for this period, likely due to the fact that it uses the MODIS SST cloud mask in the Arctic, and being an IR sensor, the MODIS instrument has limited coverage under cloudy conditions.

For the quiescent period following DOY 260, the results change dramatically and the L4 satellite products generally show better agreement with the buoy. The L3 SST products in the Taylor diagram (Figure 6b) align in an arc to the right of $\hat{\sigma}_{obs} = 1$ at a radial distance of $\hat{\sigma}_{L3} = 1.13$, whereas the L4 products cluster to the left, at a radial distance of ~ 0.80, indicating an overall agreement in L4 product performance for the second period. Corresponding statistics (Table 4) also indicate that the sampling biases, although still present, are significantly smaller (< 0.5°C) for all but the MWIR and WindSat products.

The convergence in performance among L4 product is substantiated by the *TS* scoring system, which finds no poorly skilled products for the latter half of the Louis 2012-03 deployment. Sorted *TS* scores in descending order of skill (Table 4), indicate that: $\mathrm{TS}_{LAC} > \mathrm{TS}_{OI} > \mathrm{TS}_{K10} > \mathrm{TS}_{MUR} \ggg \mathrm{TS}_{GMPE} > \mathrm{TS}_{FNMOC} > \mathrm{TS}_{CMC} > \mathrm{TS}_{WS} > \mathrm{TS}_{MWIR} > \mathrm{TS}_{GAMSSA} > \mathrm{TS}_{OSTIA}$, whereas the sorted *SS* scores result in: $\mathrm{SS}_{LAC} > \mathrm{SS}_{K10} > \mathrm{SS}_{GMPE} > \mathrm{SS}_{MUR} \ggg \mathrm{SS}_{CMC} > \mathrm{SS}_{OI} > \mathrm{SS}_{OSTIA} > \mathrm{SS}_{FNMOC} > \mathrm{SS}_{GAMSSA} \geqslant \mathrm{SS}_{MWIR} > \mathrm{SS}_{WS}$. Under quiescent conditions, the satellite products that excel at reproducing the observations in the *TS* classifier are LAC, OISST, K10, and MUR. Improvements are especially remarkable for the OISST with the closest variance to the observations ($\hat{\sigma}_{OISST} = 0.95$). However, despite achieving the correct SST variability and being ranked second best, the OISST is demoted one category by the *SS* classifier because of its larger bias (0.37°C) relative to its counterparts at the top of the scale. The GMPE with its small $\widehat{RMS}'$ and bias, is promoted by the *SS* classifier in lieu of the OISST, and placed between the K10 and the MUR. Given the homogeneity in L4 performance under these quiescent conditions, it is no surprise to find that the median ensemble is among the more skillful products.

It is noteworthy that WindSat drops from second best in the first period according to the *SS* scoring to last place in the second period. The good agreement between the MW SSTs from WindSat and the Louis 2012-03 surface temperatures during the first half (Figure 6a and *TS* in Table 3) suggests that, under warmer temperatures, the additional coverage enabled by MW SSTs in the presence of clouds that obscure the IR is highly beneficial in the objective analyses. After DOY 260, however, Figure 4b shows the appearance of a warm bias, also evident in Table 4, with the L3 WindSat SSTs overestimating the buoy temperatures by ~1°C. This contrasts with the fact that Table 3 indicates zero-bias for WindSat during the first period. This 1°C-bias persists in all other UpTempO buoy combinations explored hereafter where the prevailing SSTs were below 2°C. The WindSat bias is potentially related to cold SSTs, as a high-latitude bias has also been reported for MW AMSRE SSTs relative to ship-based observations in the Southern Ocean (Dong et al., 2006), although of a lesser magnitude (ascending/descending: 0.42/23 K in the summer and -0.21/-0.42 K in the winter).

5.2. Cold Northerly Waters

Louis 2012-04, Louis 2012-05, and Healy 2012-07 were deployed at the beginning of September 2012 in the cold waters further north from the Alaska coast. As illustrated in Figure 2, these buoys drifted counter-clock wise toward the interior of the Canada Basin. The flow pattern suggests that the circulation of the Beaufort Gyre had reversed as a result of the low pressure system that persisted for several days in August 2012. The buoy temperatures are characterized by a narrow dynamic range with initial temperatures below 0°C, followed by a gradual cooling (at nearly identical cooling rates) until the start of fall freeze-up (**Figure 7**).

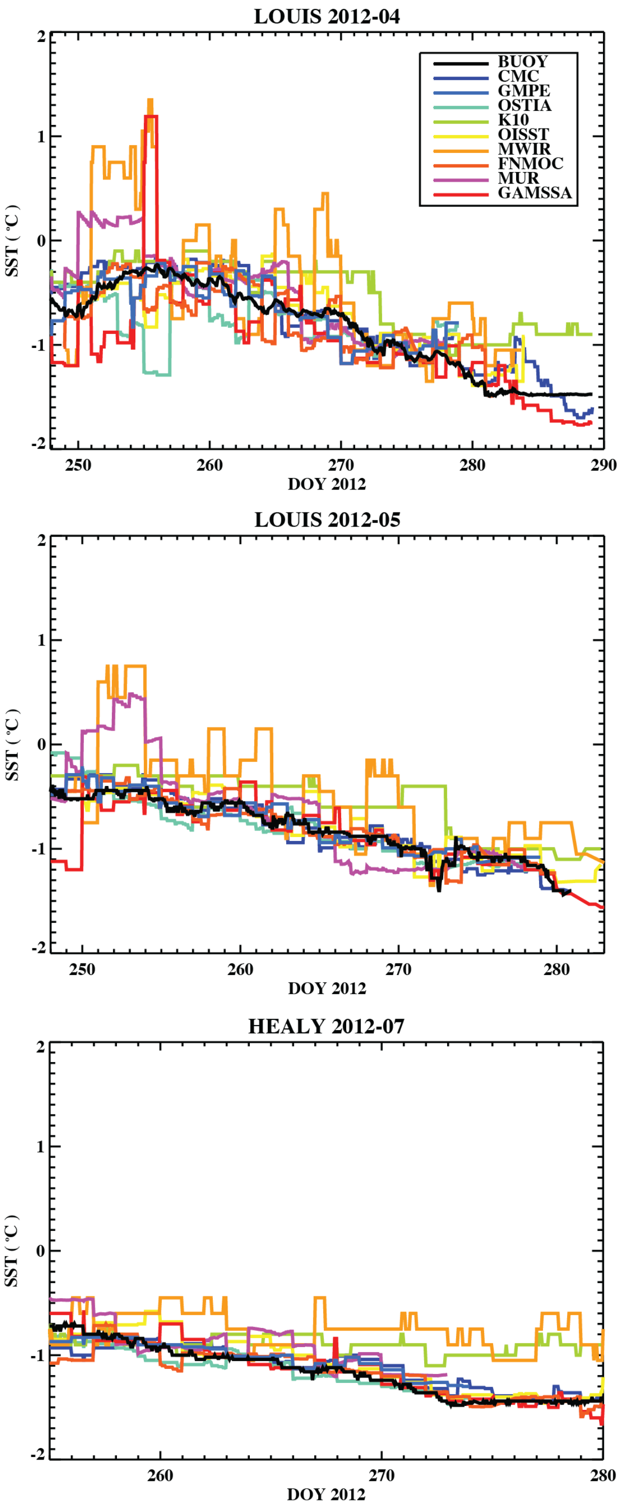

Figure 7. Time series of the selected SST products compared with the remaining UpTempO buoys deployed in 2012. These are considered the cold northerly buoys. Separate panels are shown for each buoy while the color traces correspond to the different SST products. The buoy observations are always shown with the black trace.

