

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

Article

Accepted Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

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 GHR SST Multi-Product Ensemble (GMPE) system. *Remote Sensing of Environment*, 222. pp. 18-33. ISSN 0034-4257 doi: 10.1016/j.rse.2018.12.015 Available at <https://centaur.reading.ac.uk/81206/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

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

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

1 Intercomparison of long-term sea surface temperature
2 analyses using the GHR SST Multi-Product Ensemble
3 (GMPE) system

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

7 ^a*Met Office, Exeter, UK*

8 ^b*NOAA/NCEI, Asheville, NC, USA*

9 ^c*Canadian Meteorological Centre, Dorval, Quebec, Canada*

10 ^d*Japan Meteorological Agency, Tokyo, Japan*

11 ^e*University of Leicester, Leicester, UK*

12 ^f*University of Reading, Reading, UK*

13 ^g*National Centre for Earth Observation, UK*

14 ^h*ESA/ESTEC (EOP-SME), Noordwijk, The Netherlands*

15 **Abstract**

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

*Corresponding author

28 suitable for their proposed application.

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

30 **1. Introduction**

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

45 Multiple realisations of SST timeseries using different data combinations and
46 techniques can not only be used to highlight problems, but can also be bene-
47 ficial by providing users with a choice of product to best suit their needs. For
48 example, climate-related applications require a homogeneous, stable timeseries
49 without the artificial temporal variability that can be introduced when including
50 non-homogenised data from additional instruments during the timeseries. How-

51 ever, the accuracy of the analysis may be improved, potentially at the expense
52 of stability, by utilising data from a wider variety of sources as they become
53 available. This sort of dataset is useful for applications such as short-range
54 model forcing and validation.

55 The aim of this study is to assess the relative strengths and weaknesses of var-
56 ious long-term SST analysis datasets. An intercomparison of the analyses will be
57 undertaken using the GMPE (Group for High Resolution SST (GHRSSST) Multi-
58 Product Ensemble) system, described by Martin et al. (2012). This system is
59 a tool that produces an ensemble median SST product from contributing SST
60 analyses as well as having the capability to generate matchups of the analyses
61 with in situ observations for validation. An important use of the GMPE median
62 product is to assess the deviations of the contributing analyses from it. The main
63 advantage of using this dataset for validation is that it provides complete and
64 consistent spatial and temporal coverage, unlike in situ reference data. Martin
65 et al. (2012) found the GMPE median product to perform better compared to
66 Argo than any of the component analyses used to generate it. A GMPE median
67 product is generated daily at the Met Office using NRT (near-real-time) SST
68 analyses as input, and is available from CMEMS (Copernicus Marine Environ-
69 ment Monitoring Service; marine.copernicus.eu). Monthly statistics of the NRT
70 input analyses compared to Argo observations are available at [http://ghrsst-](http://ghrsst-pp.metoffice.com/pages/latest_analysis/sst_monitor/argo)
71 [pp.metoffice.com/pages/latest_analysis/sst_monitor/argo](http://ghrsst-pp.metoffice.com/pages/latest_analysis/sst_monitor/argo) . Note that
72 results for NRT versions of the input analyses are not necessarily directly com-
73 parable to the long-term analysis versions of the same products assessed here,

74 owing to differences in methods and input data.

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

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

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

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

114 **2. Data and methods**

115 *2.1. Contributing datasets*

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

119 Temperature Climate Change Initiative) analysis long-term product v1.0 (re-
 120 ferred to herein as SST CCI analysis; Merchant et al., 2014), MyOcean OSTIA
 121 (Operational Sea Surface Temperature and Ice Analysis) reanalysis v1.0 (re-
 122 ferred to herein as OSTIA v1.0; Roberts-Jones et al., 2012), CMC (Canadian
 123 Meteorological Center) 0.2 degree analysis (referred to herein as CMC; Brasnett,
 124 2012), AVHRR (Advanced Very High Resolution Radiometer)_ONLY Daily 1/4
 125 degree OISST (Optimal Interpolation Sea Surface Temperature) v2.0 (referred
 126 to herein as AVHRR-OI; Reynolds et al., 2007; Reynolds, 2009; Banzon et al.,
 127 2016), HadISST2.1.0.0 (Hadley Centre Ice and Sea Surface Temperature) reali-
 128 sation 396 (referred to herein as HadISST2; Kennedy et al., 2018; Rayner et al.,
 129 2018) and MGDSSST (Merged satellite and in situ data Global Daily Sea Sur-
 130 face Temperature) analysis (Kurihara et al., 2006). Data were obtained directly
 131 from the producers, with the exception of AVHRR-OI, which was downloaded
 132 via ftp from PO.DAAC (NASA JPL Physical Oceanography Distributed Active
 133 Archive Data Center). Access locations for all the datasets are provided in the
 134 “Data Access” section at the end of this paper.

135 The SST CCI analysis was produced using different input data and an up-
 136 graded version of the OSTIA system previously used to produce the OSTIA
 137 v1.0 reanalysis. Updates to the system to produce the new analysis are de-
 138 scribed in Roberts-Jones et al. (2013). HadISST2.1.0.0 realisation 396 was ran-
 139 domly selected from the available set of 10 interchangeable realisations, which
 140 are intended to provide information about the likely spread arising from un-
 141 certainty in the measurements and reconstruction. The dataset is based on a

142 5-day, 1° resolution dataset that has been interpolated to 1-day, $1/4^\circ$ resolu-
143 tion by the data producers. HadISST2.1.0.0 was available to 2007 at the time
144 this work was conducted. It has subsequently been made available to 2010. A
145 version of the AVHRR-OI dataset which also includes microwave data is avail-
146 able (AVHRR+AMSR Daily $1/4$ degree OISST v2.0; Reynolds et al., 2007;
147 Reynolds, 2009), but this is not used in the comparisons due to the shorter
148 length of the available timeseries (just over 9 years) compared to other datasets
149 used here (at least 20 years).

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

162 Although the datasets are all “SST” products, they are intended to be valid
163 at a variety of near-surface depths, for use in different applications. The SST
164 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.

Analysis and Citation	Time period and SST depth/time	AVHRR	Infra-red sensors ATSR-series	AMSR-E	Microwave sensors TMI	WindSat	In situ	Ice source	data	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 v1.0]	[CCI, ATSR-1,2, AATSR [CCI, v1.0]	None	None	None	None	OSI 409 (1991-Oct 2009), OSI-401-a v1.2 (Oct 2009-2010)	SAF v1.1	1/20	ATSR-1,2, AATSR
MyOcean OSTIA re-analysis 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 409 v1.0	SAF 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 409 (1991-Oct 1998), CMC (Oct 1998-2011)	SAF v1.0	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-2 unspecified, 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°, 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

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

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

185 *2.2. Methods*

186 The methods used in the GMPE (Group for High Resolution SST Multi-
 187 Product Ensemble) system will be briefly described here. For further details the

188 reader is referred to Martin et al. (2012). The SST analyses are first regridded to
 189 a regular latitude-longitude, $1/4^\circ$ GMPE grid using a bilinear interpolation. An
 190 ensemble median SST (referred to herein as the “GMPE median”) and standard
 191 deviation for each grid box are calculated from the contributing analyses. The
 192 use of a median rather than a mean minimises the effect of potential outliers
 193 in the data on the ensemble value. If there are an even number of analyses,
 194 the mean of the two centre analyses is taken. The production of a median SST
 195 using all the datasets provides a new SST product that potentially has smaller
 196 errors than any of the component analyses, as was found for the GMPE median
 197 product generated using NRT SST analysis datasets as input (Martin et al.,
 198 2012).

199 When the GMPE system is run using NRT analysis datasets, the land mask
 200 and updated sea ice mask for each day are taken from the Met Office NRT OS-
 201 TIA (Operational Sea Surface Temperature and Ice Analysis) product (Donlon
 202 et al., 2012). Here they will be taken from the SST CCI analysis, also pro-
 203 cessed at the Met Office using the OSTIA system. The sea ice data used for
 204 the SST CCI product is sourced from EUMETSAT OSI SAF products OSI-409
 205 v1.1 (used for 1991 - October 2009) and OSI-401-a v1.2 (October 2009 - 2010)
 206 (Table 1). Using a linear interpolation method, files were created to fill gaps
 207 in the OSI SAF timeseries using the method described in Roberts-Jones et al.
 208 (2013). The data were regridded from the native 10 km polar stereographic grid
 209 to the regular latitude-longitude $1/20^\circ$ OSTIA grid and bilinear interpolation
 210 was used to perform spatial filling around coasts. For use in the GMPE system,

211 the sea ice was then regridded to the same $1/4^\circ$ grid used for SST.

