



<b>Customer</b>	: ESRIN	<b>Document Ref</b>	: SST_CCI-PVIR-UoL-001
<b>WP No</b>	: 40320	<b>Issue Date</b>	: 22 January 2014
		<b>Issue</b>	: 1

**Project** : CCI Phase 1 (SST)

**Title** : Product Validation and Intercomparison Report (PVIR)

**Abstract** : This document contains the Product Validation and Intercomparison Report (PVIR) for the ESA SST\_CCI project

  
\_\_\_\_\_  
**Authors** : Gary Corlett, Chris Atkinson,  
Nick Rayner, Simon Good,  
Emma Fiedler, Alison McLaren,  
Jacob Hoeyer, Claire Bulgin.

  
\_\_\_\_\_  
**Approved** : Chris Merchant  
Science Leader

\_\_\_\_\_  
**Accepted** : Craig Donlon  
ESA

**Distribution** : SST\_CCI team members  
Craig Donlon

**EUROPEAN SPACE AGENCY  
CONTRACT REPORT**

The work described in this report was done under ESA contract.  
Responsibility for the contents resides in the author or organisation  
that prepared it.

## AMENDMENT RECORD

This document shall be amended by releasing a new edition of the document in its entirety. The Amendment Record Sheet below records the history and issue status of this document.

### AMENDMENT RECORD SHEET

ISSUE	DATE	REASON FOR CHANGE
A	31 Oct 2013	Initial Issue for internal review
B	02 Dec 2013	Issued to ESA following internal review
C	13 Jan 2014	Updated following ESA RIDS
D	22 Jan 2014	Updated ATSR-1 plots and statistics in Appendix B3
1	22 Jan 2014	Issued

## TABLE OF CONTENTS

<b>1. INTRODUCTION</b> .....	<b>6</b>
1.1 Purpose and Scope.....	6
1.2 Structure of the Document.....	6
1.3 Referenced Documents.....	7
1.4 Acronyms and abbreviations.....	10
<b>2. DEFINITIONS</b> .....	<b>12</b>
<b>3. SUMMARY OF ACTIVITIES</b> .....	<b>13</b>
3.1 Multi-sensor match-up database.....	14
3.2 SST_CCI Products.....	14
3.3 Uncertainties.....	16
3.4 Independence of validation activities.....	16
3.5 Getting Endorsements.....	17
3.6 Release of Products.....	17
<b>4. PRODUCT VALIDATION</b> .....	<b>18</b>
4.1 Introduction.....	18
4.1.1 Definitions.....	18
4.1.2 Reference data.....	18
4.1.3 Rules and responsibilities for objective independent product validation.....	18
4.1.4 Validation criteria.....	19
4.1.5 Depth/time adjustments.....	19
4.1.6 Uncertainty verification.....	19
4.1.7 Classes of validation.....	20
4.1.8 Types of validation.....	20
4.1.9 Analysis procedures.....	21
4.2 Reference dataset.....	21
4.2.1 Introduction.....	21
4.2.2 Overview of data sources.....	22
4.2.2.1 SST at approximately 0.2m depth from drifting buoys.....	23
4.2.2.1.1 Background.....	23
4.2.2.1.2 Accuracy.....	23
4.2.2.1.3 Stability.....	24
4.2.2.2 SST at approximately 1 m depth from moored buoys.....	26
4.2.2.2.1 Background.....	26
4.2.2.2.2 Accuracy.....	27
4.2.2.3 SST <sub>skin</sub> from shipborne radiometers.....	28
4.2.2.3.1 Background.....	28
4.2.2.3.2 Accuracy.....	29
4.2.2.4 Near-surface temperature measurements from Argo.....	30
4.2.2.4.1 Background.....	30
4.2.2.4.2 Accuracy.....	31
4.2.2.4.3 Stability.....	32
4.2.3 Criteria for selection.....	32
4.2.4 Additional quality control.....	33
4.2.5 Content of Reference Dataset for Product Validation.....	34
4.3 Validation of AVHRR products using the MMS.....	35
4.4 Validation of ATSR Products using the MMS.....	37
4.5 Validation of analysis products using the MMS.....	40
4.6 Validation of SST_CCI analysis using re-matching.....	42
4.6.1 The long term product.....	42
4.6.2 The demonstration product.....	44
4.7 High-latitude validation using the MMS.....	46
4.7.1 Spatial and temporal coverage of match-ups.....	46

4.7.2	General error numbers.....	52
4.7.3	Detailed validation.....	53
4.7.3.1	Solar zenith angle.....	54
4.7.3.2	Total column water vapour.....	55
4.7.3.3	Day in year.....	56
4.7.3.4	Distance to Ice.....	56
4.7.4	Summary.....	58
4.8	Summary of validation results.....	58
<b>5.</b>	<b>VALIDATION AND VERIFICATION OF SST_CCI UNCERTAINTY ESTIMATES .....</b>	<b>60</b>
5.1	Introduction.....	60
5.1.1	Uncertainty validation for AVHRR and ATSR data using the MMS.....	60
5.1.2	Uncertainty validation using re-matching.....	61
5.2	Results for SST_CCI AVHRR products using the MMS.....	62
5.3	Results for SST_CCI ATSR products using the MMS.....	62
5.4	Results for SST_CCI analysis products using the MMS.....	63
5.5	Results for SST_CCI analysis products using re-matching.....	64
5.5.1	The long term product.....	64
5.5.2	The demonstration product.....	65
5.6	Uncertainty verification using the MMS.....	67
5.6.1	SST_CCI AVHRR example verification map.....	68
5.6.2	SST_CCI ATSR example verification map.....	68
5.6.3	SST_CCI analysis example verification map.....	69
<b>6.</b>	<b>PRODUCT INTERCOMPARISON .....</b>	<b>70</b>
6.1	Introduction.....	70
6.2	Intercomparison datasets.....	70
6.3	Inter-comparison of reanalyses.....	71
6.3.1	Validation of 10-year time series against independent data from Argo.....	71
6.3.2	Effect of including microwave data in OSTIA CCI reanalysis.....	77
6.3.3	Comparison of OSTIA CCI to OSTIA v1.0.....	79
6.4	Intercomparison using GMPE data.....	80
6.4.1	Reanalysis anomaly to GMPE median.....	80
6.4.2	Analysis contribution to the median.....	83
6.5	Feature resolution.....	84
6.6	Summary of product intercomparison results.....	84
<b>7.</b>	<b>SUMMARY AND CONCLUSIONS.....</b>	<b>86</b>
<b>APPENDIX A</b>	<b>DETAILED AVHRR PRODUCT VALIDATION RESULTS .....</b>	<b>88</b>
A.1	AVHRR MTA.....	89
A.2	AVHRR 18.....	94
A.3	AVHRR 17.....	99
A.4	AVHRR 16.....	104
A.5	AVHRR 15.....	109
A.6	AVHRR 14.....	114
A.7	AVHRR 12.....	119
<b>APPENDIX B</b>	<b>DETAILED ATSR PRODUCT VALIDATION RESULTS .....</b>	<b>124</b>
B.1	AATSR.....	125
B.2	ATSR-2.....	130
B.3	ATSR-1.....	135
<b>APPENDIX C</b>	<b>DETAILED ANALYSIS PRODUCT VALIDATION RESULTS .....</b>	<b>140</b>
C.1	SST_CCI analysis long-term product.....	141
<b>APPENDIX D</b>	<b>ASSESSMENT OF USER REQUIREMENTS .....</b>	<b>144</b>



**APPENDIX E    ADHERENCE TO CCI PROJECT GUIDELINES .....146**

## 1. INTRODUCTION

The SST\_CCI project is part of the ESA Climate Change Initiative (CCI), which aims to produce and validate improved sea surface temperature (SST) products, produced by combining retrievals of SST from different satellite sensors, which will contribute to the SST essential climate variable (ECV).

In order to identify the best performing retrieval algorithm or combination of algorithms, the SST\_CCI project held an open algorithm selection exercise. This consisted of an algorithm intercomparison exercise (described in ESA documents as the “Round Robin”, (RR)) followed by selection of algorithms following criteria defined in the SST\_CCI Product Validation Plan (PVP; RD.272). Results from the SST\_CCI Round Robin exercise are given in the SST\_CCI Algorithm Selection Report (ASR; RD.226) and the final algorithms are described in detail in the SST\_CCI Algorithm Theoretical Basic Document (ATBD; RD.305). The chosen algorithm(s) were then implemented in an end-to-end system to generate the first SST\_CCI data records. Further details of the processing system can be found in the SST\_CCI System Specification Document (SSD; RD.259) and results from system verification activities are given in the SST\_CCI System Verification Report (SVR; RD.329).

Following selection and implementation the SST\_CCI L2, L3 and L4 products have been validated using high quality SST measurements made in situ from a number of sources. This validation is independent in that (1) the validation is undertaken by a team that is independent of the algorithm development team, and (2) fully independent in situ validation data have been used as much as possible (PVP; RD.272). In addition the SST\_CCI L4 products have been compared to other L4 products using an implementation of the Group for High Resolution SST (GHRSSST) Multi Product Ensemble (GMPE).

### 1.1 Purpose and Scope

This document is the SST\_CCI Product Validation and Intercomparison Report (PVIR). It describes the approach to product validation and intercomparison for the SST\_CCI products as described in the PVP (RD.272).

### 1.2 Structure of the Document

After this introduction, the document is divided into a number of major sections that are briefly described below:

Section	Contains
Section 2	Important definitions
Section 3	Overview of PVIR activities
Section 4	SST CCI product validation results
Section 5	SST CCI uncertainty validation and verification results
Section 6	SST CCI L4 analysis product intercomparison results
Section 7	Summary and conclusions
Appendix A	Detailed SST CCI L2P validation results per sensor

Appendix B	Detailed SST CCI L3U validation results per sensor
Appendix C	Detailed SST CCI L4 analysis validation results
Appendix D	A summary of how user requirements are addressed in the PVIR
Appendix E	A summary of how the PVIR adheres to CCI project guidelines

### 1.3 Referenced Documents

The following is a list of documents with a direct bearing on the content of this report. Where referenced in the text, these are identified as RD.n, where 'n' is the number in the list below:

ID	Title
RD.047	Donlon, C., Robinson, I.S., Reynolds, M., Wimmer, W., Fisher, G., Edwards, R., Nightingale, T.J., (2008). An infrared sea surface temperature autonomous radiometer (ISAR) for deployment aboard volunteer observing ships (VOS). <i>Journal of Atmospheric and Oceanic Technology</i> , 25, 93-113.
RD.050	Barton, I.J., Minnett, P.J., Maillet, K.A., Donlon, C.J., Hook, S.J., Jessup, A.T., Nightingale, T.J., (2004). The Miami2001 Infrared Radiometer Calibration and Intercomparison. Part II: Shipboard Results, <i>Journal of Atmospheric and Oceanic Technology</i> , 21, 268-283.
RD.058	Lumpkin, R., and Pazos, M.: Measuring surface currents with Surface Velocity Program drifters: the instrument, its data, and some recent results. In: <i>Lagrangian Analysis and Prediction of Coastal and Ocean Dynamics (LAPCOS)</i> , ed. A. Griffa, A. D. Kirwan, A. J. Mariano, T. Ozgokmen, and T. Rossby, 500pp
RD.072	Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. J. Ansell, and S. F. B. Tett (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset, <i>J. Climate</i> , 19, 446-469, doi:10.1175/JCLI3637.1.
RD.076	Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S. and M.G. Schlax, 2007. Daily high resolution blended analyses for sea surface temperature. <i>J. Climate</i> 20, 5473-5496.
RD.150	Systematic Observation Requirements for Satellite-based Products for Climate: Supplemental Details to the satellite-based component of the "Implementation Plan for the Global Observing System for Climate in support of the UNFCCC (GCOS-92)", GCOS-107, September 2006 (WMO/TD No.1338)
RD.164	SST_CCI Phase I Statement of Work
RD.169	ESA CCI Project Guidelines V1, EOP-DTEX-EOPS-SW-10-0002, Issue 1, Revision 0
RD.171	SST_CCI User Requirements Document
RD.172	SST_CCI Data Access Requirements Document

ID	Title
RD.175	SST_CCI Product Specification Document
RD.191	Bureau International des Poids et Mesures, Guide to the Expression of Uncertainty in Measurement (GUM), JCGM 100:2008, 2008. Available online at <a href="http://www.bipm.org/en/publications/guides/gum.html">http://www.bipm.org/en/publications/guides/gum.html</a>
RD.210	Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011b). Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 1: measurement and sampling errors. <i>J. Geophys. Res.</i> , 116, D14103, doi:10.1029/2010JD015218
RD.211	Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011c). Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 2: biases and homogenisation. <i>J. Geophys. Res.</i> , 116, D14104, doi:10.1029/2010JD015220
RD.226	SST_CCI Algorithm Selection Report
RD.227	Fairall, C., E. Bradley, J. Godfrey, G. Wick, J. Edson, and G. Young (1996), Cool-skin and warm-layer effects on sea surface temperature, <i>J. Geophys. