Intercomparison of long-term sea surface temperature analyses using the GHRSST Multi-Product Ensemble (GMPE) system

Article

Accepted Version

Fiedler, E. K., McLaren, A., Banzon, V., Brasnett, B., Ishizaki, S., Kennedy, J., Rayner, N., Roberts-Jones, J., Corlett, G., Merchant, C. J. ORCID: https://orcid.org/0000-0003-4687-9850 and Donlon, C. (2019) Intercomparison of long-term sea surface temperature analyses using the GHRSST Multi-Product Ensemble (GMPE) system. Remote Sensing of Environment, 222. pp. 18-33. ISSN 0034-4257 doi: https://doi.org/10.1016/j.rse.2018.12.015 Available at https://centaur.reading.ac.uk/81206/

To link to this article DOI: http://dx.doi.org/10.1016/j.rse.2018.12.015

Publisher: Elsevier

Intercomparison of long-term sea surface temperature analyses using the GHRSST Multi-Product Ensemble (GMPE) system

Emma K. Fiedler[a,*], Alison McLaren[a], Viva Banzon[b], Bruce Brasnett[c], Shiro Ishizaki[d], John Kennedy[a], Nick Rayner[a], Jonah Roberts-Jones[a], Gary Corlett[e], Christopher J. Merchant[f,g], Craig Donlon[h]

[a] *Met Office, Exeter, UK*
[b] *NOAA/NCEI, Asheville, NC, USA*
[c] *Canadian Meteorological Centre, Dorval, Quebec, Canada*
[d] *Japan Meteorological Agency, Tokyo, Japan*
[e] *University of Leicester, Leicester, UK*
[f] *University of Reading, Reading, UK*
[g] *National Centre for Earth Observation, UK*
[h] *ESA/ESTEC (EOP-SME), Noordwijk, The Netherlands*

Abstract

Six global, gridded, gap-free, daily sea surface temperature (SST) analyses covering a period of at least 20 years have been intercompared: ESA SST CCI analysis long-term product v1.0, MyOcean OSTIA reanalysis v1.0, CMC 0.2 degree, AVHRR_ONLY Daily 1/4 degree OISST v2.0, HadISST2.1.0.0 and MGDSST. A seventh SST product of the ensemble median of all six has also been produced using the GMPE (Group for High Resolution SST Multi-Product Ensemble) system. Validation against independent near-surface Argo data, a long timeseries of moored buoy data from the tropics and anomalies to the GMPE median have been used to examine the temporal and spatial homogeneity of the analyses. A comparison of the feature resolution of the analyses has also been undertaken. A summary of relative strengths and weaknesses of the SST datasets is presented, intended to help users to make an informed choice of which analysis is most

*Corresponding author
Email address: emma.fiedler@metoffice.gov.uk (Emma K. Fiedler)

suitable for their proposed application.

Keywords: SST, analysis, compare, global, L4, dataset

1. Introduction

Long-term analyses, also known as reanalyses, of sea surface temperature (SST) based on satellite observations are useful for a variety of applications, including as boundary conditions in atmospheric models and for long-term monitoring of SST. Several long-term, daily SST analyses covering at least a 20-year period exist. Despite the use of similar input data (e.g. observations from AVHRR (Advanced Very High Resolution Radiometer) and ATSR (Along-Track Scanning Radiometer)-series instruments) it is well known that there are differences between them, particularly in high gradient regions such as western boundary currents (e.g. Reynolds & Chelton, 2010; Roberts-Jones et al., 2012). This is due to differing processing methods including analysis grid size, bias-correction techniques, and analysis procedures including selection of horizontal background error correlation length scales. There are also differences resulting from variations in the resolution and processing of the input data, including different retrieval methods and techniques for obtaining uncertainty estimates.

Multiple realisations of SST timeseries using different data combinations and techniques can not only be used to highlight problems, but can also be beneficial by providing users with a choice of product to best suit their needs. For example, climate-related applications require a homogeneous, stable timeseries without the artifical temporal variability that can be introduced when including non-homogenised data from additional instruments during the timeseries. However, the accuracy of the analysis may be improved, potentially at the expense of stability, by utilising data from a wider variety of sources as they become available. This sort of dataset is useful for applications such as short-range model forcing and validation.

The aim of this study is to assess the relative strengths and weaknesses of various long-term SST analysis datasets. An intercomparison of the analyses will be undertaken using the GMPE (Group for High Resolution SST (GHRSST) Multi-Product Ensemble) system, described by Martin et al. (2012). This system is a tool that produces an ensemble median SST product from contributing SST analyses as well as having the capability to generate matchups of the analyses with in situ observations for validation. An important use of the GMPE median product is to assess the deviations of the contributing analyses from it. The main advantage of using this dataset for validation is that it provides complete and consistent spatial and temporal coverage, unlike in situ reference data. Martin et al. (2012) found the GMPE median product to perform better compared to Argo than any of the component analyses used to generate it. A GMPE median product is generated daily at the Met Office using NRT (near-real-time) SST analyses as input, and is available from CMEMS (Copernicus Marine Environment Monitoring Service; marine.copernicus.eu). Monthly statistics of the NRT input analyses compared to Argo observations are available at `http://ghrsst-pp.metoffice.com/pages/latest_analysis/sst_monitor/argo` . Note that results for NRT versions of the input analyses are not necessarily directly comparable to the long-term analysis versions of the same products assessed here, owing to differences in methods and input data.

Near-surface data from Argo floats will be used to determine global and regional performance of the analyses, based on the mean and standard deviations of matchup differences generated using the GMPE system. A long and stable timeseries of observations from tropical moored buoys will be used to assess the temporal homogeneity of the datasets. A comparison of feature resolution will also be undertaken. Characteristics of the individual analyses will thus be evaluated and intercompared, the results of which will allow users to make informed choices about which analysis is most suitable for their purpose.

The GMPE system has not previously been used to intercompare long-term analyses, and a systematic intercomparison of all available long-term daily SST analyses has not previously been conducted. Other SST intercomparison projects have previously taken place, notably the Global Climate Observing System (GCOS) SST-Sea Ice intercomparison project (`https://www.nodc.noaa.gov/SatelliteData/ghrsst/intercomp.html`), but this focused on weekly and monthly datasets with lower spatial resolutions, rather than the daily, high resolution datasets used here. Other intercomparison projects organised through the framework of GHRSST include L4-SQUAM (SST Quality Monitor; Dash et al., 2012) which monitors global SST analysis quality, and HR-DDS (High Resolution Diagnostic Data Set; Poulter et al., 2008) and its more recent ESA evolution, Felyx (Taberner et al., 2013), which compare datasets at pre-defined locations. However, these projects are mainly concerned with intercomparison of short-term SST analyses on a NRT basis, and not long timeseries.

This work was conducted under the ESA SST CCI (European Space Agency Sea Surface Temperature Climate Change Initiative) project, as part of the validation stage of the ESA SST CCI analysis long-term product. The long-term GMPE median SST product (Fiedler et al., 2015) used in this study has been made freely available, and can be accessed through CEDA (Centre for Environmental Data Analysis) at `http://catalogue.ceda.ac.uk/uuid/e0659b01259145c8bfb0de6eb12c2690`.

The structure of this paper is as follows. Section 2 provides information on the analysis datasets and methods used in this study. In section 3.1, the performance of each analysis is assessed against near-surface Argo data. A long and stable timeseries of observations from tropical moored buoys at 1 m depth are then used to compare the temporal homogeneity of the analyses over the whole time period in section 3.2. In section 3.3, the six analyses are intercompared in terms of their anomaly to the GMPE median, and their relative contributions to the GMPE median are evaluated. Finally, a comparison of the analysis SST gradients is presented in section 3.4, followed by conclusions and a summary in section 4.

2. Data and methods

2.1. Contributing datasets

Six internationally-produced, daily, global, L4 (“level-4”: gap-free, gridded) SST analyses with at least 20 years’ worth of data and a minimum spatial resolution of $1/4^o$ have been used: ESA SST CCI (European Space Agency Sea Surface

Temperature Climate Change Initiative) analysis long-term product v1.0 (referred to herein as SST CCI analysis; Merchant et al., 2014), MyOcean OSTIA (Operational Sea Surface Temperature and Ice Analysis) reanalysis v1.0 (referred to herein as OSTIA v1.0; Roberts-Jones et al., 2012), CMC (Canadian Meteorological Center) 0.2 degree analysis (referred to herein as CMC; Brasnett, 2012), AVHRR (Advanced Very High Resolution Radiometter)_ONLY Daily 1/4 degree OISST (Optimal Interpolation Sea Surface Temperature) v2.0 (referred to herein as AVHRR-OI; Reynolds et al., 2007; Reynolds, 2009; Banzon et al., 2016), HadISST2.1.0.0 (Hadley Centre Ice and Sea Surface Temperature) realisation 396 (referred to herein as HadISST2; Kennedy et al., 2018; Rayner et al., 2018) and MGDSST (Merged satellite and in situ data Global Daily Sea Surface Temperature) analysis (Kurihara et al., 2006). Data were obtained directly from the producers, with the exception of AVHRR-OI, which was downloaded via ftp from PO.DAAC (NASA JPL Physical Oceanography Distributed Active Archive Data Center). Access locations for all the datasets are provided in the "Data Access" section at the end of this paper.

The SST CCI analysis was produced using different input data and an upgraded version of the OSTIA system previously used to produce the OSTIA v1.0 reanalysis. Updates to the system to produce the new analysis are described in Roberts-Jones et al. (2013). HadISST2.1.0.0 realisation 396 was randomly selected from the available set of 10 interchangeable realisations, which are intended to provide information about the likely spread arising from uncertainty in the measurements and reconstruction. The dataset is based on a

5-day, 1^o resolution dataset that has been interpolated to 1-day, $1/4^o$ resolution by the data producers. HadISST2.1.0.0 was available to 2007 at the time this work was conducted. It has subsequently been made available to 2010. A version of the AVHRR-OI dataset which also includes microwave data is available (AVHRR+AMSR Daily 1/4 degree OISST v2.0; Reynolds et al., 2007; Reynolds, 2009), but this is not used in the comparisons due to the shorter length of the available timeseries (just over 9 years) compared to other datasets used here (at least 20 years).

Information on these datasets is summarised in Table 1, including references that provide detailed descriptions of the datasets and the methods used to generate them. All of these analysis datasets use optimal interpolation assimilation methods. The SST CCI analysis is the only long-term dataset not to use in situ data as an input, and is based on infra-red satellite data only. All datasets include observations derived from AVHRR sensors and, with the exception of MGDSST and AVHRR-OI, the analyses all use data from the ATSR-series of instruments. Only MGDSST and CMC include data from microwave instruments. Different data sources given in Table 1 for the same instruments mean the retrievals will have undergone different processing. Input data to all the analyses undergo bias correction, either to ATSR-series data or in situ observations, or a combination of both (Table 1).

Although the datasets are all "SST" products, they are intended to be valid at a variety of near-surface depths, for use in different applications. The SST CCI analysis uses input data specifically adjusted to 20 cm depth and to lo-

Table 1: Information on analysis datasets. [ ] indicates data source. Acronyms: Data Providers: ARC = AATSR Reprocessing for Climate, CCI = Climate Change Initiative, ESA = European Space Agency, GSFC = Goddard Space Flight Center, JMA = Japan Meteorological Agency, NAVO = U.S. Naval Oceanographic Office, NCEP = National Centers for Environmental Prediction, NEODC = Natural Environment Research Council Earth Observation Centre, NESDIS = National Environmental Satellite, Data, and Information Service, OSI SAF = Ocean and Sea Ice Satellite Application Facility, REMSS = Remote Sensing Systems. Instruments: AMSR-E = Advanced Microwave Scanning Radiometer - Earth observing system, ATSR = Along-Track Scanning Radiometer, AVHRR = Advanced Very High Resolution Radiometer, TMI = Tropical rainfall measuring mission Microwave Imager. Datasets: GTS = Global Telecommunications System, ICOADS = International Comprehensive Ocean-Atmosphere Data Set. *Now available 1961-2010.