212 **3. Intercomparison of analyses**

213 *3.1. Validation of SST analyses using independent Argo data*

214 *3.1.1. Argo matchup statistics*

215 Temperature data from Argo profiling floats have been used here for vali-
216 dation of the six SST analyses and their ensemble (GMPE) median. The Argo
217 dataset is suitable for use as a reference for validation since it is both accurate
218 and stable (Oka & Ando, 2004). It is also the only in situ dataset from which
219 SST analysis products are kept independent, for validation purposes. This is by
220 mutual agreement through GHRSSST. Near-surface (3-5 m depth) Argo measure-
221 ments are used here, which provide an estimate of foundation SST (the pre-dawn
222 temperature, i.e. without the effects of diurnal warming). This is demonstrated
223 by Figure 1, which illustrates the close match between 3-5 m depth Argo data
224 and nighttime measurements from drifting buoys at 20 cm depth. The mean
225 difference of the matchups is 0.004 K, with a standard deviation of 0.60 K. The
226 rather large standard deviation is a result of the inclusion of matchups in high
227 gradient regions such as western boundary currents, but the global distribu-
228 tion is shown in Figure 1 for completeness. Matchup criteria for Figure 1 are
229 within 3 hours and 50 km, for Argo and drifter data between 2005-2013. Ob-
230 servations were extracted from the HadIOD database v1.0.0.0 (Atkinson et al.,
231 2014), where the data undergo rigorous quality control procedures.

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

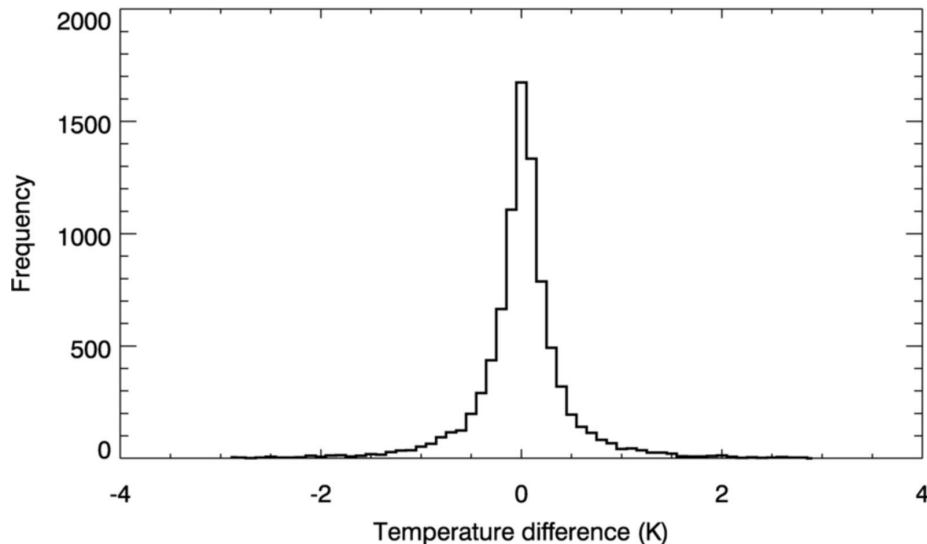


Figure 1: Distribution of nighttime Argo minus drifting buoy differences, 2005 - 2013, 0.1 K bins. Mean difference 0.004 K, standard deviation 0.60 K. Differences are taken from matchups within 3 hours and 50 km.

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

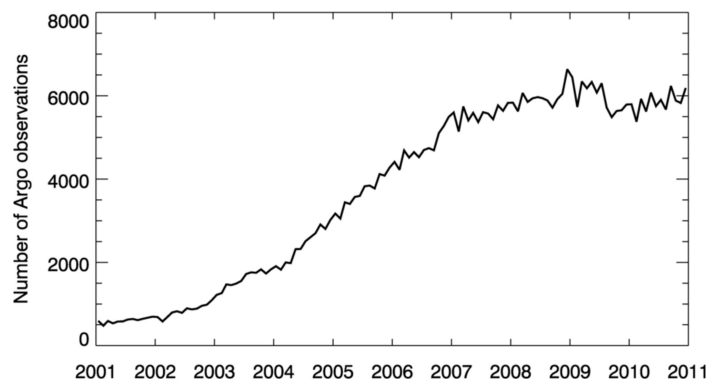
236 Daytime and nighttime Argo observations have been extracted from the EN4
 237 dataset (Good et al., 2013), where they have undergone quality control proce-
 238 dures to remove suspect observations. For each available profile, the shallowest
 239 observation between 3-5 m was obtained. A minimum depth of 3 m is used based
 240 on the assumption that this is the depth at which the effects of diurnal warming
 241 can be neglected (Zeng & Beljaars, 2005; Gentemann et al., 2009; Takaya et al.,
 242 2010). The number of Argo observations increases over time (Figure 2(a)). The
 243 dataset matures by 2007, having spread to almost cover the global ocean except
 244 for marginal seas and continental shelves (Figure 2(b)-(d)).

245 The GMPE system was used to produce matchups between the analyses and
 246 the Argo data, by interpolating the analyses from their native resolutions to the
 247 observation locations using a bilinear interpolation. The error on this interpola-
 248 tion is negligible, owing to the high resolution of the analyses. Monthly means
 249 and standard deviations of the analysis differences to Argo were calculated for
 250 2001 - 2010 (or 2001 - 2007 for HadISST2 and OSTIA v1.0) and a timeseries of
 251 the results is shown in Figure 3.

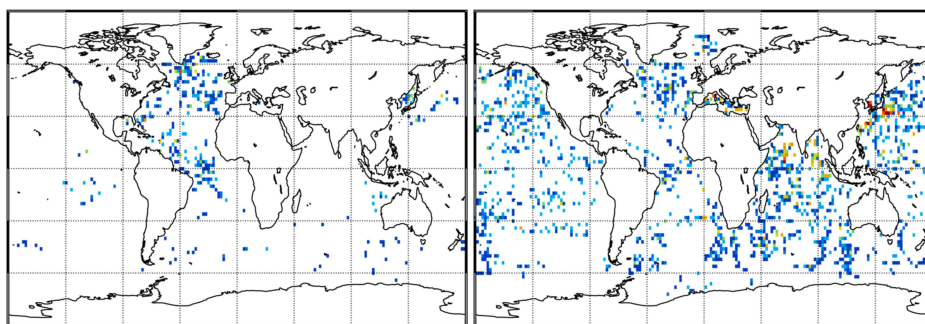
252 Figure 3 demonstrates that the CMC, SST CCI and the GMPE median
 253 datasets have the smallest monthly global standard deviation of the differences
 254 to Argo (Figure 3). MGSST and OSTIA v1.0 are in the centre of the spread,
 255 and AVHRR-OI and HadISST2 have the largest global standard deviations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

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

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

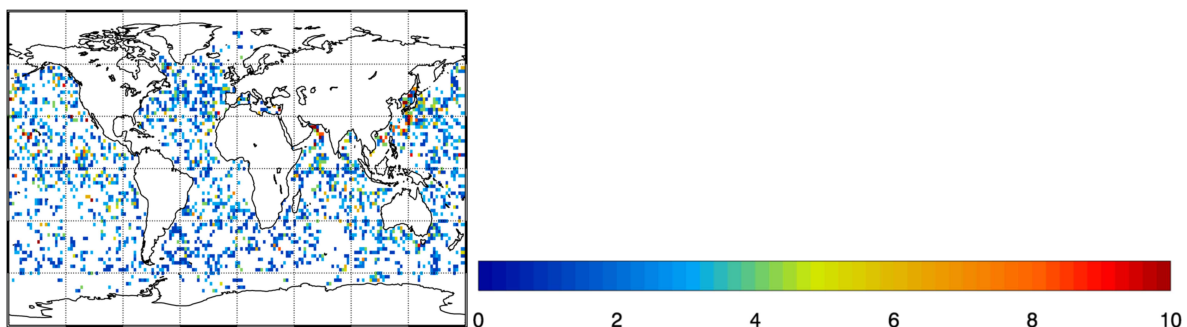


(a) Monthly total global near-surface Argo observations



(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.