The location of these three buoys (Figure 2) is especially interesting because they allow us to look at the impact that the different ice masks is having on the L4 SST products. Being farther north, these buoys were closer to the main ice pack and experienced refreeze earlier than the buoys closer to the coast. Maps for all the products for a randomly chosen day within this period (DOY 256, 2012) are shown in **Figure 8**. Buoy positions on this particular day are displayed in Figure 8 as circles, color-coded by the buoy measured temperature. Although the maps again show pronounced differences among the satellite products, the buoys are located far off the region with the largest discrepancies, which happens to be near the coast. The GAMSSA product appears to have experienced some difficulties with the ice mask around this time, as indicated by the corresponding map in Figure 8, where the region that should have been flagged as ice is, in fact, shown as water. This issue was corrected after DOY 258.

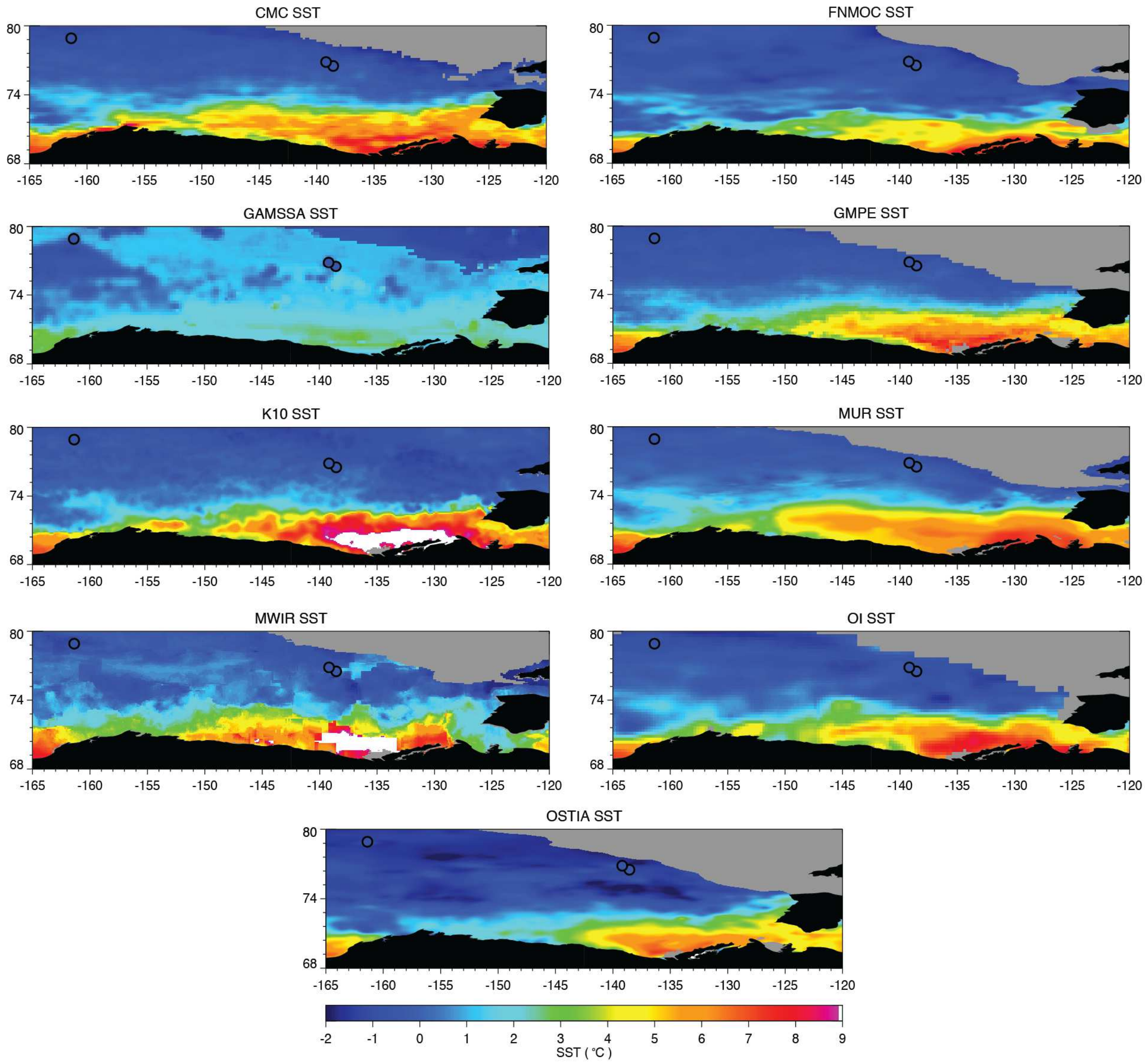

Figure 8. Graphical comparison of the selected L4 analyses on DOY 256 of 2012 corresponding to the period sampled by the cold northern buoys. The separate buoy positions are indicated with the enclosed circles with the color corresponding to the buoy temperature. The buoys, in order from west to east, are Healy 2012-07, Louis 2012-05, and Louis 2012-04. The ice masks are again indicated by the gray regions. The white areas in the K10 and MWIR analyses correspond to where the temperature exceeds the maximum value on the color scale.

In order to explore the impact that a less conservative ice mask (one that delays freezing) had on the analyses, we considered two Taylor diagrams: one based on the whole extent of each of the products' matchup time series, and a second one in which the most conservative ice mask is used to truncate the time series to a common period during which all the products were unambiguously reporting temperatures above the freezing temperature, i.e., a period during which we were highly confident that the SSTs were not being influenced by ice. A byproduct of the truncation is that, whereas the former diagram deals with time series of unequal lengths, the latter one is more balanced in terms of the number of counts. The L3 products are excluded from these comparisons, since the LAC coverage did not extend that far north, and the WindSat SSTs appeared to be decorrelated from the buoys (not shown). This could have been the result of cold absolute temperatures and/or the buoys being within 75-km from the ice, where MW SSTs cannot be retrieved.

The OSI-SAF ice mask is the most conservative of the ice masks used in the L4s considered here (see Table 1), with a much earlier freeze cut-off at the end of the melt season. Of the products using the EUMETSAT OSI-SAF ice mask, the MUR SST matchups defaulted to the freezing temperature on days 278, 279, and 273 for Louis 2012-04, -05, and Healy 2012-07, respectively. OSTIA and GMPE, also using the OSI-SAF ice mask, followed suit one day later. The other L4s, continued to report SSTs for about 10 more days. Freezing-up last were CMC, K10 and GAMSSA. The ice flagging in the UpTempO buoys agreed extremely well with the OSI-SAF ice mask indicating possible ice effects on days 279, 280, and 274 for Louis 2012-04, -05, and Healy 2012-07, respectively. It is important to emphasize that while the UpTempO ice indicator is based on the National Snow and Ice Data Center (NSIDC) SIC, it uses a lower ice concentration (SIC ≥ 0.15) and a -1.2°C threshold on the uppermost thermistor to indicate that the buoy is in/near ice. Thus, the extent of the time series of the MUR–UpTempO matchups determined the truncation dates for all the other products, as this period was considered ice-free in both, the satellite and the UpTempO records.