		Infra-red sensors		Microwave sensors						
Analysis and Citation	Time period and SST depth/time	AVHRR	ATSR-series	AMSR-E	TMI	WindSat	In situ	Ice data source	Grid resolution (degrees)	Bias-correction reference
ESA SST CCI analysis long-term product (SST CCI); Merchant et al. (2014)	1991-2010 daily mean at 20 cm	NOAA12-19 [CCI, v1.0]	ATSR-1,2, AATSR [CCI, v1.0]	None	None	None	None	OSI SAF OSI-409 v1.1 (1991-Oct 2009), OSI-401-a v1.2 (Oct 2009-2010)	1/20	ATSR-1,2, AATSR
MyOcean OSTIA reanalysis v1.0 (OSTIA v1.0); Roberts-Jones et al. (2012)	1985-2007 foundation	Pathfinder V5.0/V5.1 (1985-2007)	ATSR-1,2, AATSR [ESA/NEODC, v2.0]	None	None	None	ICOADS v2.0	OSI SAF OSI-409 v1.0	1/20	ATSR-2, AATSR, in situ
CMC 0.2 degree (CMC); Brasnett (2012)	1991-2011 1 m (referenced to ship and buoy data)	NOAA16-19 (2001-2011) [NAVO]; MetOp-A (2007-2011) [NAVO]	ATSR-1,2, AATSR [ESA, v2.0]	2002-2011 [REMSS]	1998-2002 [REMSS]	2003-2011 [REMSS]	ICOADS v2.5; GTS (after 2006)	OSI SAF OSI-409 v1.0 (1991-Oct 1998), CMC (Oct 1998-2011)	1/5	In situ (separate day and night)
AVHRR_ONLY Daily 1/4 degree OISST v2.0 (AVHRR-OI); Reynolds et al. (2007); Reynolds (2009); Banzon et al. (2016)	1981-present mean	Pathfinder V5.0/V5.1 (1981-2005); NOAA-unspecified, 2 sensors at a time (2006-present) [NAVO]	None	None	None	None	ICOADS v2.4; GTS (after 2006)	GSFC NASA NSIDC-0051 (1981-2004), NCEP (2005-present)	1/4	In situ
HadISST2.1.0.0, realisation 396 (HadISST2); Kennedy et al. (2018); Rayner et al. (2018)	1961-2007* 20 cm	Pathfinder V5.0/V5.1 (1981-2006)	ATSR-1 (3-channel retrievals only), ATSR-2, AATSR [ARC, v1.1]	None	None	None	ICOADS v2.5	HadISST2 (Titchner & Rayner, 2014)	1/4 (daily, interpolated from 1^o, 5-day product)	ATSR-1 (3-channel retrievals only), ATSR-2, AATSR, in situ
MGDSST; Kurihara et al. (2006)	1982-2011 foundation	Pathfinder V5.0/V5.1 (1982-2006); NOAA17-19 (2007-2011) [NESDIS]; MetOp-A (2010-2011) [NESDIS]	None	2003-2011 [JAXA]	None	2011 [JAXA]	GTS	JMA	1/4	In situ

cal times of 1030 hrs and 2230 hrs, producing an estimate of the daily mean temperature at this depth (Merchant et al., 2014). This is the only analysis to use methods for producing data valid for both a specified depth and local time. The HadISST2 dataset is also valid for a nominal depth of 20 cm. The OSTIA v1.0 and MGDSST reanalyses are foundation temperatures, i.e. pre-dawn temperatures without the effects of diurnal warming. This is achieved for the OSTIA v1.0 reanalysis by including daytime data only when the windspeed is greater than 6 m s^{-1} (Donlon et al., 2002), in addition to nighttime data. For MGDSST, satellite data are rejected when the diurnal SST amplitude is greater than 3 K. AVHRR-OI is a mean temperature in the sense that all available data are used but, depending on data availability, an actual daily mean temperature is not necessarily produced. The satellite data used in the CMC analysis is referenced to ship and buoy data which is stated to have a typical depth of 1 m, although no particular method is applied to the analysis to adjust data to a specified depth.

As different SST analyses are designed with slightly different specifications in mind it is not necessarily appropriate to try to determine which is “correct”. However, an intercomparison of a number of different datasets can give an idea of outliers and of which analyses perform well, especially when compared with independent data.

2.2. Methods

The methods used in the GMPE (Group for High Resolution SST Multi-Product Ensemble) system will be briefly described here. For further details the reader is referred to Martin et al. (2012). The SST analyses are first regridded to a regular latitude-longitude, $1/4^o$ GMPE grid using a bilinear interpolation. An ensemble median SST (referred to herein as the "GMPE median") and standard deviation for each grid box are calculated from the contributing analyses. The use of a median rather than a mean minimises the effect of potential outliers in the data on the ensemble value. If there are an even number of analyses, the mean of the two centre analyses is taken. The production of a median SST using all the datasets provides a new SST product that potentially has smaller errors than any of the component analyses, as was found for the GMPE median product generated using NRT SST analysis datasets as input (Martin et al., 2012).

When the GMPE system is run using NRT analysis datasets, the land mask and updated sea ice mask for each day are taken from the Met Office NRT OSTIA (Operational Sea Surface Temperature and Ice Analysis) product (Donlon et al., 2012). Here they will be taken from the SST CCI analysis, also processed at the Met Office using the OSTIA system. The sea ice data used for the SST CCI product is sourced from EUMETSAT OSI SAF products OSI-409 v1.1 (used for 1991 - October 2009) and OSI-401-a v1.2 (October 2009 - 2010) (Table 1). Using a linear interpolation method, files were created to fill gaps in the OSI SAF timeseries using the method described in Roberts-Jones et al. (2013). The data were regridded from the native 10 km polar stereographic grid to the regular latitude-longitude $1/20^o$ OSTIA grid and bilinear interpolation was used to perform spatial filling around coasts. For use in the GMPE system, the sea ice was then regridded to the same $1/4^o$ grid used for SST.

3. Intercomparison of analyses

3.1. Validation of SST analyses using independent Argo data

3.1.1. Argo matchup statistics

Temperature data from Argo profiling floats have been used here for validation of the six SST analyses and their ensemble (GMPE) median. The Argo dataset is suitable for use as a reference for validation since it is both accurate and stable (Oka & Ando, 2004). It is also the only in situ dataset from which SST analysis products are kept independent, for validation purposes. This is by mutual agreement through GHRSST. Near-surface (3-5 m depth) Argo measurements are used here, which provide an estimate of foundation SST (the pre-dawn temperature, i.e. without the effects of diurnal warming). This is demonstrated by Figure 1, which illustrates the close match between 3-5 m depth Argo data and nighttime measurements from drifting buoys at 20 cm depth. The mean difference of the matchups is 0.004 K, with a standard deviation of 0.60 K. The rather large standard deviation is a result of the inclusion of matchups in high gradient regions such as western boundary currents, but the global distribution is shown in Figure 1 for completeness. Matchup criteria for Figure 1 are within 3 hours and 50 km, for Argo and drifter data between 2005-2013. Observations were extracted from the HadIOD database v1.0.0.0 (Atkinson et al., 2014), where the data undergo rigorous quality control procedures.

The various analyses are intended to be valid for different depths (Table 1

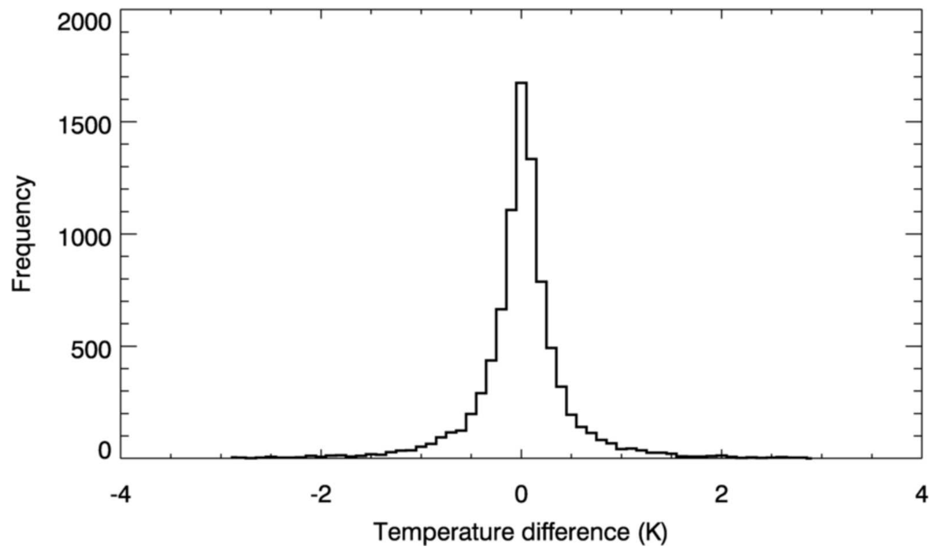


Figure 1: Distribution of nighttime Argo minus drifting buoy differences, 2005 - 2013, 0.1 K bins. Mean difference 0.004 K, standard deviation 0.60 K. Differences are taken from matchups within 3 hours and 50 km.

and section 2.2). We would therefore expect to find differences compared to the Argo foundation temperature and this should be taken into account when comparing the following results.

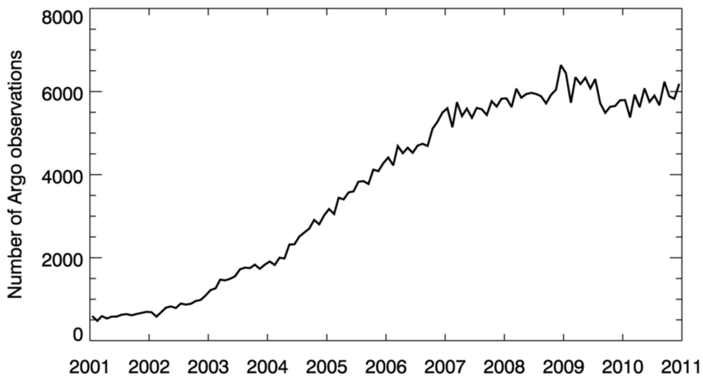
Daytime and nighttime Argo observations have been extracted from the EN4 dataset (Good et al., 2013), where they have undergone quality control procedures to remove suspect observations. For each available profile, the shallowest observation between 3-5 m was obtained. A minimum depth of 3 m is used based on the assumption that this is the depth at which the effects of diurnal warming can be neglected (Zeng & Beljaars, 2005; Gentemann et al., 2009; Takaya et al., 2010). The number of Argo observations increases over time (Figure 2(a)). The dataset matures by 2007, having spread to almost cover the global ocean except for marginal seas and continental shelves (Figure 2(b)-(d)).

The GMPE system was used to produce matchups between the analyses and the Argo data, by interpolating the analyses from their native resolutions to the observation locations using a bilinear interpolation. The error on this interpolation is negligible, owing to the high resolution of the analyses. Monthly means and standard deviations of the analysis differences to Argo were calculated for 2001 - 2010 (or 2001 - 2007 for HadISST2 and OSTIA v1.0) and a timeseries of the results is shown in Figure 3.

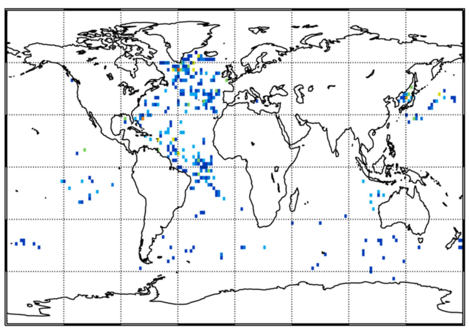
Figure 3 demonstrates that the CMC, SST CCI and the GMPE median datasets have the smallest monthly global standard deviation of the differences to Argo (Figure 3). MGDSST and OSTIA v1.0 are in the centre of the spread, and AVHRR-OI and HadISST2 have the largest global standard deviations (Figure 3).

The noisy statistics in Figure 3 prior to 2003 demonstrate the detrimental effect of a reduced matchup data volume on the robustness of monthly statistics, and illustrate that the number of floats necessary for a robust result is approximately 1000 (c.f. Figure 2(a)). Therefore results were only considered for the period 2003 and later.

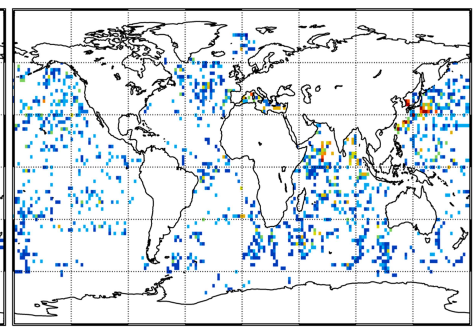
The global mean standard deviation of the differences, weighted by the number of observations, for each of the analyses compared to Argo over the time period 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2) indicates the analysis with the smallest mean standard deviation is CMC (Table 2). At 0.41 K, this is very similar to that of the GMPE median (0.42 K). This is unexpected, given that the GMPE median was found to have a smaller global standard de-



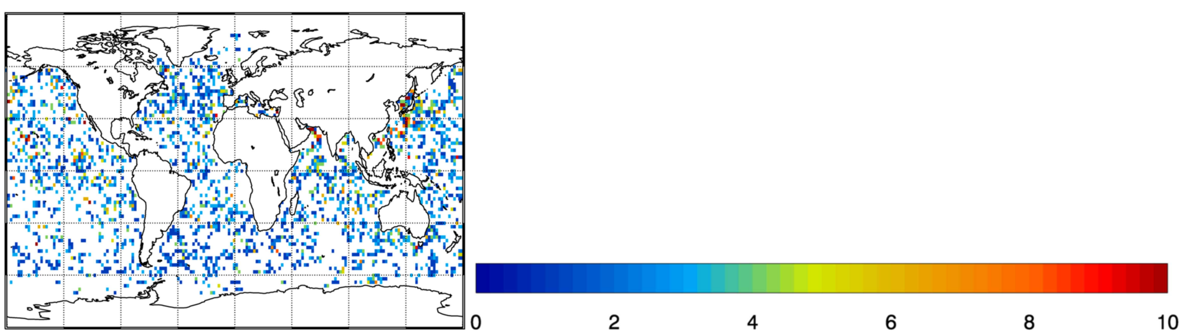
(a) Monthly total global near-surface Argo observations



(b) January 2001



(c) January 2005



(d) January 2010

Figure 2: (a) Timeseries of monthly total number of global near-surface Argo observations, using shallowest observation between 3 m and 5 m depth, and (b-d) spatial map of same for given month binned in 2x2 degree grid boxes.

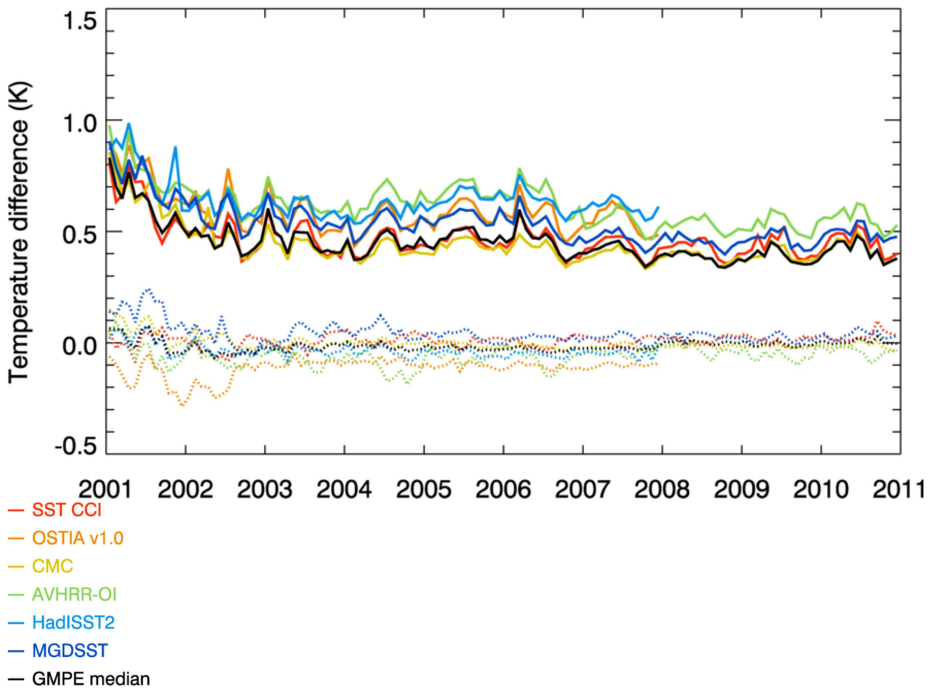


Figure 3: Timeseries of monthly analysis-minus-Argo SST differences: mean (dashed line) and standard deviation (solid line) for six analyses and their ensemble (GMPE) median, 2001-2010. All analyses independent from Argo.

viation of the differences to Argo by at least 0.05 K than all its component analyses in the NRT version of GMPE (Martin et al., 2012). A possible reason for the good performance is that CMC is the only contributing analysis to use two sets of microwave data (AMSR-E and WindSat) in addition to infra-red data (from AVHRR and ATSR) for the time period of this Argo comparison (Table 1).

SST CCI also performs well against Argo data, with a small standard deviation of the differences (0.44 K) compared to other analyses (Table 2). This is despite the SST CCI analysis being a satellite-only product, unlike the other analyses which also assimilate in situ observations. SST CCI (which uses dif-

Table 2: Global analysis-minus-Argo SST differences: mean difference, mean absolute difference and standard deviation, in K, for six analyses and their ensemble (GMPE) median, 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2).

Analysis	STD	Mean diff	Mean absolute diff	Number of Argo observations
SST CCI	0.44	0.01	0.28	430936
OSTIA v1.0	0.56	-0.10	0.36	216306
CMC	0.41	-0.01	0.25	430935
AVHRR-OI	0.58	-0.05	0.40	430938
HadISST2	0.62	-0.05	0.42	213383
MGDSST	0.49	0.03	0.31	430921
GMPE median	0.42	-0.01	0.26	429219

ferent input data (Table 1) and an upgraded version of the OSTIA system) is clearly an improvement over the OSTIA v1.0 reanalysis (Figure 3, Table 2). This will be discussed further in section 3.1.2.

CMC, the GMPE median and SST CCI datasets all have the lowest magnitude global differences to Argo (Figure 3, Table 2). The mean absolute difference (Table 2) is also the smallest for CMC, the GMPE median and SST CCI. Regional differences to Argo data have also been examined. Figure 4 shows the weighted mean spatial analysis-minus-Argo differences in 2x2 degree grid boxes for 2003-2010 (or 2003-2007 for HadISST2 and OSTIA v1.0). Figure 5 gives the mean values weighted by number of observations for various ocean regions as defined by MyOcean (now CMEMS; e.g. McLaren et al. (2014)) of the analysis-minus-Argo differences and standard deviations.

As well as having the smallest global differences to Argo (Figure 3, Table 2), CMC and the GMPE median perform well in all regions (Figures 4, 5). Although the global average of the mean difference to Argo for SST CCI is small (Table 2) the SST in the tropical Pacific is around 0.1 K too warm compared to Argo (Figure 5), with some regional variation (Figure 4). This bias is related to problems with the SST CCI input data in this region (Corlett et al., 2014). However, along with CMC and the GMPE median, SST CCI performs well regionally in terms of the standard deviation of the differences to Argo data (Figure 5).

Figure 6 shows the mean difference of each analysis to Argo, on a 5x5 degree grid and averaged zonally. The 5x5 degree grid was used instead of the noisier 2x2 degree grid used above, to avoid obscuring the main patterns of spatial homogeneity. Data for 2003 to 2007 was used for all analyses for a direct comparison of results. Figure 6 demonstrates the mean difference of the CMC analysis to Argo is small and noticeably more uniform compared to the mean differences for the other analyses, including the GMPE median. The use of observations of foundation temperature from Argo as reference data means that analyses which are intended to represent shallower depths may be warmer than Argo (e.g. SST CCI and HadISST2 at 20 cm, and CMC at 1 m (Table 1)) and this difference may vary both seasonally and latitudinally. However, only MGDSST (and CCI at low latitudes) are warm compared to Argo (Figure 6; see also Figures 4 and 5(a)). Nevertheless, this mismatch of depths may contribute to the variation in mean differences with latitude seen in Figure 6, and in Figures 4 and 5. However, the difference of MGDSST and OSTIA v1.0 foundation temperatures to those measured by Argo indicates the depth effect is not the

Table 3: Regional analysis-minus-Argo SST differences: mean difference and standard deviation, in K, for selected analyses, 2003-2007.

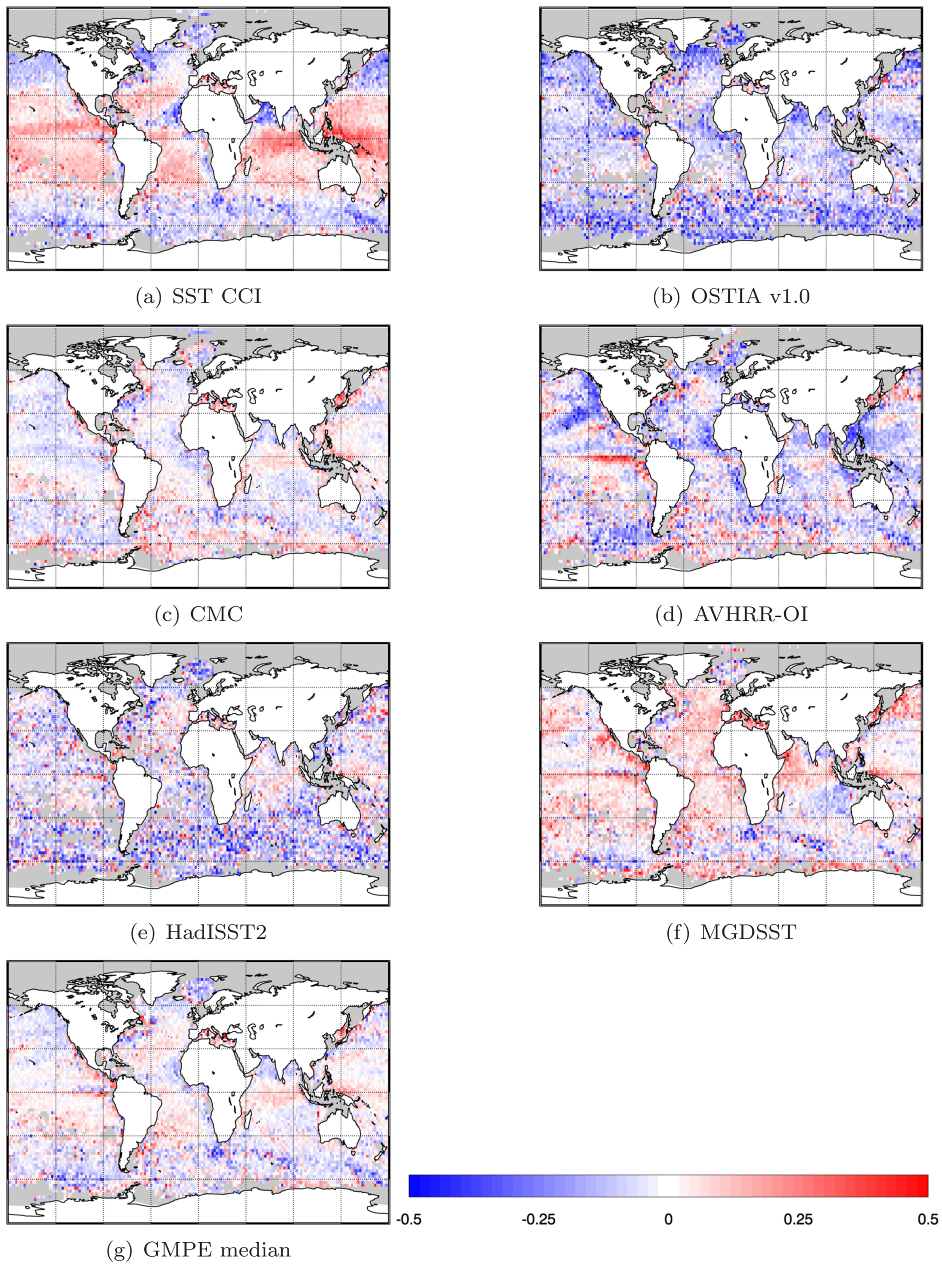
	SST CCI		OSTIA v1.0		GMPE median		CMC	
Region	STD	Mean diff	STD	Mean diff	STD	Mean diff	STD	Mean diff
Global	0.45	0.01	0.56	-0.10	0.44	-0.03	0.41	-0.01
N Atlantic	0.53	-0.01	0.67	-0.11	0.53	-0.02	0.48	-0.01
Tr Atlantic	0.41	0.03	0.45	-0.09	0.36	0.00	0.33	-0.01
S Atlantic	0.49	-0.02	0.67	-0.11	0.53	-0.04	0.47	-0.01
N Pacific	0.48	0.01	0.60	-0.08	0.47	-0.02	0.47	-0.02
Tr Pacific	0.30	0.10	0.35	-0.09	0.27	0.00	0.26	0.00
S Pacific	0.34	0.01	0.43	-0.12	0.35	-0.03	0.34	-0.02
Indian Ocean	0.37	0.06	0.41	-0.08	0.33	-0.01	0.33	-0.01
Southern Ocean	0.44	-0.07	0.62	-0.16	0.49	-0.07	0.45	-0.02

only factor influencing the differences.

It should be noted that conclusions drawn from these results regarding the relative performance of the analyses are only strictly valid for the period of the timeseries from 2003. In particular, the validation does not cover the period where most analyses use observations from the problematic ATSR-1 sensor. However, as demonstrated in section 3.3.2, assessment of the relative contribution of the analyses to the GMPE median throughout the whole timeseries produced similar conclusions to those provided by the Argo validation for the latter part only.

3.1.2. Comparison of OSTIA v1.0 and SST CCI analyses

The SST CCI analysis (Merchant et al., 2014) was produced using new input data and an updated version of the Met Office OSTIA system used to produce the OSTIA v1.0 reanalysis (Roberts-Jones et al., 2012). Roberts-Jones



(a) SST CCI

(b) OSTIA v1.0

(c) CMC

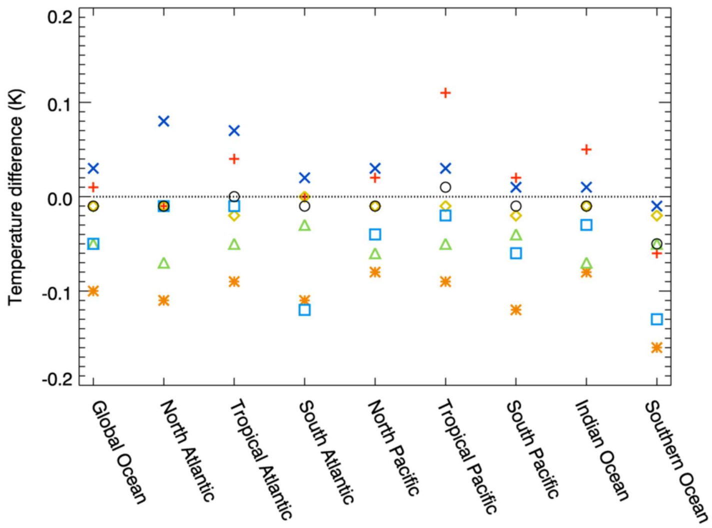
(d) AVHRR-OI

(e) HadISST2

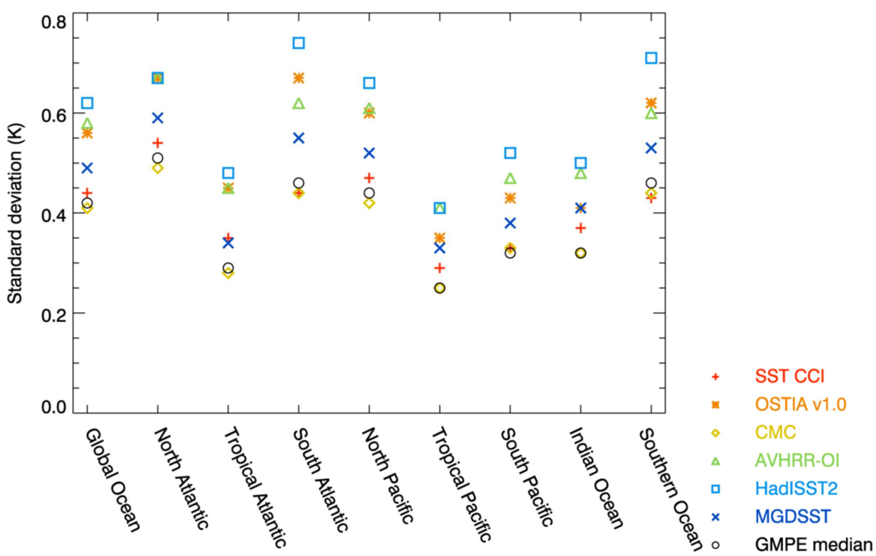
(f) MGDSST

(g) GMPE median

Figure 4: Spatial maps of mean global analysis-minus-Argo SST differences (K) for 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2) in 2x2 degree gridboxes, for six analyses and their ensemble (GMPE) median. Areas with no data shown in grey. All analyses independent from Argo.



(a) Mean difference to Argo



(b) Standard deviation of difference to Argo

Figure 5: Regional analysis-minus-Argo SST differences: (a) mean and (b) standard deviation for six analyses and their ensemble (GMPE) median, 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2). All analyses independent from Argo.

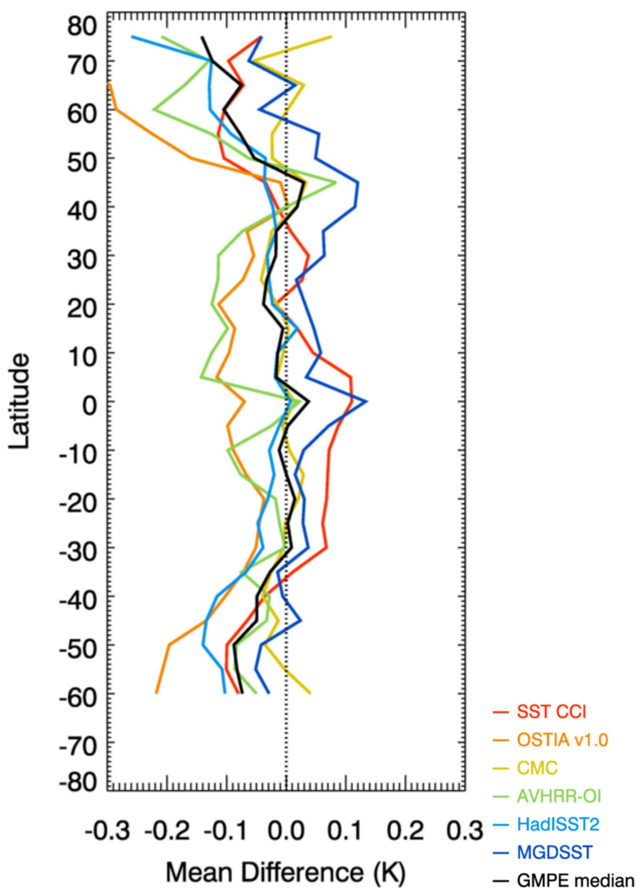


Figure 6: Zonal average of analysis-minus-Argo mean differences, 2003-2007, on a 5x5 degree grid for six analyses and their ensemble (GMPE) median. Minimum number of matchups in each grid box is 50. All analyses independent from Argo.

et al. (2013) and Roberts-Jones et al. (2016) give details of these updates. The OSTIA v1.0 reanalysis is a foundation temperature and the SST CCI analysis is a daily mean temperature at 20 cm depth. The SST CCI analysis would therefore be expected to have a small diurnal warming component compared to the foundation temperature. The magnitude of this is highly dependent on time of year and latitude, but on a global scale can be quantified as around 0.15 K for low wind speeds (0-3 m s^{-1}) and around 0.05 K for wind speeds above 7 m s^{-1} (Merchant et al., 2014).

Table 3 shows regional analysis-minus-Argo mean differences and standard deviations for the OSTIA v1.0 and SST CCI datasets. For comparison, the GMPE median and CMC statistics are also shown in Table 3. Statistics are all shown for 2003-2007 for direct comparison with the OSTIA v1.0 reanalysis, as this dataset ends in 2007. In all regions the standard deviation of differences to Argo is improved (reduced in magnitude) for the new SST CCI analysis compared to the OSTIA v1.0 reanalysis. With the exception of the tropical Pacific, the mean difference to Argo is also improved. Outside of the tropics, the results for SST CCI are much closer to the statistics for the GMPE median and CMC than are those for OSTIA v1.0. This demonstrates that the newer OSTIA reanalysis product, SST CCI, is now in line with the best-performing SST products, using Argo as a validation reference.

3.2. Assessment of temporal homogeneity of SST analyses using moored buoy data

Temperature observations from the GTMBA (Global Tropical Moored Buoy Array) dataset (McPhaden et al., 2009) were used as a reference to assess the temporal homogeneity of the six SST analyses and their ensemble (GMPE) median in tropical regions. This dataset was chosen as a complement to Argo for validation of the SST analyses due to its long timeseries, from the 1980s to present, which has been shown to possess a high degree of temporal stability (Merchant et al., 2012). The buoys are routinely maintained and pre- and post-calibrated, thus supplying high quality data.

All the analyses used here, with the exception of the SST CCI analysis, assimilate in situ observations (Table 1), sourced either from ICOADS (International Comprehensive Ocean-Atmosphere Data Set; Worley et al., 2005) or received over the GTS (Global Telecommunications System). These datasets include observations from the GTMBA array, meaning the dataset is not independent from the analyses, with the exception of SST CCI. However, it is still useful to use these data in context with other results and to compare findings for independent and non-independent datasets.

The GTMBA dataset was obtained from NOAA PMEL (Pacific Marine Environmental Laboratory). Observations at a depth of 1 m were used. The data have a sampling period of either 5, 10 or 60 minutes and the highest available temporal resolution was always used if multiple sampling periods were available. This means an average of daily matchups between a GTMBA buoy and an analysis should approximate the daily mean difference from the GTMBA buoy. All available observations, both daytime and nighttime, were used in order to maximise the number of matchups. No further quality control was applied to the data prior to their comparison with the SST analyses.

The number of observations available over the analysis period increases with time (Figure 7) due to changes in reporting frequency and further deployments, including the addition of the PIRATA (Prediction and Research Moored Array in the Atlantic) and RAMA (Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction) arrays, in 1998 and 2008 respectively. In order to avoid aliasing effects on the stability of the GTMBA dataset, only buoys which were available for more than 75% of the timeseries were included in this assessment. The number of observations used is also shown in Figure 7, which indicates that a large proportion of the total number of observations is retained despite this constraint. The locations of all GTMBA moorings (109 locations; indicated by the blue and red dots) and the reduced set used here (65 locations; indicated by the blue dots only) are shown in Figure 8. The buoys used are primarily from the TAO/TRITON (Tropical Atmosphere Ocean/Triangle Trans-Ocean Buoy Network) array in the Pacific Ocean, as these provide the longest records. Therefore the locations of the GTMBA observations used do not change greatly over time. However, this does mean the validation reported here is only directly applicable to the tropical Pacific Ocean and thus does not demonstrate any global homogeneity of these analysis datasets. However, alternative datasets, for example the drifting buoy network, are not known to be as accurate or stable in time as GTMBA.

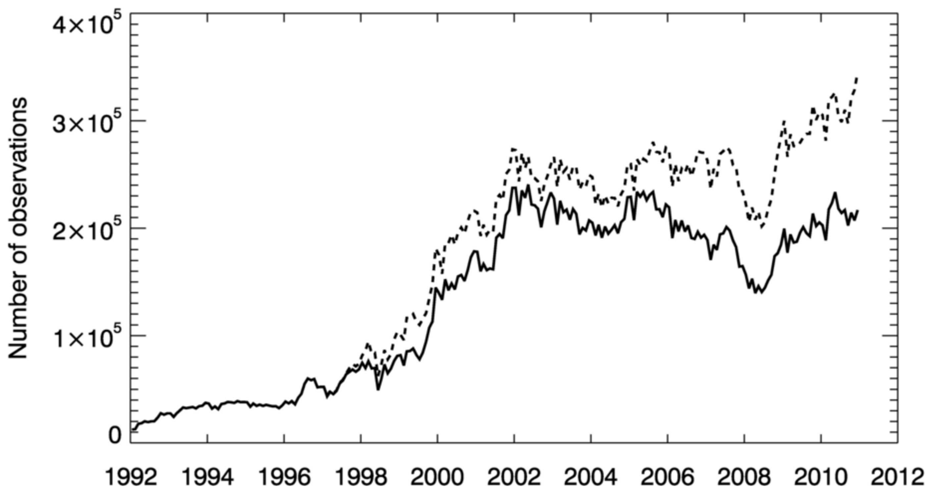


Figure 7: Monthly total number of GTMBA observations for January 1992 to December 2010. Dashed line shows all available observations, solid line those observations from buoys covering at least 75% of the timeseries (see text).