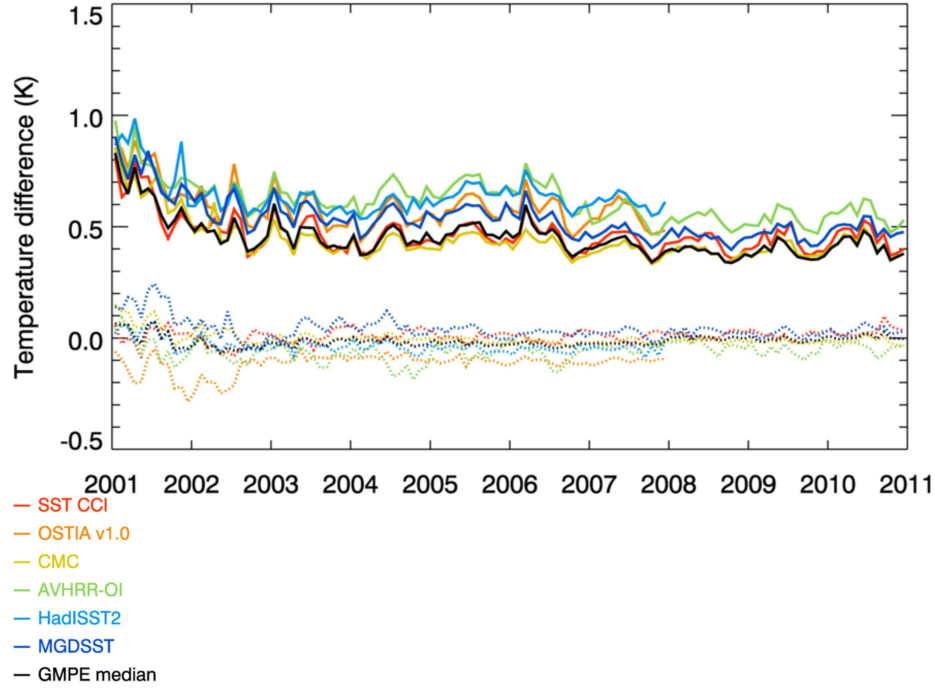


Figure 3: Timeseries of monthly analysis-minus-Argo SST differences: mean (dashed line) and standard deviation (solid line) for six analyses and their ensemble (GMPE) median, 2001-2010. All analyses independent from Argo.

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

274 SST CCI also performs well against Argo data, with a small standard de-
 275 viation of the differences (0.44 K) compared to other analyses (Table 2). This
 276 is despite the SST CCI analysis being a satellite-only product, unlike the other
 277 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). Al-

292 though the global average of the mean difference to Argo for SST CCI is small
 293 (Table 2) the SST in the tropical Pacific is around 0.1 K too warm compared
 294 to Argo (Figure 5), with some regional variation (Figure 4). This bias is re-
 295 lated to problems with the SST CCI input data in this region (Corlett et al.,
 296 2014). However, along with CMC and the GMPE median, SST CCI performs
 297 well regionally in terms of the standard deviation of the differences to Argo data
 298 (Figure 5).

299 Figure 6 shows the mean difference of each analysis to Argo, on a 5x5 de-
 300 gree grid and averaged zonally. The 5x5 degree grid was used instead of the
 301 noisier 2x2 degree grid used above, to avoid obscuring the main patterns of
 302 spatial homogeneity. Data for 2003 to 2007 was used for all analyses for a di-
 303 rect comparison of results. Figure 6 demonstrates the mean difference of the
 304 CMC analysis to Argo is small and noticeably more uniform compared to the
 305 mean differences for the other analyses, including the GMPE median. The use
 306 of observations of foundation temperature from Argo as reference data means
 307 that analyses which are intended to represent shallower depths may be warmer
 308 than Argo (e.g. SST CCI and HadISST2 at 20 cm, and CMC at 1 m (Table 1))
 309 and this difference may vary both seasonally and latitudinally. However, only
 310 MGDSSST (and CCI at low latitudes) are warm compared to Argo (Figure 6; see
 311 also Figures 4 and 5(a)). Nevertheless, this mismatch of depths may contribute
 312 to the variation in mean differences with latitude seen in Figure 6, and in Fig-
 313 ures 4 and 5. However, the difference of MGDSSST and OSTIA v1.0 foundation
 314 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.

Region	SST CCI		OSTIA v1.0		GMPE median		CMC	
	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

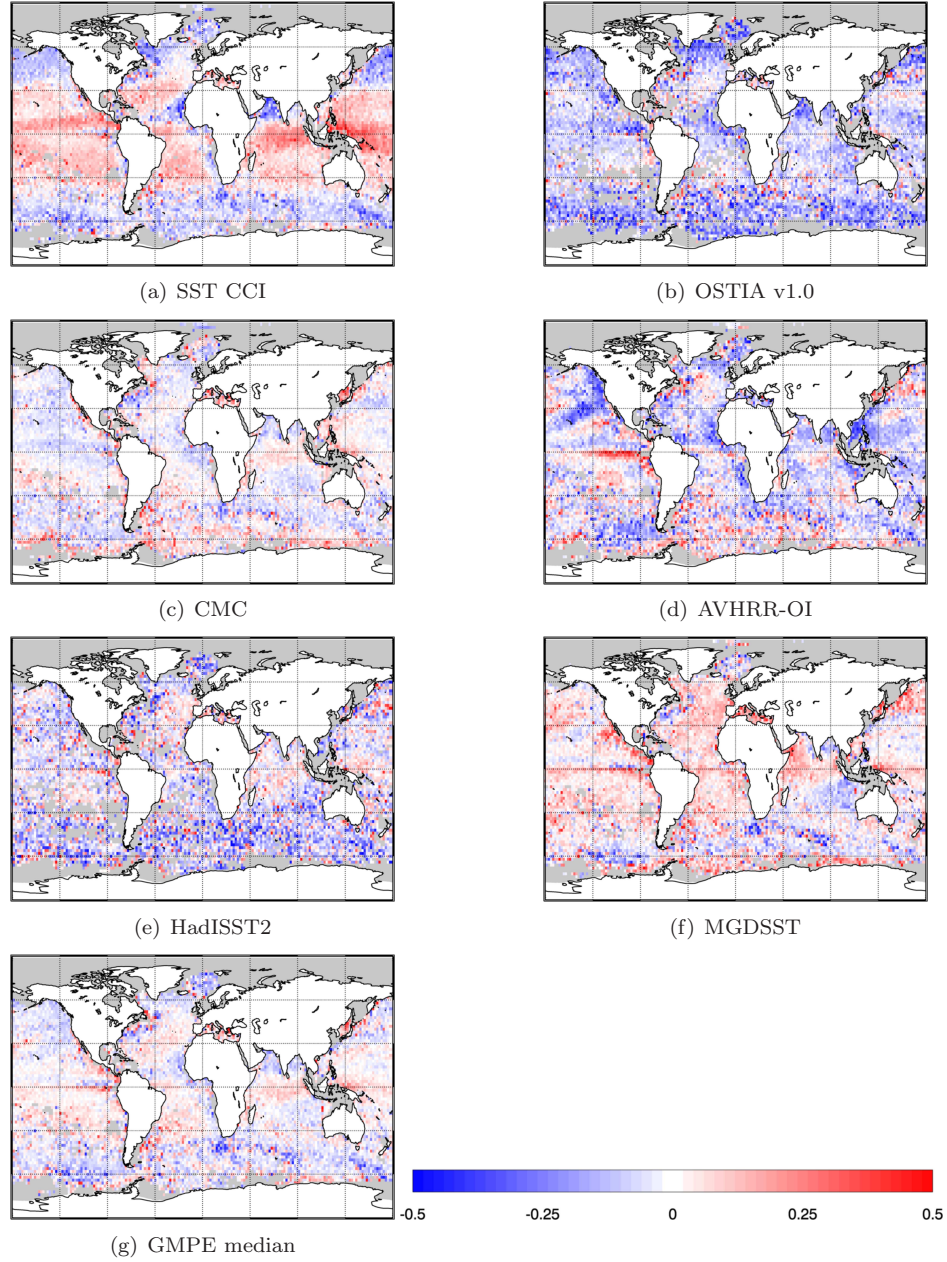


Figure 4: Spatial maps of mean global analysis-minus-Argo SST differences (K) for 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2) in 2x2 degree gridboxes, for six analyses and their ensemble (GMPE) median. Areas with no data shown in grey. All analyses independent from Argo.