The normalized Taylor diagram for the period free of ice effects is shown in **Figure 9a** and corresponding statistics are shown in **Table 5**. A visual inspection of this diagram indicates that the majority of the products tightly align with the arc for $\hat{\sigma}_{obs} = 1$, with a spread in azimuthal direction characterized by an increase in $\widehat{RMS}'$ (decrease in ρ) from 0.42 (0.91) for GMPE to 0.75 (0.70) for OSTIA.

The two products that do not conform with the others are the MUR and the MWIR. Their location to the right of the $\hat{\sigma}_{obs}$-arc, indicates substantially larger $\hat{\sigma}$ and $\widehat{RMS}'$ than the rest. For this region with narrow dynamic SST range, the scores in Table 5 had a wider range than the scores obtained for the latter part of Louis 2012-03 (Table 4), but several products demonstrated good skill. The TS ranking of the L4 products in decreasing order of skill, $\mathrm{TS}_{GMPE} > \mathrm{TS}_{CMC} > \mathrm{TS}_{OISST} > \mathrm{TS}_{FNMOC} \ggg \mathrm{TS}_{K10} > \mathrm{TS}_{GAMSSA} > \mathrm{TS}_{OSTIA} \ggg \mathrm{TS}_{MUR} > \mathrm{TS}_{MWIR}$, agrees with the *SS* ranking except for the placement of the K10, which moves from fifth to seventh place, and the threshold delimiters shifting two positions to the left, leaving GMPE and CMC at the top of the ranking, followed by OISST, FNMOC, and GAMSSA in the intermediate category, and OSTIA, K10, MUR and MWIR at the bottom. The positioning of the GAMSSA product follows after careful removal of data for DOY 255, for which this analysis reported unrealistically warm SSTs relative to Louis 2012-04 (see Figure 7a). Among the products showing poor skill under the sampled conditions, MUR and the MWIR differ from the others mainly in that they are the only ones that assimilate MODIS SSTs in the analysis. We speculate that biases due to residual cloud contamination at the highest quality MODIS L2 SSTs might have negatively impacted these products.

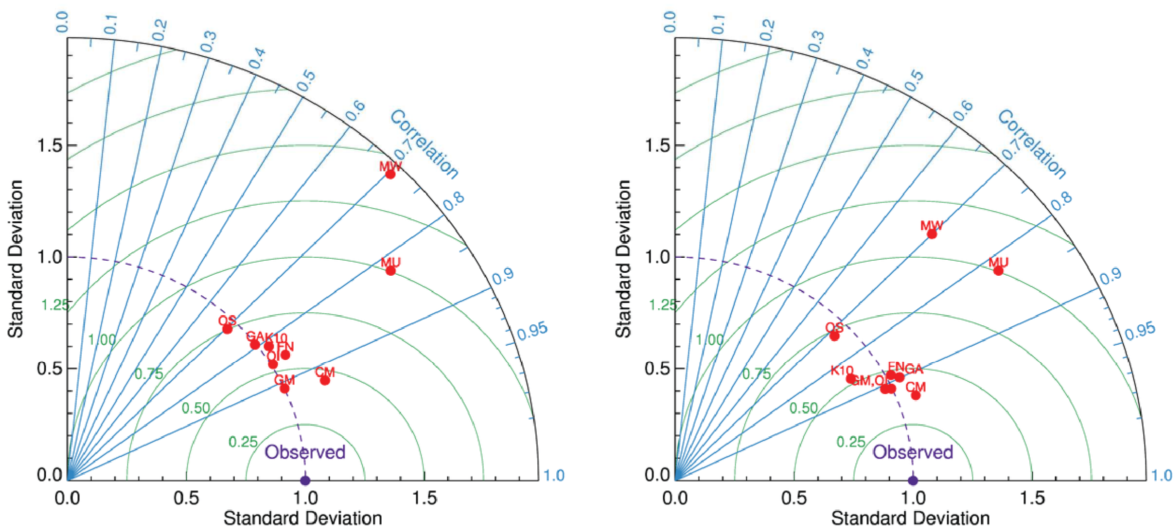

Figure 9. Normalized Taylor diagrams for the combination of the cold northerly buoys including Louis 2012-04, Louis 2012-05 and Healy 2012-07. The left panel shows the results for the common truncated

time series when the OSI-SAF ice mask indicated the region was ice-free while the right panel shows the results for the whole extent of each individual product time series.

Table 5. Statistics for satellite SSTs matched to temperatures from the 2012 northerly buoys after truncating irregular intervals at the start/end of the summer period to eliminate possible ice effects.

L4 Product	No. Pts	σ_{obs} (K)	σ_{sat} (K)	Bias (K)	STDEV (K)	RMS (K)	ρ	SS	TS
CMC	1761	0.29	0.33	0.04	0.13	0.14	0.92	0.77	0.83
FNMOC	1714	0.27	0.30	-0.04	0.16	0.16	0.85	0.65	0.77
GAMSSA	1736	0.28	0.28	-0.04	0.18	0.19	0.79	0.57	0.68
GMPE	1705	0.28	0.28	0.02	0.12	0.12	0.91	0.82	0.89
K10	1761	0.29	0.30	0.24	0.18	0.30	0.86	-0.08	0.72
MUR	1680	0.27	0.45	0.14	0.28	0.31	0.82	-0.27	0.31
MWIR	1761	0.29	0.55	0.35	0.40	0.53	0.70	-2.49	0.14
OISST	1756	0.29	0.29	0.05	0.15	0.16	0.86	0.68	0.79
OSTIA	1660	0.28	0.27	-0.10	0.21	0.23	0.70	0.31	0.56

The normalized Taylor diagram using the whole extent of the time series (**Figure 9b**) differs from the truncated one (Figure 9.a) in that the products that previously aligned with the arc for $\hat{\sigma}_{obs} = 1$ re-emerge clustered around the arc for $\widehat{RMS}' = 0.45$ and the radius for $\rho = 0.90$. The ranking of the L4s based on the TS classifier (**Table 6**) is: $\mathrm{TS}_{CMC} > \mathrm{TS}_{OISST} > \mathrm{TS}_{GMPE} > \mathrm{TS}_{GAMSSA} > \mathrm{TS}_{FNMOC} \ggg \mathrm{TS}_{OSTIA} > \mathrm{TS}_{K10} \gg \mathrm{TS}_{MUR} > \mathrm{TS}_{MWIR}$. Note that with the exception of GAMSSA, which benefited from more data to compensate for the issues it experienced at the beginning of these buoy surveys, the skill categories based on the TS classifier are comprised of the same products in both experiments; i.e., the potential skill of the L4s did not change regardless of the presence of possible ice effects. The *SS* ranking differs from the TS ranking in that K10 and OSTIA get demoted one category because of a large bias and large $\widehat{RMS}'$, respectively. Consequently, there are no products in the intermediate category under the SS classifier when the entire length of the time series is used. The K10 has a bias of 0.29°C relative to the northern UpTempO buoys, the second largest after the MWIR. A bias of the same magnitude (~0.3°C) has been reported by the authors (Castro et al., 2012) for SSTs < 8°C in validation studies of the K10 SSTs using GTS drifting buoys. This warm bias might be a manifestation of the K10 reliance on long-term climatologies that show greater ice extent than what was actually observed in a year with extreme thinning.