Matchups between the GTMBA observations and the SST analyses were produced by interpolating the analyses from their original grids to the observation locations, using the GMPE system in the same way as was performed for the Argo data (section 3.1.1), and with similarly negligible interpolation errors. The method used for the assessment itself is that of the GHRSST CDAF (Group for High Resolution Sea Surface Temperature Climate Data Assessment Framework), as described by Merchant et al. (2014) and summarised below.

Following the initial matchup process, the following method was performed separately for each analysis. First, the monthly median analysis-minus-GTMBA difference for each GTMBA location was calculated. This considers each location independently and avoids aliasing by periods with a greater number of matchups. For each month of the year and location, the multi-year average

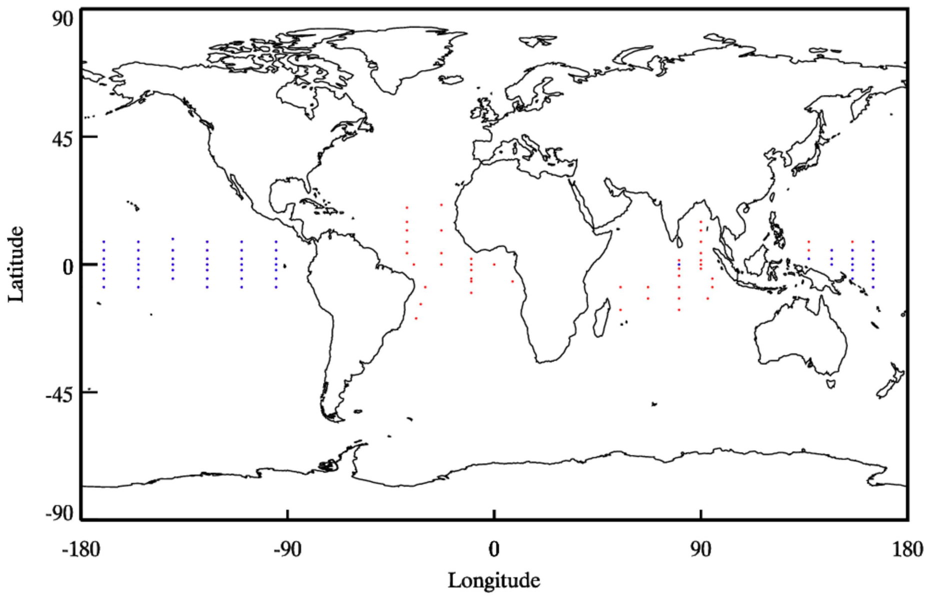


Figure 8: Nominal reference location of GTMBA buoys (red and blue dots) and the reduced set of locations (blue dots) used for validation, 1992-2010.

of the monthly median analysis-minus-GTMBA difference was then calculated. For each month the data were then deseasonalised by subtracting the multi-year average for the appropriate month of the year from each month of the timeseries. The data were deseasonalised to minimise any potential aliasing of an annual cycle in residual timeseries, following the approach of Merchant et al. (2014).

Although analysis data are available from September 1991, this validation begins in January 1992. This date was chosen both for computational efficiency reasons of working with full years, and to produce the multi-year monthly average required for deseasonalising from the same number of datapoints per month, i.e. not including part-years. Finally, the monthly mean difference across all locations was determined, producing a single analysis-minus-GTMBA timeseries for each dataset. A least squares linear fit to each timeseries of monthly mean differences was calculated and 95% confidence intervals of these fits were determined.

Deseasonalised timeseries for the monthly mean analysis-minus-GTMBA differences for each analysis are given in Figure 9, and the linear trends are given in Table 4. Trends over the full time period may not be representative of trends for shorter periods in the analysis, as can be inferred from Figure 9. Therefore, trends in Table 4 have been given for the full period (1992 - 2010), and the periods when the different ATSR-series instruments were used in the SST CCI analysis (to pick one), namely:

ATSR-1: January 1992 - May 1995

ATSR-2: July 1996 - July 2002

AATSR : August 2002 - December 2010

Note that OSTIA v1.0 and HadISST2 finish in 2007 so the AATSR period for these datasets is August 2002 - December 2007. In the gap between ATSR-1 and ATSR-2 given above, the two instruments were being swapped in the SST CCI analysis according to availability of data. Therefore this period is not included in the short-term trend calculations for simplicity. Not all the analyses use data from the ATSR series of instruments (Table 1) but the trends in analysis-minus-GTMBA difference are still calculated for the same periods to enable intercomparison between datasets.

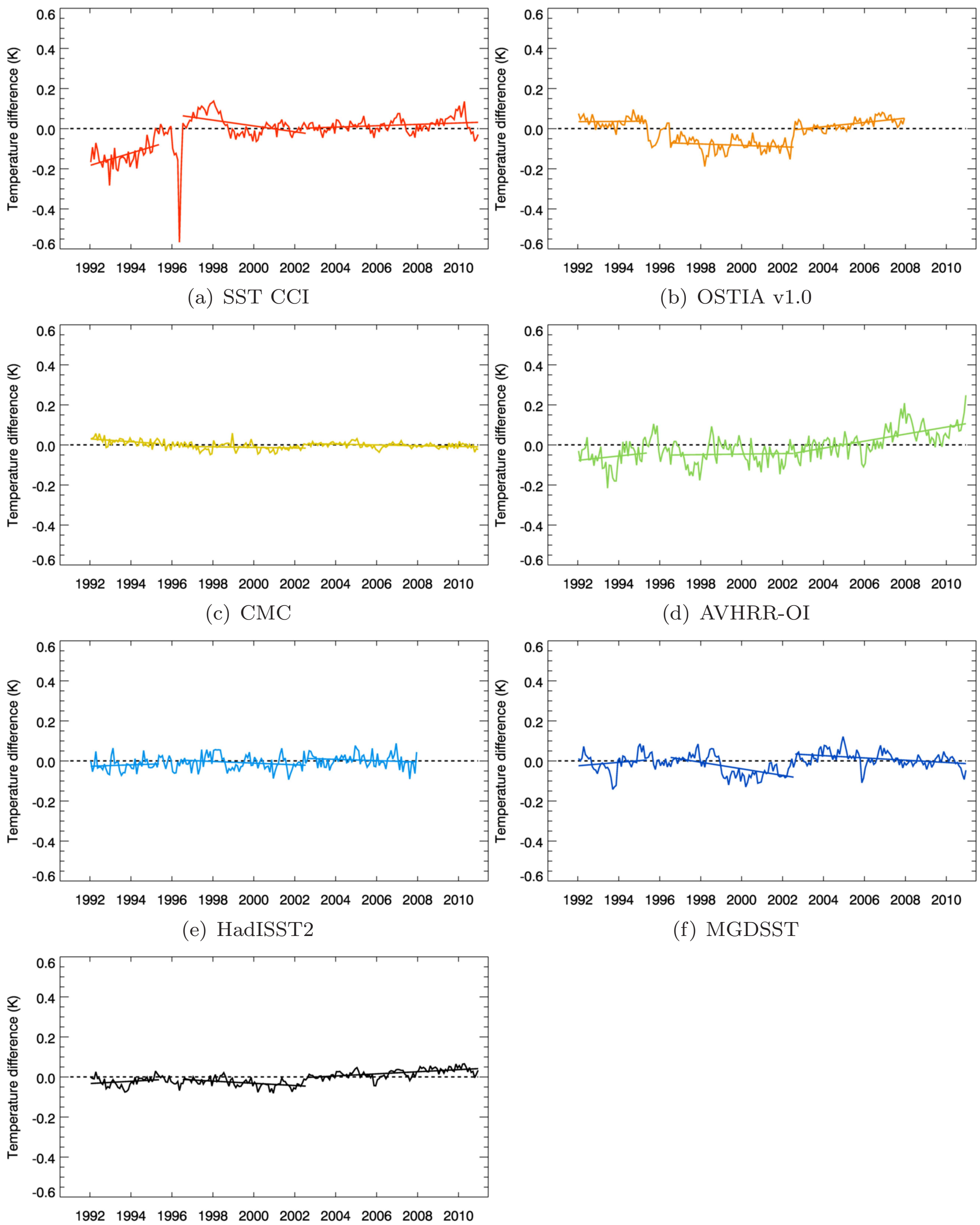
The various SST analyses in the intercomparison are intended to be valid at different depths (section 2.1) so a difference to the 1 m depth GTMBA data

Table 4: Linear trends for monthly mean analysis-minus-GTMBA differences in mK/yr for six SST analyses and their ensemble (GMPE) median. Trends given for full time period (January 1992 - December 2010, or December 2007 for OSTIA v1.0 and HadISST2), ATSR-1 period (January 1992 - May 1995), ATSR-2 period (July 1996 - July 2002) and AATSR period (August 2002 - December 2010, or December 2007 for OSTIA v1.0 and HadISST2). Quoted uncertainties on trends are 95% confidence intervals.

Analysis	Full period	ATSR-1 period	ATSR-2 period	AATSR period
SST CCI	8.0 ± 1.7	30.7 ± 15.7	-14.5 ± 5.7	3.4 ± 2.8
OSTIA v1.0	1.1 ± 1.9	0.7 ± 8.1	-3.5 ± 4.5	10.6 ± 3.5
CMC	-1.0 ± 0.4	-7.1 ± 5.3	-1.4 ± 2.4	-1.9 ± 1.0
AVHRR-OI	7.8 ± 1.5	10.8 ± 18.8	1.3 ± 7.4	17.6 ± 4.3
HadISST2	1.5 ± 1.0	3.5 ± 10.1	-4.2 ± 4.8	-3.2 ± 5.2
MGDSST	1.0 ± 1.1	8.6 ± 15.6	-16.4 ± 5.5	-5.7 ± 2.8
GMPE median	3.8 ± 0.6	5.6 ± 7.8	-5.5 ± 2.5	5.3 ± 1.5

is expected. However, any mean bias is removed by the deseasonalisation approach carried out as part of the CDAF stability method (Figure 9). It is noted that despite the non-independence of most of the analyses from the reference GTMBA dataset there is still significant variation in the trends found (Figure 9).

Trends in CMC for each of the short-term periods are very similar to each other and very small (Figure 9, Table 4). This indicates the CMC reanalysis is temporally homogeneous. HadISST2 also shows good results, with small trends which are consistent in magnitude between the ATSR periods. The HadISST2 trend is smaller than for CMC in the ATSR-1 period, although the error on the CMC trend is around half that of the error on the HadISST2 trend for the whole timeseries. Trends for the ensemble median of all the analyses, the GMPE median, are also small and fairly consistent between ATSR periods. As the GTMBA data are not independent from the CMC, HadISST2 and GMPE median products, the small trends may be related to a high weighting given to



(a) SST CCI

(b) OSTIA v1.0

(c) CMC

(d) AVHRR-OI

(e) HadISST2

(f) MGDSST

(g) GMPE median

Figure 9: Monthly mean deseasonalised analysis-minus-GTMBA differences with linear fits for different ATSR periods (see text). Only SST CCI is independent from GTMBA.

the GTMBA observations in the analyses, which continues uniformly throughout the time period. However, as analysis-minus-Argo differences for 2003-2010 in the tropical Pacific are also good for CMC, HadISST2 and the GMPE median compared to other datasets (Figure 5), this may indeed reflect a high degree of homogeneity, i.e. datasets with the smallest differences to Argo also have the smallest differences to other reference datasets. If the reference dataset is stable over time, then so is the analysis.

The stability of OSTIA v1.0 in the tropical Pacific is clearly affected by a lack of homogenisation in the ATSR-series data used, which has introduced jumps in the timeseries of analysis-minus-GTMBA data (Figure 9). However, the magnitude of the trends in the individual ATSR periods themselves are small. CMC has presumably avoided similar large jumps despite using the same ATSR dataset as OSTIA v1.0 by bias-correcting all the ATSR data to in situ (Table 1). AVHRR-OI has no large jumps in the timeseries, but a change in the magnitude of the trend of differences to GTMBA occurs in 2006 when the AVHRR data source changes from Pathfinder (Kilpatrick et al., 2001) to operational NAVO (U.S. Naval Oceanographic Office) data leading to a departure of the analysis from the reference data used here.