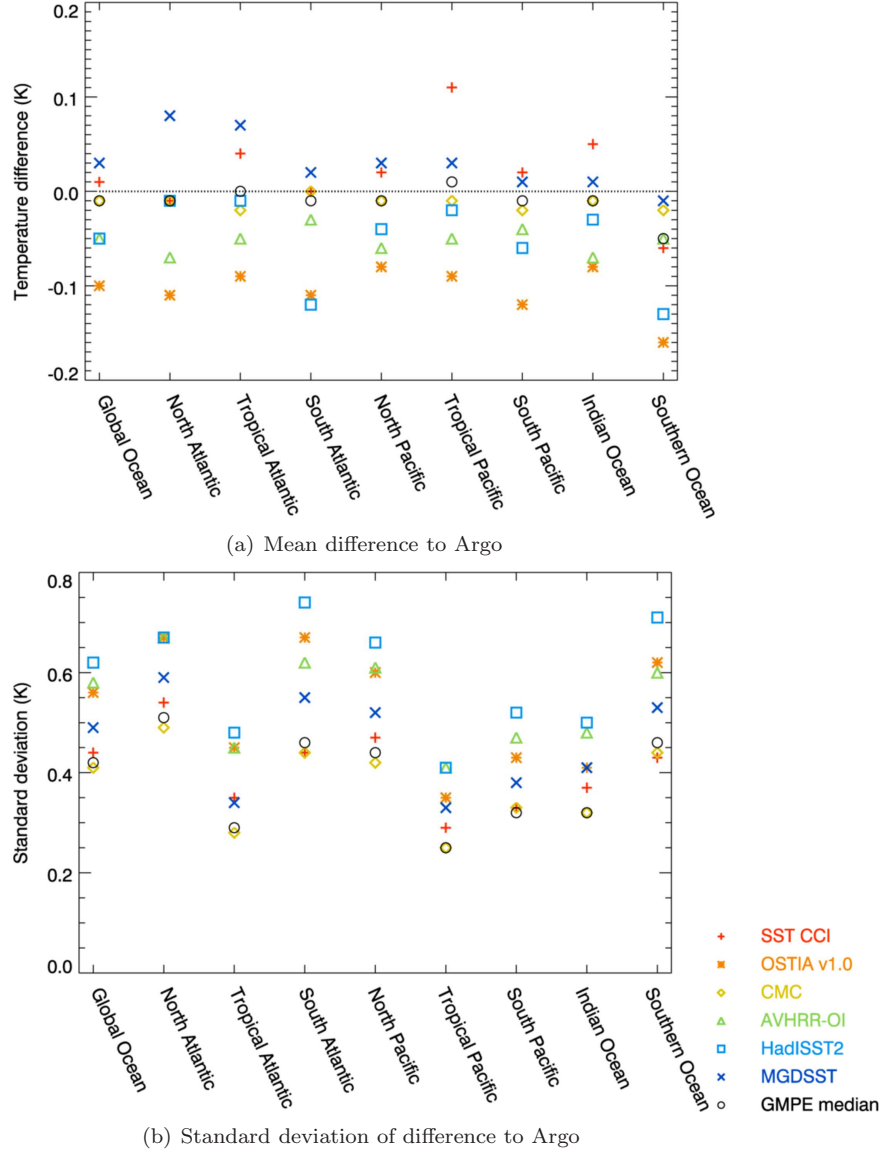


Figure 5: Regional analysis-minus-Argo SST differences: (a) mean and (b) standard deviation for six analyses and their ensemble (GMPE) median, 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2). All analyses independent from Argo.

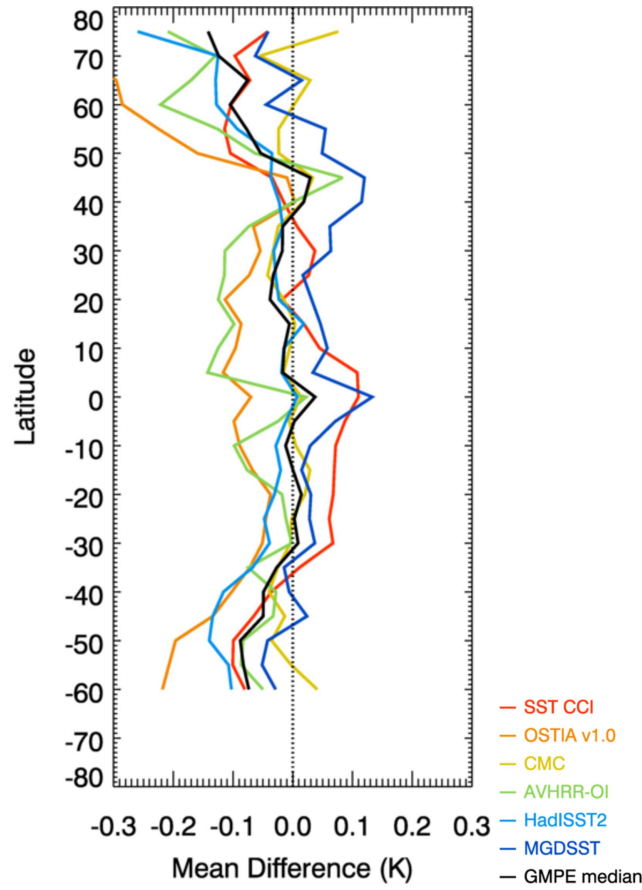


Figure 6: Zonal average of analysis-minus-Argo mean differences, 2003-2007, on a 5x5 degree grid for six analyses and their ensemble (GMPE) median. Minimum number of matchups in each grid box is 50. All analyses independent from Argo.

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 ($0-3 \text{ m s}^{-1}$) 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.

348 3.2. Assessment of temporal homogeneity of SST analyses using moored buoy

349 *data*

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

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

366 The GTMBA dataset was obtained from NOAA PMEL (Pacific Marine En-
367 vironmental Laboratory). Observations at a depth of 1 m were used. The data
368 have a sampling period of either 5, 10 or 60 minutes and the highest available
369 temporal resolution was always used if multiple sampling periods were available.
370 This means an average of daily matchups between a GTMBA buoy and an anal-

371 ysis should approximate the daily mean difference from the GTMBA buoy. All
372 available observations, both daytime and nighttime, were used in order to max-
373 imise the number of matchups. No further quality control was applied to the
374 data prior to their comparison with the SST analyses.

375 The number of observations available over the analysis period increases with
376 time (Figure 7) due to changes in reporting frequency and further deployments,
377 including the addition of the PIRATA (Prediction and Research Moored Ar-
378 ray in the Atlantic) and RAMA (Research Moored Array for African-Asian-
379 Australian Monsoon Analysis and Prediction) arrays, in 1998 and 2008 respec-
380 tively. In order to avoid aliasing effects on the stability of the GTMBA dataset,
381 only buoys which were available for more than 75% of the timeseries were in-
382 cluded in this assessment. The number of observations used is also shown in
383 Figure 7, which indicates that a large proportion of the total number of observa-
384 tions is retained despite this constraint. The locations of all GTMBA moorings
385 (109 locations; indicated by the blue and red dots) and the reduced set used
386 here (65 locations; indicated by the blue dots only) are shown in Figure 8.
387 The buoys used are primarily from the TAO/TRITON (Tropical Atmosphere
388 Ocean/Triangle Trans-Ocean Buoy Network) array in the Pacific Ocean, as these
389 provide the longest records. Therefore the locations of the GTMBA observations
390 used do not change greatly over time. However, this does mean the validation
391 reported here is only directly applicable to the tropical Pacific Ocean and thus
392 does not demonstrate any global homogeneity of these analysis datasets. How-
393 ever, alternative datasets, for example the drifting buoy network, are not known

394 to be as accurate or stable in time as GTMBA.

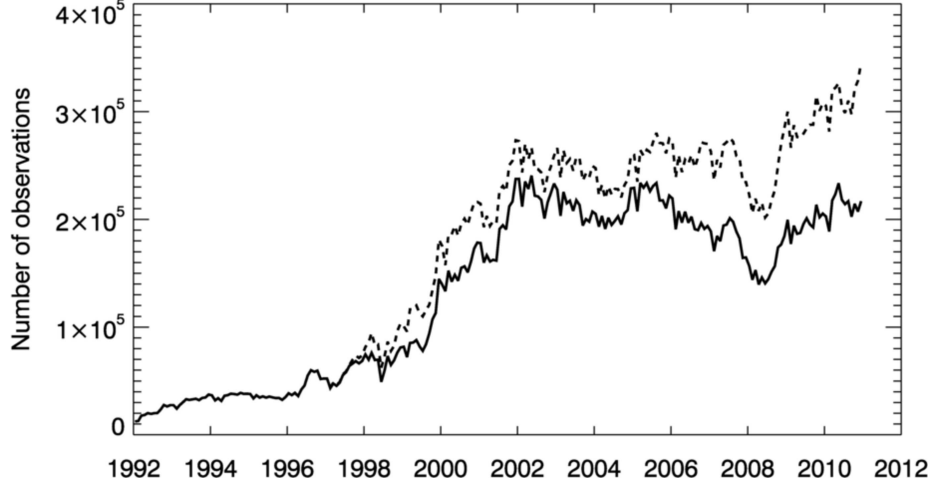


Figure 7: Monthly total number of GTMBA observations for January 1992 to December 2010. Dashed line shows all available observations, solid line those observations from buoys covering at least 75% of the timeseries (see text).