Table 6. Statistics for satellite SSTs matched to the temperatures from the 2012 northerly buoys for the entire series of matches.

L4 Product	No. Pts	σ_{obs} (K)	σ_{sat} (K)	Bias (K)	STDEV (K)	RMS (K)	ρ	SS	TS
CMC	2232	0.36	0.39	0.05	0.14	0.15	0.94	0.84	0.91
FNMOC	2015	0.32	0.33	-0.04	0.15	0.16	0.89	0.76	0.84
GAMSSA	2404	0.37	0.38	-0.05	0.17	0.18	0.90	0.77	0.86
GMPE	1776	0.29	0.29	0.03	0.13	0.13	0.91	0.81	0.88
K10	2394	0.37	0.32	0.29	0.19	0.35	0.85	0.10	0.72
MUR	1680	0.27	0.45	0.14	0.28	0.31	0.82	-0.27	0.31
MWIR	2124	0.34	0.53	0.38	0.38	0.54	0.70	-1.42	0.28
OISST	2148	0.35	0.35	0.06	0.15	0.16	0.91	0.79	0.89
OSTIA	1731	0.29	0.27	-0.09	0.21	0.23	0.72	0.38	0.57

Looking at the small differences between the statistics in Tables 5 and 6, it becomes apparent that one statistical measure alone cannot capture the impact that the ice mask is having on SST product performance. In order to see a discernible effect, we need to look at the combined effect of all the statistics at once. This is captured by scores, which show more drastic changes between the two tables. If we ignore the products that rely on the OSI-SAF ice mask (i.e., MUR, OSTIA, GMPE, and MWIR) which should have remained mostly unchanged between experiments (the MWIR is the exception since it continued to report SSTs for 6 additional days) and look at the differences in actual performance (the SS classifier) between Table 5 and Table 6, we find that the scores for all the L4s that use less conservative ice masks improved significantly through the availability of more observations. In fact, by using longer time series, the OISST, FNMOC, and GAMSSA, which had intermediate skills according to the SS scores in Table 5, joined CMC as having excellent skills (Table 6). More data (note that the OISST and FNMOC estimate proxy SSTs until SIC = 1) resulted in smaller $\widehat{RMS}'$ values and better linear fits ($\rho \rightarrow 1$), which in turn produced higher scores that rewarded the products in the cluster closer to the observations. Thus, product performance was not degraded by using less conservative ice masks in this specific case.

5.3. Coastal Buoys

The Arctic summer of 2013 (August and September) saw less ice retreat than in the record year of 2012. The UpTempO buoys deployed in 2013 (Healy 2013-16, Healy 2013-17, Healy 2013-18, and Ukpik 2013-04) remained closer to the coast (see Figure 2) and moved westward toward the Chukchi Sea. The time series for the satellite SSTs matched to the observed temperatures from the 2013 buoys are shown in **Figure 10**. The SST dynamic range is from approximately 4 to -1°C which is rather small from a global perspective, but is larger than for the northern buoys. In fact, the observed temperatures here are warmer than for all other groupings other than the storm/gradient period from Louis 2012-03. A notable difference visible in the time series is that the satellite products tend to show a more constant, gradual cooling while the buoys suggest a more step-wise drop. Maps for a single, arbitrarily chosen day (**Figure 11**, DOY 250, 2013), suggest that once again, the buoys were drifting along thermal fronts, which might explain the choppiness in the buoy measurements, and the smoothing of the SST analyses along frontal boundaries.

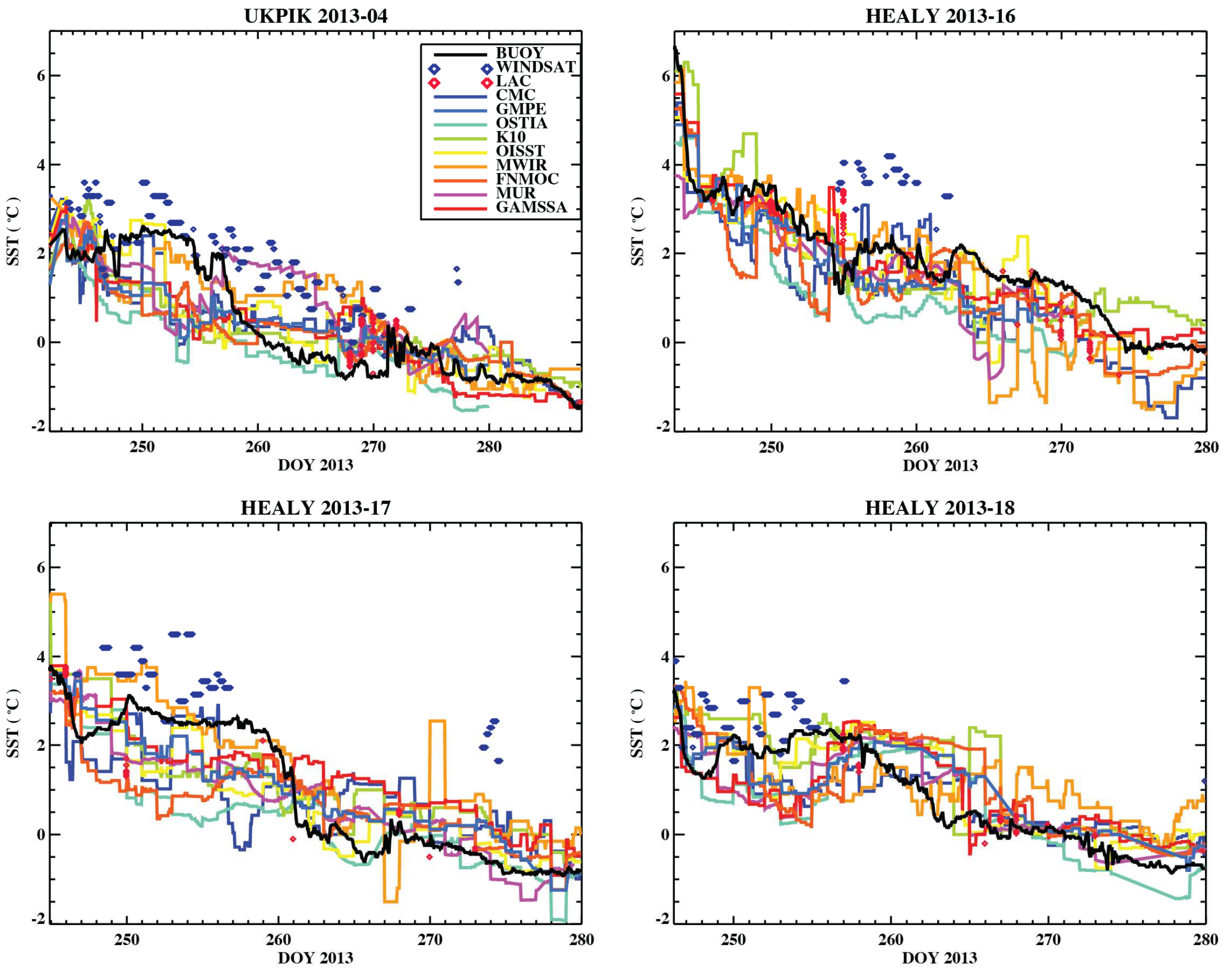

Figure 10. Time series of the selected SST products compared with the 2013 UpTempO buoys termed as coastal buoys. Separate panels are shown for each buoy while the color traces correspond to the different SST products. The buoy observations are always shown with the black trace.