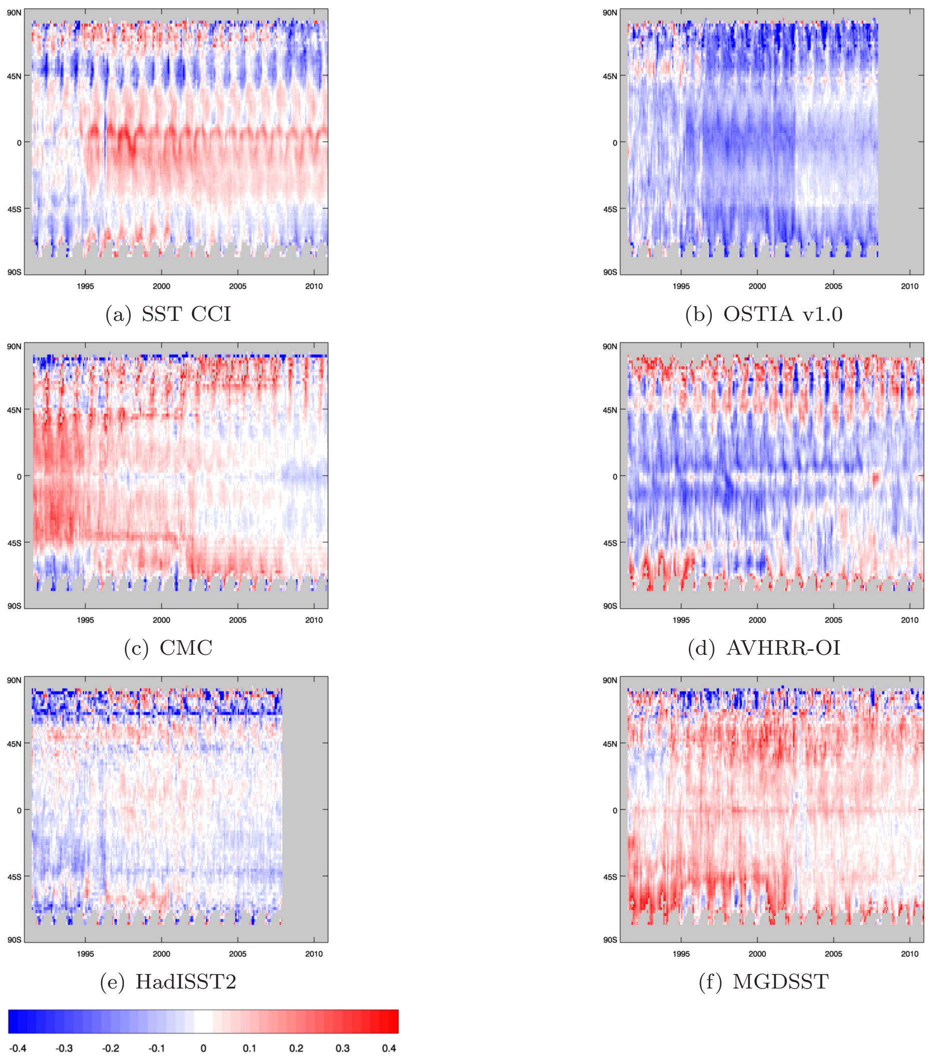
The trend for the SST CCI analysis compared to GTMBA is much larger than for the other analyses in the ATSR-1 period. In the ATSR-2 period it is marginally better than MGDSST only, but during the AATSR period the relative magnitude of the trend improves to become the third smallest (Table 4), despite being the only analysis independent of the GTMBA dataset. This indicates the comparatively large trends during the earlier periods are not solely due to the independence of the data. The reduced performance of the SST CCI analysis during the lifetime of the ATSR-1 instrument is likely due to the residual effects on SST retrieval of the Pinatubo eruption (e.g. Reynolds, 1993) and the loss of the 3.7 μm channel (e.g. Murray et al., 1998). The SST CCI analysis is the only dataset included here not to perform any bias-correction of the ATSR-1 data. Although HadISST2 uses ATSR-1 as a reference (Table 1) this only applies to the period when 3-channel retrievals were available, so much of the data are not used.

3.3. Assessment of SST analyses using the ensemble (GMPE) median

3.3.1. SST analysis anomaly to the GMPE median

With the exception of Argo, there is no global in situ dataset independent of all the SST analyses. In order to gain some insight into the relative performance of the analyses for time periods before the Argo data became available, comparisons of the anomaly of each analysis to the ensemble median have been made. The ensemble median was produced using the GMPE (Group for High Resolution SST (GHRSST) Multi-Product Ensemble) system, using the method described in section 2.2. The monthly mean anomaly to the GMPE median of each analysis is shown by latitude on a 2x2 degree grid for the period September 1991 to December 2010 in Figure 10. This method has an advantage over using observations as a reference by allowing comparisons at all latitudes (excluding ice-covered regions) instead of solely in data-rich areas and time periods.

All of the analyses show some seasonal anomalies to the GMPE median



(a) SST CCI

(b) OSTIA v1.0

(c) CMC

(d) AVHRR-OI

(e) HadISST2

(f) MGDSST

Figure 10: "Hovmöller" plots of monthly mean anomaly by latitude for six SST analyses to their ensemble (GMPE) median (analysis-minus-GMPE median) in K for 1991-2010. Areas with no data shown in grey. For reference, the ATSR-1 period is January 1992 to May 1995, ATSR-2 is July 1996 to July 2002, and AATSR is August 2002 to December 2010.

(Figure 10). The SST CCI analysis has a distinct seasonal cold difference to the GMPE median at around 50N which occurs consistently throughout the whole time period. This was found to begin in the Northern Hemisphere Spring and to deepen in Summer. A similar anomaly pattern can be seen for the AVHRR-OI analysis, although of the opposite sign. This indicates the anomaly source is the AVHRR data, since this is the only common component of both analyses. As this is a seasonal feature, the cold difference for SST CCI does not appear as strong in Figure 4 on comparison to Argo as does the more persistent warm bias in the tropics. The tropical difference is seen in Figure 10(a), and also has a seasonal component, but is smaller than that seen at 50N. The anomaly of SST CCI to the GMPE median also varies in the time periods when the different instruments of the ATSR series are used (section 3.2, Figure 10(a)). The tropical warm difference is largest when the ATSR-2 data are used (June 1995 - December 1995, July 1996 - July 2002), smaller in the AATSR period (July 2002 - December 2010) and does not appear when the ATSR-1 data are used (September 1991 - May 1995, January 1996 - June 1996). A distinct cold anomaly appears in the tropics from mid-May 1996 to early June 1996 and has been attributed to a decline in performance of the ATSR-1 instrument at the end of its life (Corlett et al., 2014).

OSTIA v1.0 shows three distinct periods of difference to the GMPE median (Figure 10(b)) seen at all latitudes and which correspond to the use of ATSR series data. This demonstrates the analysis is not homogeneous over the whole timeseries, but within these periods the difference to the GMPE median is consistent (Figure 10(b), see also section 3.2 which indentified a similar result compared to moored buoys for OSTIA v1.0 in the tropics).

In the ATSR-1 period, the CMC reanalysis has a warm anomaly to the GMPE median in the tropics (extending to 45N and S; Figure 10(c)). The difference of this analysis to the GMPE median for the ATSR-2 period is smaller in magnitude than for the ATSR-1 period, and in the AATSR period it is closer again to the GMPE median. AVHRR-OI and MGDSST do not include data from the ATSR series of instruments so do not show these same patterns (Figures 10(d),10(f)). They both however become closer to the GMPE median, particularly in the Southern Hemisphere, towards the end of the timeseries. Although it uses ATSR data, HadISST2 does not show distinct boundaries for the ATSR periods (Figure 10(e)), illustrating the homogeneity of the dataset, as previously demonstrated in section 3.2 for the tropics.

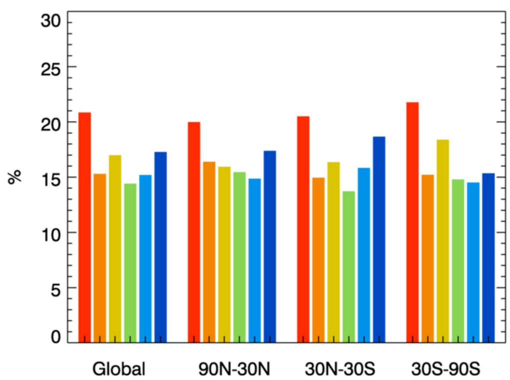
3.3.2. SST analysis contribution to the GMPE median

The GMPE median is the ensemble median of all the contributing analyses on a gridbox by gridbox basis, after they have been regridded to a $1/4^o$ grid. If an analysis is the median its contribution in that gridbox is counted as 1. If there are an even number of analyses, the mean of the two centre analyses is taken and their respective contribution to the GMPE median is counted as 0.5. Figure 11 is a summary of the contribution of each analysis to the GMPE median on a gridbox basis in various latitude bands, for the three periods of the ATSR-series instruments:

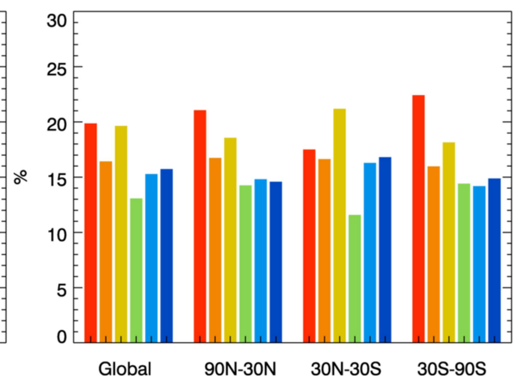
ATSR-1: January 1992 - May 1995

ATSR-2: July 1996 - July 2002

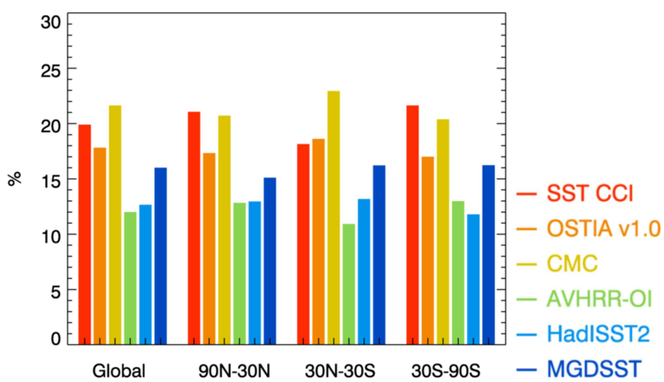
AATSR : August 2002 - December 2007



(a) ATSR-1 period (January 1992 to May 1995)



(b) ATSR-2 period (July 1996 to July 2002)



(c) AATSR period (August 2002 to December 2007)

Figure 11: Percentage gridbox contribution of different SST analyses to their ensemble (GMPE) median.

Two of the analyses finish in 2007 (HadISST2 and OSTIA v1.0) so all results are given up to and including that year in the AATSR period to aid intercomparison. The contributions are calculated as a percentage of the total number of gridboxes in that latitude band.

Those analyses with the smallest global and regional standard deviations of differences to Argo, CMC and SST CCI (Section 3.1, Figures 3 and 5), contribute the greatest percentage of gridboxes to the GMPE median (Figure 11). AVHRR-OI and HadISST2 generally have the smallest number of contributions to the median (Figure 11). These are also the analyses which have the largest global and regional standard deviations of differences to Argo (Section 3.1, Figures 3 and 5 respectively). This result indicates the level of contribution to the GMPE median can be used to give a general idea of the quality of an analysis relative to others in periods where no validation data are available.

In the ATSR-1 period, for the northern and southern extratropics (90N-30N and 30S-90S) and the tropics themselves (30N-30S), the SST CCI analysis makes the largest number of contributions to the median (Figure 11(a)). These are wide latitude bands, so the seasonal temperature cycling centred on 50N in SST CCI (Figure 10(a)) does not dominate these statistics. For the ATSR-2 period, SST CCI still has the largest percentage of contributions to the median in the northern and southern extratropics (Figure 11(b)). However, in the tropics, where the SST CCI mean and standard deviation of differences to Argo are poorer than for other regions (e.g. Table 3) the contribution to the GMPE median is smaller, and CMC has the highest percentage of contributions (Figure 11(b)). For the AATSR period, SST CCI has only the third highest contribution to the median in the tropics, behind CMC and OSTIA v1.0, with MGDSST not far behind SST CCI (Figure 11(c)). In the northern and southern extratropics in the AATSR period SST CCI still has the largest number of contributions to the median, but CMC is very close.