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

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

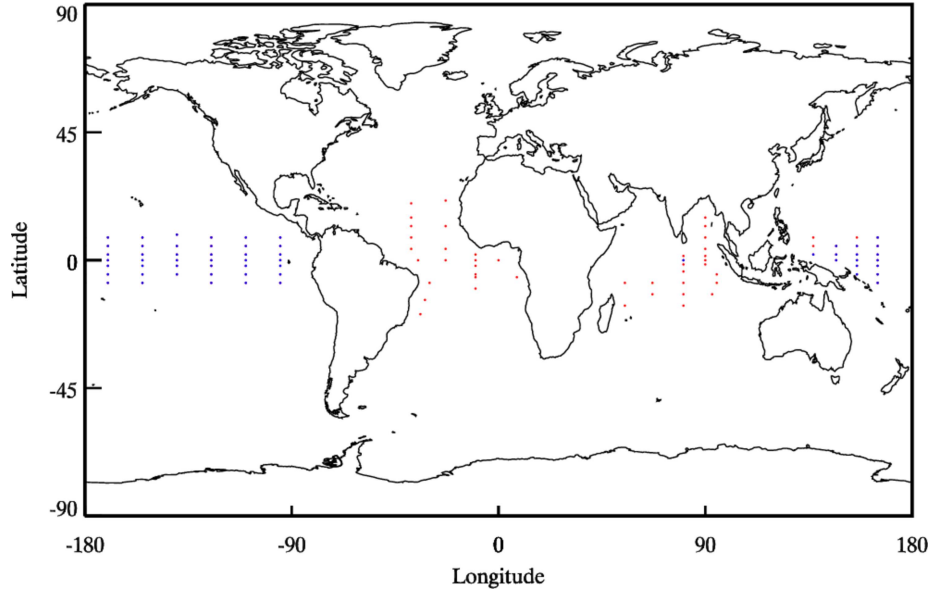


Figure 8: Nominal reference location of GTMBA buoys (red and blue dots) and the reduced set of locations (blue dots) used for validation, 1992-2010.

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

412 Although analysis data are available from September 1991, this validation
 413 begins in January 1992. This date was chosen both for computational efficiency
 414 reasons of working with full years, and to produce the multi-year monthly aver-
 415 age required for deseasonalising from the same number of datapoints per month,
 416 i.e. not including part-years. Finally, the monthly mean difference across all lo-
 417 cations was determined, producing a single analysis-minus-GTMBA timeseries
 418 for each dataset. A least squares linear fit to each timeseries of monthly mean

419 differences was calculated and 95% confidence intervals of these fits were deter-
420 mined.

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

428 ATSR-1: January 1992 - May 1995
429 ATSR-2: July 1996 - July 2002
430 AATSR : August 2002 - December 2010

431

432 Note that OSTIA v1.0 and HadISST2 finish in 2007 so the AATSR period
433 for these datasets is August 2002 - December 2007. In the gap between ATSR-
434 1 and ATSR-2 given above, the two instruments were being swapped in the
435 SST CCI analysis according to availability of data. Therefore this period is
436 not included in the short-term trend calculations for simplicity. Not all the
437 analyses use data from the ATSR series of instruments (Table 1) but the trends
438 in analysis-minus-GTMBA difference are still calculated for the same periods
439 to enable intercomparison between datasets.

440 The various SST analyses in the intercomparison are intended to be valid
441 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

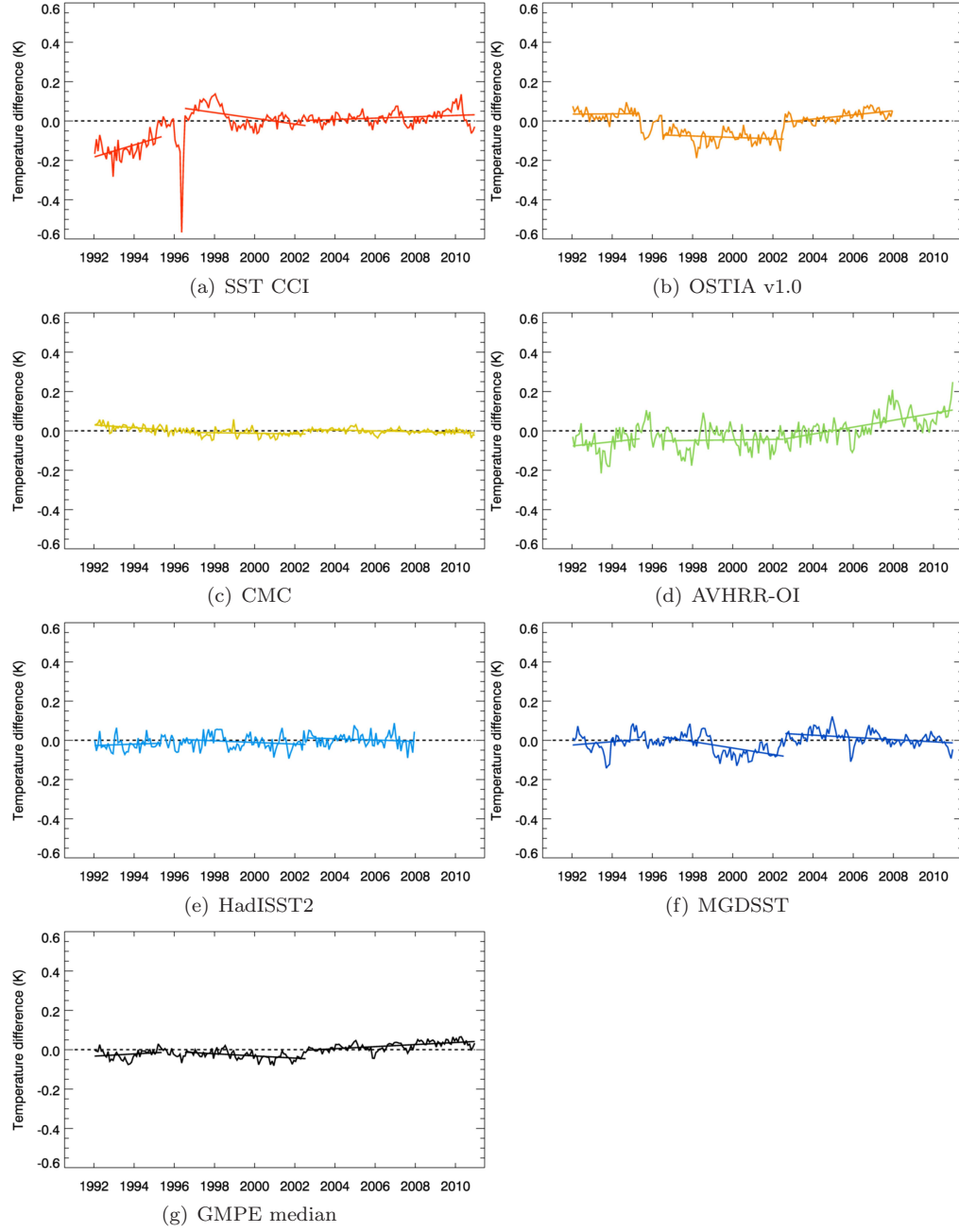


Figure 9: Monthly mean deseasonalised analysis-minus-GTMBA differences with linear fits for different ATSR periods (see text). Only SST CCI is independent from GTMBA.

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 MGD SST 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 in-

479 dicates the comparatively large trends during the earlier periods are not solely
 480 due to the independence of the data. The reduced performance of the SST
 481 CCI analysis during the lifetime of the ATSR-1 instrument is likely due to the
 482 residual effects on SST retrieval of the Pinatubo eruption (e.g. Reynolds, 1993)
 483 and the loss of the $3.7 \mu m$ channel (e.g. Murray et al., 1998). The SST CCI
 484 analysis is the only dataset included here not to perform any bias-correction of
 485 the ATSR-1 data. Although HadISST2 uses ATSR-1 as a reference (Table 1)
 486 this only applies to the period when 3-channel retrievals were available, so much
 487 of the data are not used.

488 *3.3. Assessment of SST analyses using the ensemble (GMPE) median*

489 *3.3.1. SST analysis anomaly to the GMPE median*

490 With the exception of Argo, there is no global in situ dataset independent
 491 of all the SST analyses. In order to gain some insight into the relative perfor-
 492 mance of the analyses for time periods before the Argo data became available,
 493 comparisons of the anomaly of each analysis to the ensemble median have been
 494 made. The ensemble median was produced using the GMPE (Group for High
 495 Resolution SST (GHRSSST) Multi-Product Ensemble) system, using the method
 496 described in section 2.2. The monthly mean anomaly to the GMPE median of
 497 each analysis is shown by latitude on a 2×2 degree grid for the period September
 498 1991 to December 2010 in Figure 10. This method has an advantage over using
 499 observations as a reference by allowing comparisons at all latitudes (excluding
 500 ice-covered regions) instead of solely in data-rich areas and time periods.