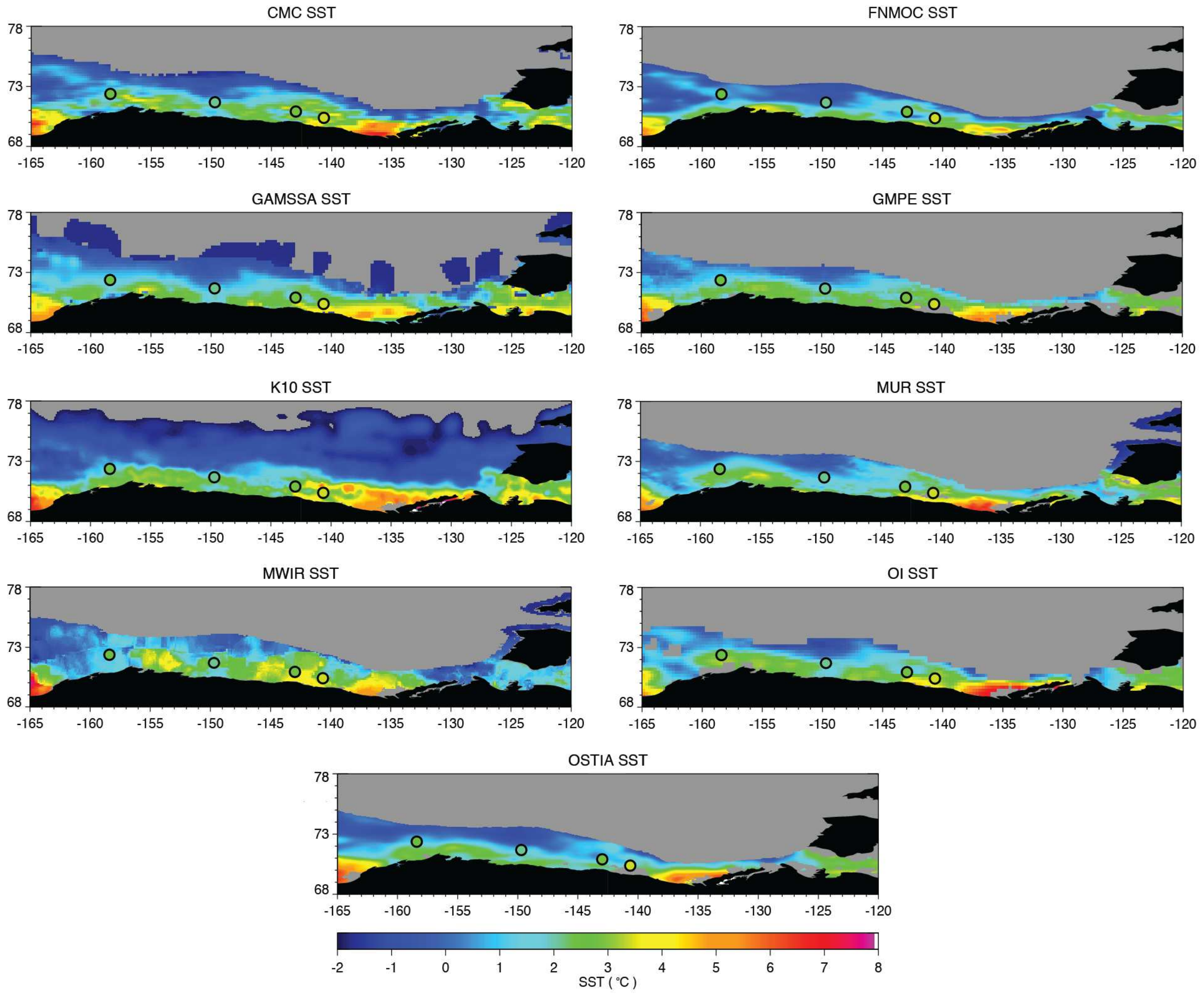

Fig 11. Graphical comparison of the selected L4 analyses on DOY 250 of 2013 corresponding to the period sampled by the coastal buoys. The separate buoy positions are indicated with the enclosed circles with the color corresponding to the buoy temperature. The buoys, in order from west to east, are Ukpik 2013-04, Healy 2013-18, Healy 2013-17, and Healy 2013-16. The ice masks are again indicated by the gray regions.

The normalized Taylor diagram for the coastal buoys, including the latter portion of Louis 2012-03, is shown in **Figure 12** and the corresponding statistics are given in **Table 7**. The diagram shows similar skill for all SST products, with most products clustering in a narrow region of the parameter space delimited by the $\widehat{RMS}'$-arcs for 0.5 and 0.65, the $\hat{\sigma}$-arcs for 0.8 and 1.0, and ρ -radii between 0.8 and 0.9. The L3 SST products have daily SST amplitudes that are in excellent agreement with the observed (they lie on the arc for $\hat{\sigma}_{obs} = 1$), whereas the L4s have slightly smoother amplitudes, as expected.

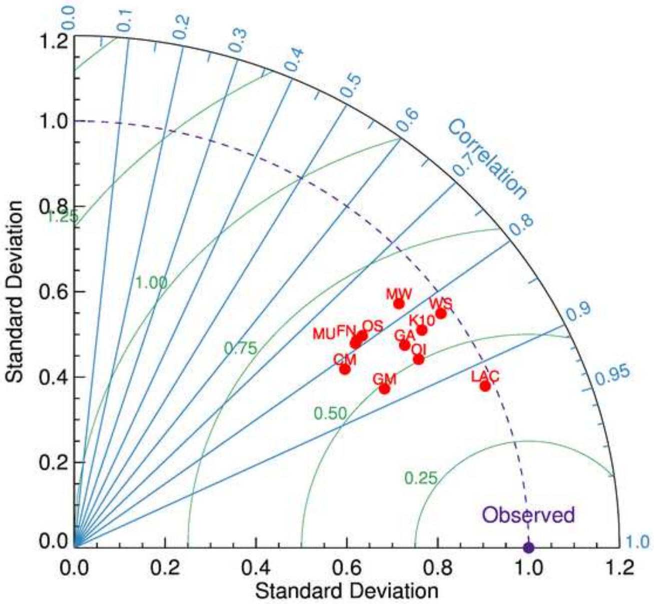

Figure 12. Normalized Taylor diagram for the combination of the coastal buoys including all of the 2013 buoys plus the observations from Louis 2012-03 excluding the period of the storm.