Although overall the CMC and SST CCI analyses make up the largest number of contributions to the GMPE median, neither analysis contributes more than 24% of the gridboxes in any of the three latitude bands investigated (Figure 11). Therefore, the GMPE median is not dominated by any one analysis, but is made up of significant contributions from all the analyses.

3.4. Feature resolution

Accurate feature resolution in SST analyses is important due to its influence on aspects of atmospheric forecasting (e.g. Maloney & Chelton, 2006). Feature resolution, which is not necessarily related to grid size but rather analysis parameters and data limitations (Reynolds & Chelton, 2010), can be determined and quantified using spectral analysis techniques (e.g. Reynolds et al., 2013). However, a full investigation into this aspect of the contributing SST analyses is beyond the scope of this work. Instead, following Martin et al. (2012), horizontal SST gradients will be examined for each analysis, where a greater number and magnitude of the gradients illustrates the ability of the analysis to capture high-resolution features.

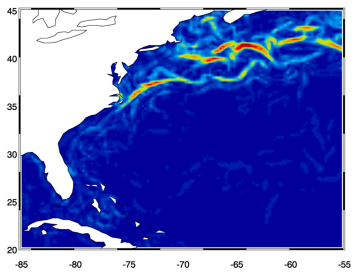
Horizontal gradients were calculated for each analysis on its native grid, by finding the vector sum of SST gradients in the North-South and East-West directions for each grid point. This was only calculated when all four of the neighboring North-South and East-West points were available, i.e. when there was no land or ice in the immediate proximity. The gradients for each analysis were then interpolated to the $1/4^o$ GMPE grid before plotting.

Figure 12 shows horizontal SST gradients in the Gulf Stream region on 01

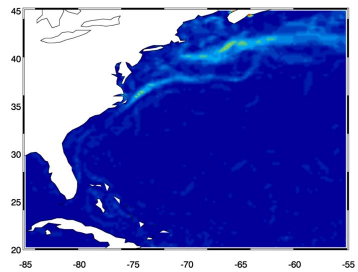
July 2007 as an example date for the six contributing analyses and their ensemble (GMPE) median. Animations of the gradients throughout the timeseries for all the analyses were visually assessed, and indicate the features seen on the example date shown in Figure 12 are coherent and persistent. This means they are likely to be an accurate representation of fronts and unlikely to be noise. All of the products are able to capture the main SST features of the region, but show differing levels of smoothness. Figure 12 is representative of the relative smoothness of features between different analyses over the whole timeseries.

The grid resolution for each of the analyses is given on Figure 12 and illustrates that the feature resolution capability of each analysis relative to the other analyses is not necessarily related to its grid size. For example, the $1/20^o$ OSTIA v1.0 analysis (Figure 12(b)) actually has the smoothest gradients and there is notable variation in feature resolution between those analyses on a $1/4^o$ grid (Figures 12(d)-(g)). Nevertheless, the sharpest gradients are seen for the SST CCI analysis (Figure 12(a)), which clearly utilises more of the potential of the $1/20^o$ grid than does the OSTIA v1.0 analysis (Figure 12(b)).

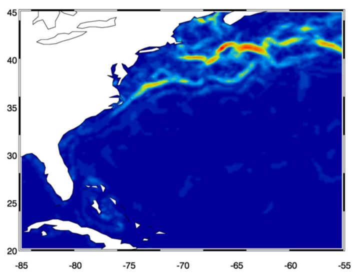
SST CCI uses an upgraded version of the OSTIA system used to produce the OSTIA v1.0 analysis, including updates to the background error covariances and an increase in the number of iterations performed by the analysis scheme (Roberts-Jones et al., 2016). CMC (Figure 12(c)) also compares well against the other analyses in terms of feature resolution. MGDSST also has sharp gradients although some noise can be discerned, manifesting as angular shapes which can be seen within the SST features (Figure 12(f); may require



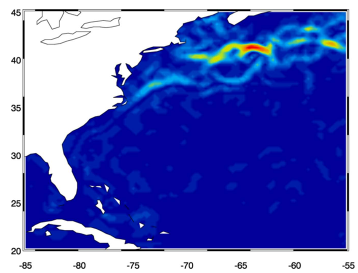
(a) SST CCI (1/20^o)



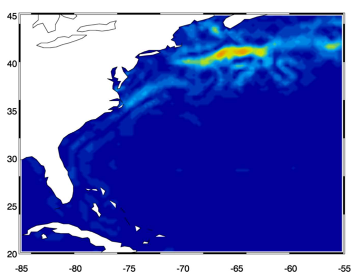
(b) OSTIA v1.0 (1/20^o)



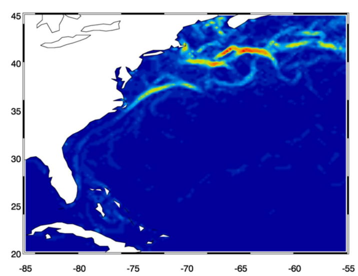
(c) CMC (1/5^o)



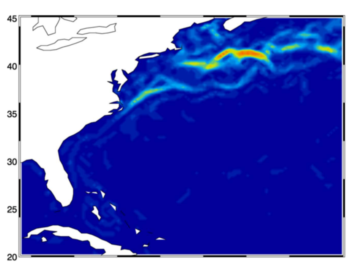
(d) AVHRR-OI (1/4^o)

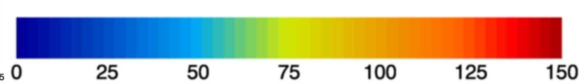


(e) HadISST2 (1/4^o)



(f) MGDSST (1/4^o)





(g) GMPE median (1/4^o)

Figure 12: Horizontal SST gradients (vector sum of North-South and East-West differences) given in mK per km, on 01 July 2007 for the Gulf Stream region. Shown are six analyses and their ensemble (GMPE) median, with their grid resolutions.

zooming in on electronic copy).

The GMPE median (Figure 12(g)) has smoother features than SST CCI and CMC, but given that it is an ensemble median of 6 different analyses it captures features well. Despite the source of the GMPE median potentially varying from gridbox to gridbox, artificial gradients and noise do not appear to be introduced. SST gradients for the GMPE median are spatially coherent, with sharper features than some of the contributing datasets (cf. Figure 12). A similar result was also found by Martin et al. (2012) in an assessment of the GMPE median for near-real-time analyses.

4. Summary and Conclusions

Six global, gridded, daily SST analyses of at least 20 years in length have been intercompared: ESA SST CCI (European Space Agency Sea Surface Temperature Climate Change Initiative) analysis long-term product v1.0, MyOcean OSTIA (Operational Sea Surface Temperature and Ice Analysis) reanalysis v1.0, CMC (Canadian Meteorological Center) 0.2 degree analysis, AVHRR (Advanced Very High Resolution Radiometter)_ONLY Daily 1/4 degree OISST (Optimal Interpolation Sea Surface Temperature) v2.0, HadISST2.1.0.0 (Hadley Centre Ice and Sea Surface Temperature) realisation 396 and MGDSST (Merged satellite and in situ data Global Daily Sea Surface Temperature) analysis. A seventh SST product, an ensemble median of all six analyses, has been produced using the GMPE (Group for High Resolution Sea Surface Temperature Multi-Product Ensemble) system.

The performance and spatial homogeneity of the seven datasets has been assessed for the period 2003-2010 using near-surface Argo data, which are independent from all the analyses. The temporal homogeneity of the analyses has been investigated for the period 1992-2010 using a long and stable timeseries of GTMBA (Global Tropical Moored Buoy Array) observations. Comparisons to the GMPE median provide a method for assessment of both spatial and temporal homogeneity. The feature resolution for all the products has also been compared using horizontal SST gradients. Table 5 is a summary of all the results from these investigations. The rankings 1 to 3 (where 1 is best) for different criteria given in the table are intended to give an idea of the relative performance of each of the analyses and are based on global and regional results where applicable. Particular characteristics of different analyses have also been highlighted.

None of the analyses performs badly. The rankings in Table 5 are therefore not intended to be added up and used as an overall "score" for performance as the intended use of the analysis should still inform which will be the most suitable. For example, if a long-term, temporally homogeneous analysis is required, with reduced emphasis on feature resolution, the user might select HadISST2. If a foundation temperature is required, with good all-round performance in temporal and spatial homogeneity, standard deviation, bias and feature resolution criteria, MGDSST might be selected. If a daily mean temperature at 20 cm depth, with excellent feature resolution is required, then SST CCI would be the most suitable product. Thus the choice of analysis is dependent on which

Table 5: Summary of strengths and weaknesses of different analyses. Relative ranks are 1, 2 or 3; 1 being best. Ranking of standard deviation of differences to Argo assessed from Figs 3 and 5(b), mean difference from Figs 5(a) and 6. Temporal homogeneity assessed from Figs 9 and 10. Spatial homogeneity assessed from Figs 4, 6 and 10. Feature resolution assessed from Fig 12. *Now available 1961-2010.

Analysis	Relative rank	Key strengths and weaknesses compared to other analyses
SST CCI	1	small standard deviation of difference to Argo
daily mean, 20 cm	2	moderate mean difference to Argo
1991 - 2010	3	reduced temporal homogeneity
	2	moderate spatial homogeneity
	1	good feature resolution
		independent from in situ observations
OSTIA v1.0	2	moderate standard deviation of difference to Argo
foundation	3	larger mean difference to Argo
1985 - 2007	3	reduced temporal homogeneity
	3	reduced spatial homogeneity
	3	reduced feature resolution
CMC	1	small standard deviation of difference to Argo
1 m	1	small mean difference to Argo
1991 - 2011	1	good temporal homogeneity
	1	good spatial homogeneity
	1	good feature resolution
		includes microwave data
AVHRR-OI	3	larger standard deviation of difference to Argo
daily mean (all data)	3	larger mean difference to Argo
1981 - present	2	moderate temporal homogeneity
	3	reduced spatial homogeneity
	2	moderate feature resolution
		independent from ATSRs
		single sensor product
HadISST2	3	larger standard deviation of difference to Argo
20 cm	2	moderate mean difference to Argo
1961 - 2007*	1	good temporal homogeneity
	1	good spatial homogeneity
	3	reduced feature resolution
		uncertainty information from multiple realisations
		very long time period
MGDSST	2	moderate standard deviation of difference to Argo
foundation	2	moderate mean difference to Argo
1982 - 2011	2	moderate temporal homogeneity
	2	moderate spatial homogeneity
	2	moderate feature resolution
		independent from ATSRs
		includes microwave data
GMPE median	1	small standard deviation of difference to Argo
No specific depth	1	small mean difference to Argo
1991 - 2007 (6 products)	2	moderate temporal homogeneity
2008 - 2010 (4 products)	1	good spatial homogeneity
	2	moderate feature resolution
		source potentially varies from gridbox to gridbox

criteria are most important to the proposed application.

Clearly CMC performs extremely well relative to the other analyses (Table 5), and is equivalent in performance to the GMPE median in terms of standard deviation and mean of the difference to independent Argo observations. In a previous study using NRT (near-real-time) data, Martin et al. (2012) found that the GMPE median had a smaller standard deviation on comparison to Argo than any of its component analyses (although more recently improvements to NRT products have been closing the gap, see `http://ghrsst-pp.metoffice.com/pages/latest_analysis/sst_monitor/argo`). However, as the GMPE median is constructed from different analyses on a gridbox by gridbox basis, spatial or temporal discontinuities could potentially be introduced into the SST field. Despite the similar results, the GMPE median is not composed mainly of the CMC analysis but has been shown to be made up of significant contributions from all the analyses.

The analyses with the largest contributions to the GMPE median are those with the smallest standard deviations of differences to Argo. This result means that the relative contributions of an analysis to the ensemble median could be used to provide a general idea of the accuracy of an analysis relative to others in periods when no reference data are available. Seasonal anomalies to the GMPE median were identified for all analyses, occurring throughout the time period and demonstrating that comparison to the GMPE median also allows an in-depth assessment of analysis quality for all regions and time periods. Indeed, the patterns seen in the Hovmöller plot of the SST CCI analysis anomaly to the GMPE median (Figure 10(a)) are qualitatively similar to those seen in Hovmöller plots of the SST CCI analysis anomaly to drifter data in Corlett et al. (2014).