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

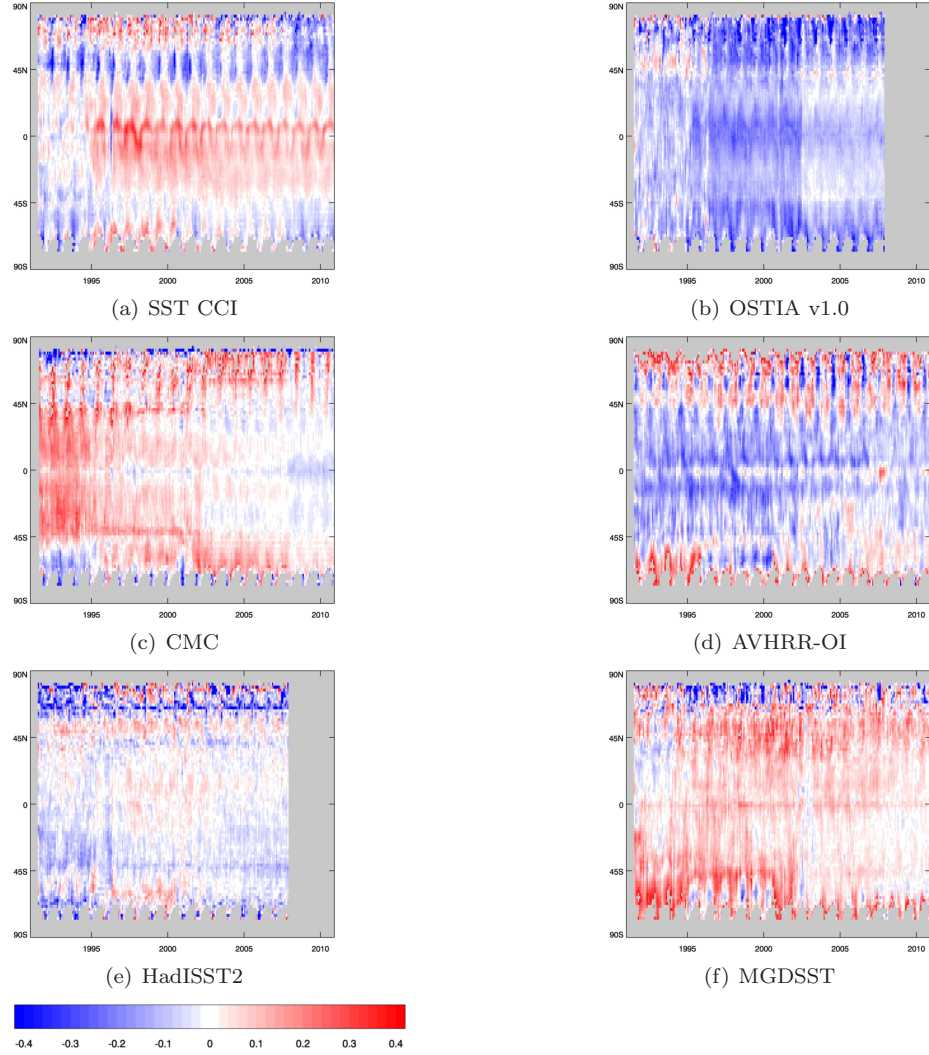


Figure 10: “Hovmöller” plots of monthly mean anomaly by latitude for six SST analyses to their ensemble (GMPE) median (analysis-minus-GMPE median) in K for 1991-2010. Areas with no data shown in grey. For reference, the ATSR-1 period is January 1992 to May 1995, ATSR-2 is July 1996 to July 2002, and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

502 (Figure 10). The SST CCI analysis has a distinct seasonal cold difference to the
 503 GMPE median at around 50N which occurs consistently throughout the whole
 504 time period. This was found to begin in the Northern Hemisphere Spring and to
 505 deepen in Summer. A similar anomaly pattern can be seen for the AVHRR-OI
 506 analysis, although of the opposite sign. This indicates the anomaly source is
 507 the AVHRR data, since this is the only common component of both analyses.
 508 As this is a seasonal feature, the cold difference for SST CCI does not appear
 509 as strong in Figure 4 on comparison to Argo as does the more persistent warm
 510 bias in the tropics. The tropical difference is seen in Figure 10(a), and also
 511 has a seasonal component, but is smaller than that seen at 50N. The anomaly
 512 of SST CCI to the GMPE median also varies in the time periods when the
 513 different instruments of the ATSR series are used (section 3.2, Figure 10(a)).
 514 The tropical warm difference is largest when the ATSR-2 data are used (June
 515 1995 - December 1995, July 1996 - July 2002), smaller in the AATSR period
 516 (July 2002 - December 2010) and does not appear when the ATSR-1 data are
 517 used (September 1991 - May 1995, January 1996 - June 1996). A distinct cold
 518 anomaly appears in the tropics from mid-May 1996 to early June 1996 and has
 519 been attributed to a decline in performance of the ATSR-1 instrument at the
 520 end of its life (Corlett et al., 2014).

521 OSTIA v1.0 shows three distinct periods of difference to the GMPE median
 522 (Figure 10(b)) seen at all latitudes and which correspond to the use of ATSR
 523 series data. This demonstrates the analysis is not homogeneous over the whole
 524 timeseries, but within these periods the difference to the GMPE median is

525 consistent (Figure 10(b), see also section 3.2 which indentified a similar result
526 compared to moored buoys for OSTIA v1.0 in the tropics).

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

538 *3.3.2. SST analysis contribution to the GMPE median*

539 The GMPE median is the ensemble median of all the contributing analyses
540 on a gridbox by gridbox basis, after they have been regridded to a $1/4^\circ$ grid.
541 If an analysis is the median its contribution in that gridbox is counted as 1.
542 If there are an even number of analyses, the mean of the two centre analyses
543 is taken and their respective contribution to the GMPE median is counted as
544 0.5. Figure 11 is a summary of the contribution of each analysis to the GMPE
545 median on a gridbox basis in various latitude bands, for the three periods of the
546 ATSR-series instruments:

547 ATSR-1: January 1992 - May 1995

548 ATSR-2: July 1996 - July 2002

549 AATSR : August 2002 - December 2007

550

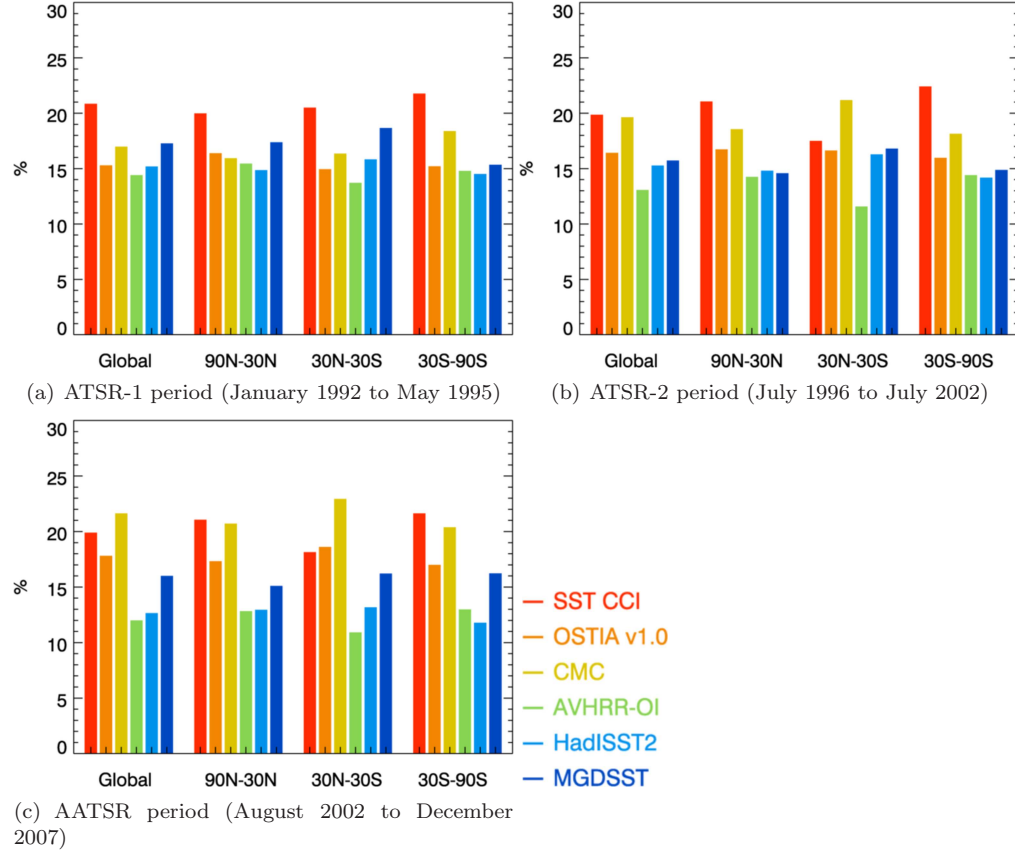


Figure 11: Percentage gridbox contribution of different SST analyses to their ensemble (GMPE) median.