Table 7. Statistics for satellite SSTs matched to temperatures from the 2012–2013 UpTempO coastal buoys.

SST Product	No. Pts	σ_{obs} (K)	σ_{sat} (K)	Bias (K)	STDEV (K)	RMS (K)	ρ	SS	TS
CMC	4215	1.39	1.01	-0.03	0.81	0.81	0.82	0.66	0.50
FNMOC	4211	1.38	1.09	-0.19	0.85	0.87	0.79	0.60	0.55
GAMSSA	4216	1.38	1.20	-0.04	0.76	0.76	0.84	0.70	0.70
GMPE	3990	1.39	1.08	-0.05	0.68	0.68	0.88	0.76	0.65
K10	4215	1.38	1.27	0.09	0.78	0.78	0.83	0.68	0.73
MUR	3979	1.39	1.09	-0.02	0.85	0.85	0.79	0.63	0.54
MWIR	4118	1.38	1.26	0.34	0.88	0.94	0.78	0.53	0.65
OISST	4163	1.39	1.22	0.09	0.70	0.70	0.86	0.74	0.75
OSTIA	4008	1.39	1.12	-0.47	0.86	0.98	0.79	0.51	0.56
LAC	876	1.30	1.27	0.06	0.51	0.51	0.92	0.85	0.91
WindSat	1892	1.22	1.19	1.03	0.71	1.25	0.83	-0.05	0.74

The sorted list of TS scores (Table 7) from high to low is: $\mathrm{TS}_{LAC} > \mathrm{TS}_{OISST} \ggg \mathrm{TS}_{WS} > \mathrm{TS}_{K10} > \mathrm{TS}_{GAMSSA} > \mathrm{TS}_{GMPE} > \mathrm{TS}_{MWIR} > \mathrm{TS}_{OSTIA} > \mathrm{TS}_{FNMOC} > \mathrm{TS}_{MUR} \gg \mathrm{TS}_{CMC}$. Similarly, the sorted list for the SS scores, is: $\mathrm{SS}_{LAC} > \mathrm{SS}_{GMPE} \ggg \mathrm{SS}_{OISST} > \mathrm{SS}_{GAMSSA} > \mathrm{SS}_{K10} > \mathrm{SS}_{CMC} > \mathrm{SS}_{MUR} > \mathrm{SS}_{FNMOC} > \mathrm{SS}_{MWIR} > \mathrm{SS}_{OSTIA} \gg \mathrm{SS}_{WS}$. The L3 AVHRR LAC SSTs, once again, agree best overall with the smallest centered RMS errors, normalized standard deviation closest to the observed, and maximum correlation ($\widehat{RMS}'_{LAC} = 0.39, \hat{\sigma}_{LAC} = 0.98, \text{and } \rho_{LAC} = 0.92$). The majority of the L4s have excellent to good skills in the coastal region around the Beaufort Gyre. Overall, however, the scores are generally lower than for the northern buoys and are also lower than for the latter period of Louis 2012-03. This could be a result of the dynamic range and SST variability. The limited LAC observations coincident with the buoys could also suggest more issues with cloud coverage during this period. Among the best products are the OISST ($\widehat{RMS}'_{OISST} = 0.50, \hat{\sigma}_{OISST} = 0.88, \text{and } \rho_{OISST} = 0.86$), and the GMPE ($\widehat{RMS}'_{GMPE} = 0.49, \hat{\sigma}_{GMPE} = 0.78$ and $\rho_{GMPE} = 0.88$). GAMSSA and K10, which had difficulties in the 2012 comparisons, showed good skill for this case study ($\widehat{RMS}'_{GAMSSA} = 0.55$ and $\widehat{RMS}'_{K10} = 0.56$). Both analyses underwent some changes in January 2013, when they started ingesting WindSat SSTs. The MUR, FNMOC, OSTIA and MWIR have the largest errors with $\widehat{RMS}'$ between 0.6 and 0.65. These numbers, however, are consistent with global statistics. A notable exception is the CMC product. This L4, which was at the top of the rankings for the 2012 buoy combinations, ended up last, according to the TS classifier ($TS_{CMC} = 0.5$, right below T_{51}), due to the large difference in variance relative to the coastal buoys ($\hat{\sigma}_{CMC} = 0.73$ or 73% of σ_{obs}). This places the CMC farthest to the left from the dashed arc in the Taylor diagram, implying that the product is underestimating the SST amplitude variability at these buoy locations. Note, however, that the SS classifier places the product in fifth position among the products with good skills ($\widehat{RMS}'_{CMC} = 0.58$). Given this and the fact that we are not aware of processing changes in the interim period that might have affected the skill of this analysis, we acknowledge that the CMC was perhaps unjustly penalized by the T_{51}, as there is some inherent arbitrariness in how the thresholds used to place products in discrete categories are selected.

5.4. Combined Score

For a final ranking of the L4 products, we looked at three of the groupings analyzed here, as they correspond to different operating conditions/regimes: 1) the "weather" system passing, which included matchups with Louis 2012-03 before DOY 260 (Figure 7b; Table 5); 2) The "northern buoys" (Figure 9b, Table 6); and 3) The "coastal" buoys included in Section 5.3 (Figure 12; Table 7). Note that these categories include mutually exclusive data sets that, when combined, add up to all the matchups. We then considered the average of the TS and SS rankings after removing the L3 products from the sorted lists; i.e., if, after removing LAC and WindSat from both score rankings in the coastal regime, the OISST occupies first and second positions (highest rankings) in the TS and SS rankings, respectively, then the average ranking, $\bar{S}$, for this product is $\bar{S}_{OISST} = 1.5$. The results are shown in **Table 8**. Since there are nine L4 products being evaluated, $\bar{S}$ varies from 1 – 9. In this approach, the lower the value of $\bar{S}$, the better the two skill scores. A natural clustering emerges for each of these regimes in which L4 products with comparable skills end up having similar average ranking. Table 8 indicates that the best products have $\bar{S}$ between 1 and 3.5, the next best group between 4 and 6.5, and the poorest performing products between 7.5 and 9. For instance, the OISST, GMPE, K10, and GAMSSA are all seemingly skillful in the coastal regime ($\bar{S}$ between 1.5 and 3.5), whereas CMC, MUR, FNMOC, OSTIA, and MWIR are equally challenged ($\bar{S}$ between 7.5 and 9.0). An overall ranking followed from averaging the mean position in each of the three regimes (see "Overall" in Table 8). In the overall classification the OISST, CMC, and GMPE, occupy the top three rankings (the most skillful overall), followed by K10, FNMOC, GAMSSA, MUR, OSTIA, and MWIR. In terms of product resolution, Table 8 indicates that, in general, high resolution L4 products (<10 km) constituted the best group for frontal regions and extreme weather conditions; the second best group for the cold Northerly waters, and the poor-performing group for the coastal regions. Coarser resolution (>10 km) L4 products performed best overall for all but the frontal regions.