This study has provided an assessment of the relative performance of currently available long-term, global, gridded SST products. As newly-reprocessed input data become available, the selection of global SST analysis products will be updated. For example, ongoing work will extend the SST CCI analysis to cover a period of more than 30 years as part of the CCI Phase 2 project (`http://www.esa-sst-cci.org`). The complete dataset is expected to be released in 2019.

The aspiration of the SST community is to move away from an empirical approach to SST retrievals and reanalyses, become completely independent of in situ measurements, and use a physics-based approach. Among the analyses examined here only SST CCI is independent of in situ observations, but did not perform well during the early period. This underlines the challenge with using older-generation satellite data and correcting for biases. SST CCI did perform well in the more recent decade demonstrating the feasibility of a more physical approach as a way forward. However, for extended timeseries using older satellites, quality of the satellite analyses will likely remain dependent on in situ data.

It is envisioned that updated intercomparison studies will be useful in the future, in order to continue to provide users with the information needed to make an informed choice regarding the most appropriate analysis for their application.

Acknowledgements

This work was carried out as part of validation activities for the ESA SST CCI project. Chris Atkinson is acknowledged for providing the matchups of drifter and Argo observations from the HadIOD database used in Figure 1. Three anonymous reviewers are also acknowledged for their helpful comments, resulting in improvements to the paper.

Data Access

The datasets used in this study can be freely accessed from the following locations (may require registration):

ESA SST CCI analysis long-term product v1.0: `http://catalogue.ceda.ac.uk/uuid/916986a220e6bad55411d9407ade347c`

MyOcean OSTIA reanalysis v1.0: `http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw\&view=details\&product_id=SST_GLO_SST_L4_REP_OBSERVATIONS_010_011`

CMC 0.2 degree: `https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-L4-GLOB-v2.0`

AVHRR ONLY Daily 1/4 degree OISST v2.0: `https://podaac.jpl.nasa.gov/dataset/AVHRR_OI-NCEI-L4-GLOB-v2.0`

HadISST2.1.0.0: 1^o, 5-day and interpolated $1/4^o$, daily products will be made available from `https://www.metoffice.gov.uk/hadobs/hadisst2/data/download.html`

MGDSST: `https://ds.data.jma.go.jp/gmd/goos/data/rrtdb/file_list.`

php#a0

GMPE median from long-term analysis inputs: http://catalogue.ceda.ac.uk/uuid/e0659b01259145c8bfb0de6eb12c2690

GMPE median from near-real-time analysis inputs: http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw\&view=details\&product_id=SST_GLO_SST_L4_NRT_OBSERVATIONS_010_005

References

Atkinson, C. P., Rayner, N. A., Kennedy, J. J., & Good, S. A. (2014). An integrated database of ocean temperature and salinity observations. *J. Geophys. Res.*, *119*, 7139–7163. DOI: 10.1002/2014JC010053.

Banzon, V., Smith, T. M., Chin, T. M., Liu, C., & Hankins, W. (2016). A long-term record of blended satellite and in situ sea-surface temperature for climate monitoring, modeling and environmental studies. *Earth Syst. Sci. Data*, *8*, 165–176. DOI: doi:10.5194/essd-8-165-2016.

Brasnett, B. (2012). *A 20-year Reanalysis of Sea Surface Temperature*. Report CMC. Available on request from the author at bruce.brasnett@gmail.com.

Corlett, G. K., Atkinson, C., Rayner, N., Good, S., Fiedler, E., McLaren, A., Hoeyer, J., & Bulgin, C. (2014). *Product Validation and Intercomparison Report (PVIR)*. SST_CCI-PVIR-UoL-201 Issue 1 ESA. URL: http://www.esa-sst-cci.org/PUG/documents.

Dash, P., Ignatov, A., Martin, M., Donlon, C., Brasnett, B., Reynolds, R., Banzon, V., Beggs, H., Cayula, J.-F., Chao, Y., Grumbine, R., Maturi, E., Harris, A., Mittaz, J., Sapper, J., Chin, T. M., Vazquez-Cuervo, J., Armstrong, E. M., Gentemann, C., Cummings, J., Piolle, J.-F., Autret, E., Roberts-Jones, J., Ishizaki, S., Hoyer, J. L., & Poulter, D. (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.

Donlon, C. J., Gentemann, C., & Nykjaer, L. (2004). Using SST measurements from microwave and infrared satellite measurements. *Intl. J. Rem. Sens.*, *25*, 1331–1336.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., & Wimmer, W. (2012). The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Rem. Sens. Env.*, *116*, 140–158.

Donlon, C. J., Minnett, P. J., Gentemann, C., Nightingale, T. J., Barton, I. J., Ward, B., & Murray, M. J. (2002). Toward improved validation of satellite sea surface skin temperature measurements for climate research. *J. Climate*, *15*, 353–369.

Fiedler, E. K., McLaren, A., Merchant, C. J., & Donlon, C. (2015). *ESA Sea Surface Temperature Climate Change Initiative (ESA SST CCI): GHRSST Multi-Product ensemble (GMPE)*. Dataset NERC Earth Observation Data Centre. DOI: 10.5285/7BAF7407-2F15-406C-8F09-CB9DC10392AA.

Gentemann, C., Minnett, P., & Ward, B. (2009). Profiles of ocean surface heating (POSH): A new model of upper ocean diurnal warming. *J. Geophys. Res.*, *114*. C07017, DOI: 10.1029/2008JC004825.

Good, S. A., Martin, M. J., & Rayner, N. A. (2013). EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res.*, *118*, 6704–6716. DOI: 10.1002/2013JC009067.

Kennedy, J. J., Rayner, N. A., Millington, S. C., & Saunby, M. (2018). The Met Office Hadley Centre Sea Ice and Sea-Surface Temperature data set, version 2, Part 2: Sea Surface Temperature Analysis. *J. Geophys. Res. Atmos.*, . (in prep.).

Kilpatrick, K. A., Podesta, G. P., & Evans, R. (2001). Overview of the NOAA/NASA Advanced Very High Resolution Radiometer Pathfinder algorithm for sea surface temperature and associated matchup database. *J. Geophys. Res.*, *106*, 9179–9197.

Kurihara, Y., Sakurai, T., & Kuragano, T. (2006). Global daily sea surface temperature analysis using data from satellite microwave radiometer, satellite infrared radiometer and in situ observations. *Weather Bull.*, *73*, 1–18. (in Japanese).

Maloney, E. D., & Chelton, D. B. (2006). An assessment of the sea surface temperature influence on surface wind stress in numerical weather prediction and climate models. *J. Climate*, *19*, 2743–2762.

Martin, M., Dash, P., Ignatov, A., Banzon, V., Beggs, H., Brasnett, B., Cayula, J.-F., Cummings, J., Donlon, C., Gentemann, C., Grumbine, R., Ishizaki, S., Maturi, E., Reynolds, R. W., & Roberts-Jones, J. (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.

McLaren, A., Fiedler, E., Roberts-Jones, J., & Martin, M. (2014). *Global Ocean OSTIA Sea Surface Temperature Reprocessing, SST-GLO-SST-L4-REP-OBSERVATIONS-010-011*. Quality Information Document Copernicus Marine Enviroment Monitoring Service. URL: `http://cmems-resources.cls.fr/documents/QUID/CMEMS-OSI-QUID-010-011.pdf`.

McPhaden, M. J., Ando, K., Bourles, B., Freitag, H. P., Lumpkin, R., Masumoto, Y., Murty, V. S. N., Nobre, P., Ravichandran, M., Vialard, J., Vousden, D., & Yu, W. (2009). The Global Tropical Moored Buoy Array. In *OceanObs09: Sustained Ocean Observations and Information for Society*. ESA Publication WPP-306 volume 2. 10.5270/OceanObs09.cwp.6.

Merchant, C. J., Embury, O., Rayner, N. A., Berry, D. I., Corlett, G., Lean, K., Veal, K. L., Kent, E. C., Llewellyn-Jones, D., Remedios, J. J., & Saunders, R. (2012). A twenty-year independent record of sea surface temperature for climate from Along Track Scanning Radiometers. *J. Geophys. Res.*, *117*. DOI: 10.1029/2012JC008400.

Merchant, C. J., Embury, O., Roberts-Jones, J., Fiedler, E. K., Bulgin, C., Corlett, G. K., Good, S., McLaren, A., Rayner, N., Morak-Bozzio, S., & Donlon, C. (2014). Sea surface temperature datasets for climate applications from Phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geoscience Data Journal*, *1*, 179–191. DOI: 10.1002/gdj3.20.

Merchant, C. J., Mittaz, J., & Corlett, G. K. (2014). *Climate Data Assessment Framework*. CDR-TAG-CDAF Version 1.0.5 GHRSST. URL: https://www.ghrsst.org/wp-content/uploads/2018/01/CDR-TAG_CDAF-v1.0.5.pdf

Murray, M. J., Allen, M. R., Mutlow, C. T., Zavody, A. M., Jones, M. S., & Forrester, T. N. (1998). Actual and potential information in dual?view radiometric observations of sea surface temperature from ATSR. *J. Geophys. Res. Oceans*, *103, C4*, 8153–8165. DOI: 10.1029/97JC02180.

Oka, E., & Ando, K. (2004). Stability of temperature and conductivity sensors of Argo profiling floats. *Journal of Oceanography*, *60*, 253–258.

Poulter, D. J. S., Donlon, C. J., & Robinson, I. S. (2008). Real time analysis with the Medspiration High Resolution Diagnostic Data Set. In *2nd MERIS/(A)ATSR User Workshop*. ESA.

Rayner, N. A., Kennedy, J. J., Smith, R. O., & Titchner, H. A. (2018). The Met Office Hadley Centre Sea Ice and Sea-Surface Temperature data set, version 2, Part 3: The Combined Analysis. *J. Geophys. Res. Atmos.*, . (in prep.).

Reynolds, R. W. (1993). Impact of Mount Pinatubo aerosols on satellite-derived sea surface temperatures. *J. Climate*, *6*, 768–774.

Reynolds, R. W. (2009). *What's New in Version 2*. Report NCDC. URL: http://www.ncdc.noaa.gov/sst/papers.

Reynolds, R. W., & Chelton, D. B. (2010). Comparisons of daily sea surface temperature analyses for 2007-08. *J. Climate*, *23*, 3545–3562.

Reynolds, R. W., Chelton, D. B., Roberts-Jones, J., Martin, M. J., Menemenlis, D., & Merchant, C. J. (2013). Objective determination of feature resolution in two sea surface temperature analyses. *J. Climate*, *26*, 2514–2533.

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., & Schlax, M. G. (2007). Daily high resolution blended analyses for sea surface temperature. *J. Climate*, *20*, 5473–5496.

Roberts-Jones, J., Bovis, K., Martin, M., & McLaren, A. (2016). Estimating background error covariance parameters and assessing their impact in the OSTIA system. *Rem. Sens. Env.*, *176*, 117–138. DOI: 10.1016/j.rse.2015.12.006.

Roberts-Jones, J., Fiedler, E., & Martin, M. (2012). Daily, global, high-resolution SST and sea ice reanalysis for 1985-2007 using the OSTIA system. *J. Climate*, *25*, 6215–6232.

Roberts-Jones, J., Fiedler, E., Martin, M., & McLaren, A. (2013). *Improvements to the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system*. SST_CCI-REP-UKMO-001 Issue C ESA. URL: http://www.esa-sst-cci.org/PUG/documents.

Taberner, M., Shutler, J., Walker, P., Poulter, D., Piolle, J.-F., Donlon, C.,

& Guidetti, V. (2013). The ESA Felyx High Resolution Diagnostic Data Set system design and implementation. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. International Society for Photogrammetry and Remote Sensing.

Takaya, Y., Bidlot, J.-R., Beljaars, A., & Janssen, P. (2010). Refinements to a prognostic scheme of skin sea surface temperature. *J. Geophys. Res.*, *115*. C06009, DOI: 10.1029/2009JC005985.

Titchner, H. A., Rayner N. A., (2014). The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *J. Geophys. Res. Atmos.*, *119*, 2864–2889. DOI: 10.1002/2013JD020316.

Worley, S. J., Woodruff, S. D., Reynolds, R. W., Lubker, S. J., & Lott, N. (2005). ICOADS release 2.1 data and products. *Int. J. Climatol.*, *25*, 823–842.

Zeng, X., & Beljaars, A. (2005). A prognostic scheme of sea surface skin temperature for modeling and data assimilation. *Geophys. Res. Lett.*, *32*. L14605, DOI: 10.1029/2005GL023030.

List of Figures