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

555 Those analyses with the smallest global and regional standard deviations of

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

564 In the ATSR-1 period, for the northern and southern extratropics (90N-
 565 30N and 30S-90S) and the tropics themselves (30N-30S), the SST CCI analysis
 566 makes the largest number of contributions to the median (Figure 11(a)). These
 567 are wide latitude bands, so the seasonal temperature cycling centred on 50N in
 568 SST CCI (Figure 10(a)) does not dominate these statistics. For the ATSR-2
 569 period, SST CCI still has the largest percentage of contributions to the me-
 570 dian in the northern and southern extratropics (Figure 11(b)). However, in
 571 the tropics, where the SST CCI mean and standard deviation of differences
 572 to Argo are poorer than for other regions (e.g. Table 3) the contribution to
 573 the GMPE median is smaller, and CMC has the highest percentage of contri-
 574 butions (Figure 11(b)). For the AATSR period, SST CCI has only the third
 575 highest contribution to the median in the tropics, behind CMC and OSTIA
 576 v1.0, with MGDSST not far behind SST CCI (Figure 11(c)). In the northern
 577 and southern extratropics in the AATSR period SST CCI still has the largest
 578 number of contributions to the median, but CMC is very close.

579 Although overall the CMC and SST CCI analyses make up the largest num-
580 ber of contributions to the GMPE median, neither analysis contributes more
581 than 24% of the gridboxes in any of the three latitude bands investigated (Fig-
582 ure 11). Therefore, the GMPE median is not dominated by any one analysis,
583 but is made up of significant contributions from all the analyses.

584 3.4. Feature resolution

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

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

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

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

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

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

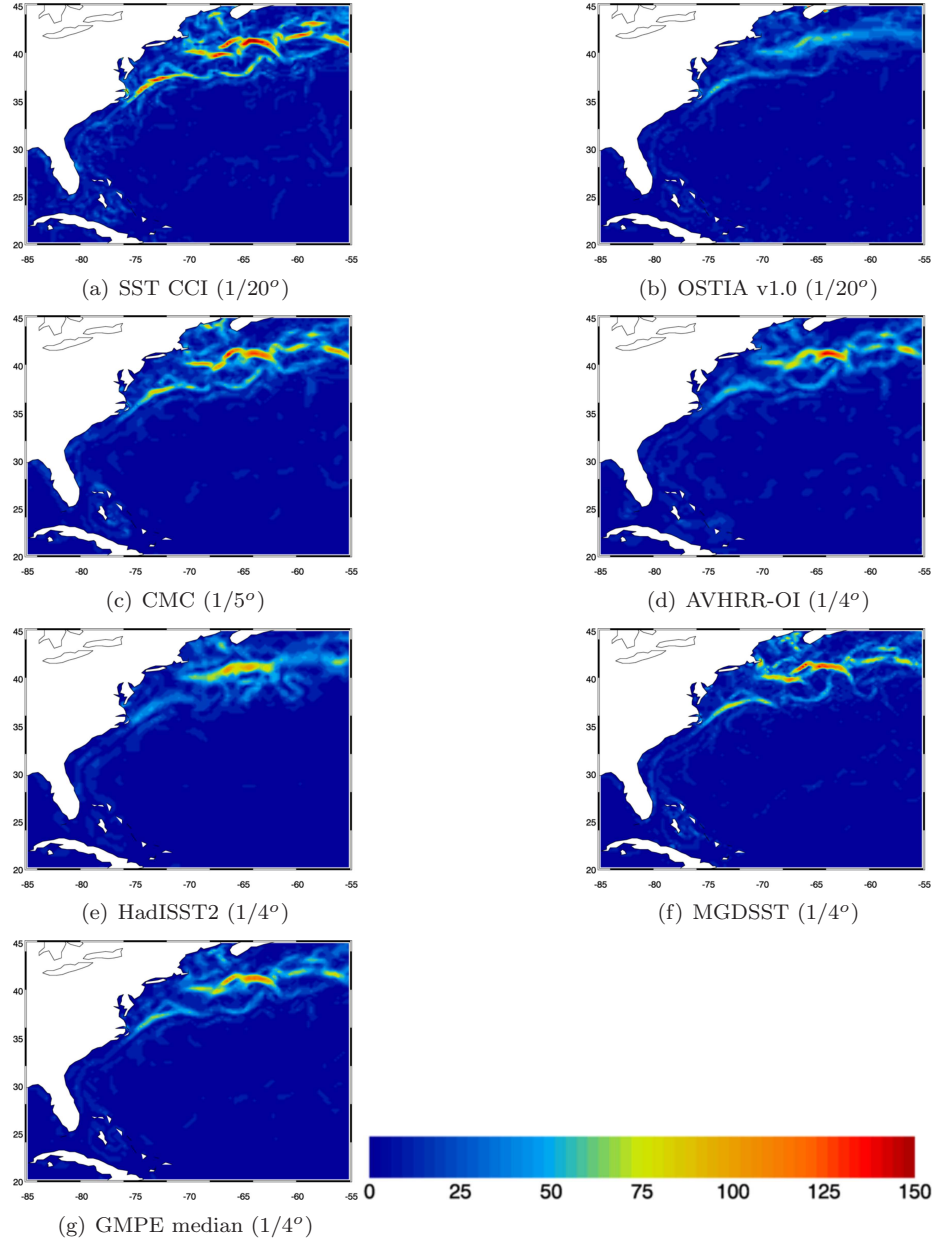


Figure 12: Horizontal SST gradients (vector sum of North-South and East-West differences) given in mK per km, on 01 July 2007 for the Gulf Stream region. Shown are six analyses and their ensemble (GMPE) median, with their grid resolutions.

625 zooming in on electronic copy).

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

634 4. Summary and Conclusions

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

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

660 None of the analyses performs badly. The rankings in Table 5 are therefore
 661 not intended to be added up and used as an overall “score” for performance as
 662 the intended use of the analysis should still inform which will be the most suit-
 663 able. For example, if a long-term, temporally homogeneous analysis is required,
 664 with reduced emphasis on feature resolution, the user might select HadISST2. If
 665 a foundation temperature is required, with good all-round performance in tem-
 666 poral and spatial homogeneity, standard deviation, bias and feature resolution
 667 criteria, MGDSSST might be selected. If a daily mean temperature at 20 cm
 668 depth, with excellent feature resolution is required, then SST CCI would be
 669 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
		<i>independent from in situ observations</i>
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
		<i>includes microwave data</i>
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
		<i>independent from ATSRs</i>
		<i>single sensor product</i>
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
		<i>uncertainty information from multiple realisations</i>
		<i>very long time period</i>
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
		<i>independent from ATSRs</i>
		<i>includes microwave data</i>
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
		<i>source potentially varies from gridbox to gridbox</i>

670 criteria are most important to the proposed application.

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

684 The analyses with the largest contributions to the GMPE median are those
685 with the smallest standard deviations of differences to Argo. This result means
686 that the relative contributions of an analysis to the ensemble median could be
687 used to provide a general idea of the accuracy of an analysis relative to others in
688 periods when no reference data are available. Seasonal anomalies to the GMPE
689 median were identified for all analyses, occurring throughout the time period
690 and demonstrating that comparison to the GMPE median also allows an in-
691 depth assessment of analysis quality for all regions and time periods. Indeed,
692 the patterns seen in the Hovmöller plot of the SST CCI analysis anomaly to

693 the GMPE median (Figure 10(a)) are qualitatively similar to those seen in
694 Hovmöller plots of the SST CCI analysis anomaly to drifter data in Corlett
695 et al. (2014).

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

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

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

715 **Acknowledgements**

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

721 **Data Access**

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

724 ESA SST CCI analysis long-term product v1.0: [http://catalogue.ceda.](http://catalogue.ceda.ac.uk/uuid/916986a220e6bad55411d9407ade347c)
725 [ac.uk/uuid/916986a220e6bad55411d9407ade347c](http://catalogue.ceda.ac.uk/uuid/916986a220e6bad55411d9407ade347c)
726 MyOcean OSTIA reanalysis v1.0: [http://marine.copernicus.eu/services-](http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=SST_GLO_SST_L4_REP_OBSERVATIONS_010_011)
727 [portfolio/access-to-products/?option=com_csw&view=details&product_](http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=SST_GLO_SST_L4_REP_OBSERVATIONS_010_011)
728 [id=SST_GLO_SST_L4_REP_OBSERVATIONS_010_011](http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=SST_GLO_SST_L4_REP_OBSERVATIONS_010_011)
729 CMC 0.2 degree: [https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-](https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-L4-GLOB-v2.0)
730 [L4-GLOB-v2.0](https://podaac.jpl.nasa.gov/dataset/CMC0.2deg-CMC-L4-GLOB-v2.0)
731 AVHRR ONLY Daily 1/4 degree OISST v2.0: [https://podaac.jpl.nasa.](https://podaac.jpl.nasa.gov/dataset/AVHRR_OI-NCEI-L4-GLOB-v2.0)
732 [gov/dataset/AVHRR_OI-NCEI-L4-GLOB-v2.0](https://podaac.jpl.nasa.gov/dataset/AVHRR_OI-NCEI-L4-GLOB-v2.0)
733 HadISST2.1.0.0: 1°, 5-day and interpolated 1/4°, daily products will be made
734 available from [https://www.metoffice.gov.uk/hadobs/hadisst2/data/download.](https://www.metoffice.gov.uk/hadobs/hadisst2/data/download.html)
735 [html](https://www.metoffice.gov.uk/hadobs/hadisst2/data/download.html)
736 MGDSST: [https://ds.data.jma.go.jp/gmd/goos/data/rrtdb/file_list.](https://ds.data.jma.go.jp/gmd/goos/data/rrtdb/file_list)

737 php#a0

738 GMPE median from long-term analysis inputs: [http://catalogue.ceda.ac.](http://catalogue.ceda.ac.uk/uuid/e0659b01259145c8bfb0de6eb12c2690)

739 [uk/uuid/e0659b01259145c8bfb0de6eb12c2690](http://catalogue.ceda.ac.uk/uuid/e0659b01259145c8bfb0de6eb12c2690)

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

741 [eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_](http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=SST_GLO_SST_L4_NRT_OBSERVATIONS_010_005)

742 [id=SST_GLO_SST_L4_NRT_OBSERVATIONS_010_005](http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=SST_GLO_SST_L4_NRT_OBSERVATIONS_010_005)

743 References

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

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

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

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

757 Dash, P., Ignatov, A., Martin, M., Donlon, C., Brasnett, B., Reynolds, R., Ban-
 758 zon, V., Beggs, H., Cayula, J.-F., Chao, Y., Grumbine, R., Maturi, E., Harris,
 759 A., Mittaz, J., Sapper, J., Chin, T. M., Vazquez-Cuervo, J., Armstrong, E. M.,
 760 Gentemann, C., Cummings, J., Piolle, J.-F., Autret, E., Roberts-Jones, J.,
 761 Ishizaki, S., Hoyer, J. L., & Poulter, D. (2012). Group for High Resolu-
 762 tion Sea Surface Temperature (GHR SST) analysis fields inter-comparisons -
 763 Part 2: Near real time web-based level 4 SST Quality Monitor (L4-SQUAM).
 764 *Deep-Sea Research II*, 77-80, 31–43.

765 Donlon, C. J., Gentemann, C., & Nykjaer, L. (2004). Using SST measurements
 766 from microwave and infrared satellite measurements. *Intl. J. Rem. Sens.*, 25,
 767 1331–1336.

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

771 Donlon, C. J., Minnett, P. J., Gentemann, C., Nightingale, T. J., Barton, I. J.,
 772 Ward, B., & Murray, M. J. (2002). Toward improved validation of satellite
 773 sea surface skin temperature measurements for climate research. *J. Climate*,
 774 15, 353–369.

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

- 779 Gentemann, C., Minnett, P., & Ward, B. (2009). Profiles of ocean surface
780 heating (POSH): A new model of upper ocean diurnal warming. *J. Geophys.*
781 *Res.*, *114*. C07017, DOI: 10.1029/2008JC004825.
- 782 Good, S. A., Martin, M. J., & Rayner, N. A. (2013). EN4: Quality con-
783 trolled ocean temperature and salinity profiles and monthly objective anal-
784 yses with uncertainty estimates. *J. Geophys. Res.*, *118*, 6704–6716. DOI:
785 10.1002/2013JC009067.
- 786 Kennedy, J. J., Rayner, N. A., Millington, S. C., & Saunby, M. (2018). The Met
787 Office Hadley Centre Sea Ice and Sea-Surface Temperature data set, version
788 2, Part 2: Sea Surface Temperature Analysis. *J. Geophys. Res. Atmos.*, . (in
789 prep.).
- 790 Kilpatrick, K. A., Podesta, G. P., & Evans, R. (2001). Overview of the
791 NOAA/NASA Advanced Very High Resolution Radiometer Pathfinder al-
792 gorithm for sea surface temperature and associated matchup database. *J.*
793 *Geophys. Res.*, *106*, 9179–9197.
- 794 Kurihara, Y., Sakurai, T., & Kuragano, T. (2006). Global daily sea surface
795 temperature analysis using data from satellite microwave radiometer, satellite
796 infrared radiometer and in situ observations. *Weather Bull.*, *73*, 1–18. (in
797 Japanese).
- 798 Maloney, E. D., & Chelton, D. B. (2006). An assessment of the sea surface
799 temperature influence on surface wind stress in numerical weather prediction
800 and climate models. *J. Climate*, *19*, 2743–2762.

801 Martin, M., Dash, P., Ignatov, A., Banzon, V., Beggs, H., Brasnett, B., Cayula,
 802 J.-F., Cummings, J., Donlon, C., Gentemann, C., Grumbine, R., Ishizaki,
 803 S., Maturi, E., Reynolds, R. W., & Roberts-Jones, J. (2012). Group for
 804 High Resolution Sea Surface Temperature (GHRSSST) analysis fields inter-
 805 comparisons. Part 1: A GHRSSST Multi-Product Ensemble (GMPE). *Deep-*
 806 *Sea Research II*, 77-80, 21–30.

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

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

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

822 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 GHRSSST. URL: [https://www.
ghrsst.org/wp-content/uploads/2018/01/CDR-TAG_CDAF-v1.0.5.pdf](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 anal-
ysis 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

880 List of Figures

881	1	Distribution of nighttime Argo minus drifting buoy differences,	
882		2005 - 2013, 0.1 K bins. Mean difference 0.004 K, standard devi-	
883		ation 0.60 K. Differences are taken from matchups within 3 hours	
884		and 50 km.	12

885	2	(a) Timeseries of monthly total number of global near-surface	
886		Argo observations, using shallowest observation between 3 m and	
887		5 m depth, and (b-d) spatial map of same for given month binned	
888		in 2x2 degree grid boxes.	14
889	3	Timeseries of monthly analysis-minus-Argo SST differences: mean	
890		(dashed line) and standard deviation (solid line) for six analyses	
891		and their ensemble (GMPE) median, 2001-2010. All analyses	
892		independent from Argo.	15
893	4	Spatial maps of mean global analysis-minus-Argo SST differences	
894		(K) for 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2)	
895		in 2x2 degree gridboxes, for six analyses and their ensemble (GMPE)	
896		median. Areas with no data shown in grey. All analyses inde-	
897		pendent from Argo.	19
898	5	Regional analysis-minus-Argo SST differences: (a) mean and (b)	
899		standard deviation for six analyses and their ensemble (GMPE)	
900		median, 2003-2010 (or 2003-2007 for OSTIA v1.0 and HadISST2).	
901		All analyses independent from Argo.	20
902	6	Zonal average of analysis-minus-Argo mean differences, 2003-2007,	
903		on a 5x5 degree grid for six analyses and their ensemble (GMPE)	
904		median. Minimum number of matchups in each grid box is 50.	
905		All analyses independent from Argo.	21

906	7	Monthly total number of GTMBA observations for January 1992	
907		to December 2010. Dashed line shows all available observations,	
908		solid line those observations from buoys covering at least 75% of	
909		the timeseries (see text).	25
910	8	Nominal reference location of GTMBA buoys (red and blue dots)	
911		and the reduced set of locations (blue dots) used for validation,	
912		1992-2010.	26
913	9	Monthly mean deseasonalised analysis-minus-GTMBA differences	
914		with linear fits for different ATSR periods (see text). Only SST	
915		CCI is independent from GTMBA.	29
916	10	“Hovmöller” plots of monthly mean anomaly by latitude for six	
917		SST analyses to their ensemble (GMPE) median (analysis-minus-	
918		GMPE median) in K for 1991-2010. Areas with no data shown in	
919		grey. For reference, the ATSR-1 period is January 1992 to May	
920		1995, ATSR-2 is July 1996 to July 2002, and AATSR is August	
921		2002 to December 2010.	32
922	11	Percentage gridbox contribution of different SST analyses to their	
923		ensemble (GMPE) median.	35
924	12	Horizontal SST gradients (vector sum of North-South and East-	
925		West differences) given in mK per km, on 01 July 2007 for the	
926		Gulf Stream region. Shown are six analyses and their ensemble	
927		(GMPE) median, with their grid resolutions.	39