Table 8. Final ranking of the selected L4 analyses for the individual regimes and overall period. Lower values indicate the best agreement with the UpTempO observations.

	OISST	CMC	GMPE	K10	FNMOC	GAMSSA	MUR	OSTIA	MWIR
Weather	6.5	1.5	6.5	3.5	3.5	9.0	1.5	8.0	5.0
North	2.5	1.0	2.5	6.5	5.0	4.0	8.0	6.5	9.0
Coastal	1.5	7.5	3.0	3.5	7.5	3.5	7.5	8.0	9.0
Overall	3.5	3.7	4.0	4.5	5.3	5.5	5.7	7.5	7.7

6. Conclusions

The Beaufort Sea is an extremely challenging region for SST analyses as evidenced by the very dramatic differences shown in contemporaneous scenes from the individual SST analyses. Despite the use of largely similar satellite input products and analysis procedures (they should be highly correlated), the L4 products exhibit significant differences in the amplitude and the phasing of their spatial patterns. Products found to be the best performing in other open ocean regions are not necessarily the best here. The UpTempO buoys provide a unique, and very valuable, verification data set since they are truly independent from the SST products evaluated in this study, giving great confidence in the results.

Taylor diagrams and skill scores provide very useful tools for comparative evaluation of the different SST analyses. One single statistical measure does not adequately capture all the aspects that might influence the relative skill of the different SST analyses in this challenging environment. Taylor diagrams provide a convenient way to graphically summarize the interplay among three different statistics that gauge different strengths and weaknesses in the products being evaluated. Skill scores allow emphasis of different measures and permit objective relative ranking of the different products.

The products found to be best performing varied with the region and conditions within the Beaufort Sea. Where available, the IR L3 AVHRR LAC data outperformed all of the analyses due largely to its high spatial resolution. While AMSR-E data were not available at the time of the comparison, WindSat SSTs performed well and appeared beneficial to the analyses at warmer temperatures, but exhibited biases at temperatures below about 2°C. Strong SST gradients in the region, particularly near the Alaskan coast, posed challenges for the L4 analyses and led to large differences among the products. SST analyses have lower resolutions that require upscaling the input data streams. This process can decrease the magnitude of cross-frontal SST gradients and/or slightly change the apparent location of the front, which increases the likelihood of the satellite product misrepresenting the point buoy measurement. In dynamic regions, the MUR and CMC analyses exhibited more realistic gradients. For the Louis 2012-03 buoy, the better gradient representation resulted in very strong product performance, but for other coastal buoys, slight uncertainties in the positioning of the gradients could have partially degraded the product scores. In coastal waters, also a region of high spatial variability, the OISST and the GMPE proved to be very skillful. Where temperatures were more uniform, such as the cold waters farther to the north, the products performed more similarly. Here, the CMC, OISST, and GMPE distinguished themselves from the others. The OISST performed best overall, with the best possible score when considering all buoy observations together, closely followed by the CMC and the

GMPE. Those products appear to have the best utility for applications in the challenging Beaufort Sea, at least during this period.

While inclusion of the bias in the SS implies that the score could be influenced by the effective depth of the analysis, that was not a significant factor here. Of the analyses, only FNMOC is representative of other than the foundation or daily average temperature. Given the largely isothermal conditions in the top 10 m, differences between foundation estimates and daily averages are expected to be small. Any biases associated with the skin layer were insignificant relative to other effects here such as cloud filtering.

Derived uncertainties in the analyses for the Beaufort Sea are generally greater than that observed globally. Martin et al. (2012) found that, globally, all the analyses they considered (which included all those considered here except for MUR) had standard deviations less than 0.7 K. In this study, except for the northern buoys where SSTs were very uniform, the standard deviations commonly exceeded 0.7 K. For the coastal buoys (Table 7), the analysis uncertainties exceeded the global values reported by Martin et al. (2012) by about 0.3 K. Interestingly, if we rank the L4s based on the standard deviations for the coastal buoys alone (column 5, Table 7), we obtain the same performing groupings reported by Martin et al. (2012) for their global results, despite the fact that those were estimated for an earlier period when a slightly different mix of sensors was active.

A significant difference between the SST analyses is the approach employed for ice masking. Theoretically, these differences could have an important impact on the merit of the different products near the ice. The results here suggested that the EUMETSAT OSI-SAF ice mask agreed very well with the ice flagging included with the UpTempO data, but the choice of ice mask had surprisingly little impact on the performance of the analyses relative to the buoys for this specific case.

The results here were certainly affected by lack of AMSR-E data during the study period. Several of the analyses that normally incorporate MW data did not do so during this period and others might not have performed as well with the substituted WindSat data. While MW data are of coarser spatial resolution and cannot retrieve close to land or the ice edge, they provide important independent observations in cloudy conditions that can commonly obscure IR retrievals in this region. Performing a similar analysis in a future period when both buoy data and new AMSR2 data are available would be valuable.

Acknowledgments:

This work was supported by the MIZOPEX project with funding from NASA and NOAA. Additional time for SLC was supported through the Multi-Sensor Improved Sea Surface Temperature (MISST 2) project funded through the National Ocean Partnership Project (NOPP). The UpTempO buoy SSTs from the Polar Science Center (PSC) at the Applied Physics Laboratory (APL) in the University of Washington were obtained from the UpTempO Buoy Project web site at http://psc.apl.washington.edu/UpTempO/. The CMC SSTs were provided by Bruce Brasnett at CMC. JPL MUR SSTs were provided by Mike Chin at NASA JPL through the Physical Oceanography Distributed Active Archive Center (PODAAC, podaac. jpl.nasa.gov/datasetlist). The GAMSSA, K10, OISST, MWIR, and OSTIA analyses were accessed through the NOAA NCEI Long Term Stewardship and Reanalysis Facility (LTSRF) (ww.nodc.noaa.gov/sog/GHRSST/accessdata.html). FNMOC SSTs for GHRSST were obtained from the USGODAE server (www.usgodae.org), hosted by the U.S. Navy at the Naval Research Laboratory Marine Meteorology Division (http://www.usgodae.org/pub/outgoing/fnmoc/models/GHRSST/). The GMPE SST data were provided by GHRSST, the Met Office, and MyOcean (www.myocean.eu). The MWIR and WindSat SST retrievals were provided by Remote Sensing Systems, and the AVHRR LAC data were provided by NAVOCEANO.

References

Andersen, S., Breivik, L. A., Eastwood, S., Godøy, Ø., Lind, M., Porcires, M. & Schyberg, H. (2007). OSI SAF Sea Ice Product Manual – v3.5, EUMETSAT OSI SAF – Ocean and Sea Ice Satellite Application Facility, Tech. Rep. SAF/OSI/met.no/TEC/MA125, (pp. 36), available from http://sat.met.no/docs/ss2_pmseaice_v3p5.pdf.

Beggs, H., Zhong, A., Warren, G., Alves, O., Brassington, G., & Pugh, T. (2011). RAMSSA –an operational, high-resolution, multi-sensor sea surface temperature analysis over the Australian region, *Australian Meteorological and Oceanographic Journal*, *61*, 1 – 22.

Brasnett, B. (2008). The impact of satellite retrievals in a global sea-surface-temperature analysis, *Quarterly Journal of the Royal Meteorological Society*, *134*, 1745 – 1760, doi:10.1002/qj.319.

Buehner, M., Caya A., Pogson L., Carrieres, T., & Pestieau, (2013). A new Environment Canada regional ice analysis system, *Atmosphere - Ocean, 51*, 18 – 34, doi: 10.1080/07055900.2012.747171.

Castro, S. L., Wick, G. A., & Emery, W. J. (2012). Evaluation of the relative performance of sea surface temperature measurements from different types of drifting and moored buoys using satellite-derived reference products, *Journal of Geophysical Research*, *117*, doi:10.1029/2011JC007472.

Castro, S. L., Wick, G. A., & Buck, J. J. H. (2014). Comparison of diurnal warming estimates from unpumped Argo data and SEVIRI satellite observations, *Remote Sensing of Environment*, *140*, 789 – 799, doi:10.1016/j.rse.2013.08.042.

Chin, T. M., Milliff, R.F. & Large, W.G. (1998). Basin-scale, high-wavenumber sea surface wind fields from a multiresolution analysis of scatterometer data, *Journal of Atmospheric and Oceanic Technology*, *15*, 15,741 – 763.

Cummings, J. A. & Smedstad, O. M. (2013). Variational Data Assimilation for the Global Ocean, Chapter 13, pp 303 – 343. In Data Assimilation for Atmospheric, Oceanic and Hydrological Applications (Vol. II). S. Park and L. Xu (eds), doi:10.1007/978-3-642-35088-7 13, Springer-Verlag, Berlin, Heidelberg.

Daley, R. (1991). Atmospheric Data Analysis, Cambridge University Press, Cambridge, UK.

Dash, P., Ignatov, A., Martin, M., Donlon, C., Brasnett, B., Reynolds, R., & Coauthors (2012). Group for High Resolution Sea Surface Temperature (GHRSST) analysis fields inter-comparisons. Part 2: Near-real time web-based Level 4 SST Quality Monitor (L4-SQUAM), *Deep-Sea Research II*, *77-80*, 31 – 43, doi:10.1016/j.dsr2.2012.04.002.

Dong, S., Guille, S.T., Sprintall, J., & Gentemann, C.L. (2006). Validation of the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) sea surface temperature in the Southern Ocean, *Journal of Geophysical Research*, *111*, C04002, doi:10.1029/2005JC002934.

Donlon, C. J., I. Robinson, K. S. Casey, J. Vazquez-Cuervo, E. Armstrong, O. Arino & Coauthors (2007). The Global Ocean Data Assimilation Project (GODAE) High Resolution Sea Surface Temperature Pilot Project (GHRSST-PP), *Bulletin of the American Meteorological Society*, 88, 1197-1213.

Donlon, C. J., Martin, M., Stark, J. D., Robert-Jones, J., Fiedler, E., & Wimmer, W. (2012). The operational sea surface temperature and sea ice analysis (OSTIA) system, *Remote Sensing of Environment*, *116*, 140 – 158, doi:10.1016/j.rse.2010.10.017.

Gandin, L. (1965). Objective Analysis of Meteorological Fields. *Israel Program for Scientific Translations*, 22 pp.

Gentemann, C.L., Wentz, F.J., & DeMaria, M. (2006). Near real time global optimum interpolated microwave SSTs: Applications to hurricane intensity forecasting. Presented at 27th Conference on Hurricanes and Tropical Meteorology, Monterrey, CA.

Grumbine, R. W. (1996). Automated passive microwave sea ice concentration analysis at NCEP. NOAA Tech. Note 120, 13 pp. [available from NCEP/NWS/NOAA, 5200 Auth Road, Camp Springs, MD 20746.]

Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: translating threshold choice into expected classification loss, *Journal of Machine Learning Research*, *13*, 2813 – 2869.

Hoyer, J. L., Karagali, I., Dybkjaer, G., & Tonboe, R. (2012). Multi sensor validation and error characteristics of Arctic satellite sea surface temperature observations, *Remote Sensing of Environment*, *121*, 335 – 346, doi:10.1016/j.rse.2012.01.013.

Mallat, S.G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Analysis Machine Intelligence*, *11*, 674–693.

Markus, T., & Cavalieri, D. J. (2000). An enhancement of the NASA Team sea ice algorithm, *IEEE Transactions on Geoscience and Remote Sensing*, *38*, 1387 – 1398.

Martin, M., Dash, P., Ignatov, A., Banzon, V., Beggs, H., Brasnett, B., & Coauthors (2012). Group for High Resolution Sea Surface Temperature (GHRSST) analysis fields inter-comparisons. Part 1: A GHRSST multi-product ensemble (GMPE), *Deep-Sea Research II*, *77-80*, 21 – 30, doi:10.1016/j.dsr2.2012.04.013.

May, D. A., M. M. Parmeter, D. S. Olszewski, and B. D. McKenzie (1998). Operational processing of satellite sea surface temperature retrievals at the Naval Oceanographic Office, Bulletin of the American Meteorological Society, 79, 397-407.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient, *Monthly Weather Review*, *116*, 2417 – 2424.

Przybylak, R. (2003). The Climate of the Arctic. Series: Atmospheric and Oceanographic Sciences Library, Vol. 26, Kluwer Academic Publishers, Springer, XI, 272 p. ISBN 1-4020-1134-2

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. C., & Schlax, M. G. (2007). Daily high-resolution blended analyses for sea surface temperature, *Journal of Climate*, *23*, 5473 – 5496.

Reynolds, R. W. & Chelton, D. B., (2010). Comparisons of daily sea surface temperature analyses for 2007 – 08, *Journal of Climate*, *23*, 3545 – 3563, doi:10.1175/2010JCLI3294.1.

Simmonds, I. & Rudeva, I. (2012). The great Arctic cyclone of August 2012, *Geophysical Research Letters*, *39*, L23709, doi:10.1029/2012GL054259.

Smith, G.C.., Roy, F., Reszka, M., Surcel-Colan, D., He, Z., Deacu, D., & Coauthors (2015). Sea ice forecast verification in the Canadian Global Ice Ocean Prediction System, *Quarterly Journal of the Royal Meteorological Society*, *142*, 659 – 671, doi: 10.1002/qj.2555.

Steele, M., Ermold, W., Colburn, K., & Rigor, I. (2016). UpTempO buoy, *Journal of Atmospheric and Oceanic Technology*, in preparation.

Taylor, K. E., (2001). Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research*, *106*, 7183 – 7192.

Zhang, J., Lindsay, R., Schweiger, A., & Steel, M. (2013). The impact of an intense summer cyclone on 2012 Arctic sea ice retreat, *Geophysical Research Letters*, *40,* 720 – 726, doi:10.1002/grl.50190.

Zhong, A. & Beggs, H. (2008). Operational Implementation of Global Australian Multi-Sensor Sea Surface Temperature Analysis, *Analysis and Prediction Operations Bulletin No. 77*, Bureau of Meteorology, Australia, 2 October 2008, http://www.bom.gov.au/australia/charts/bulletins/apob77.pdf.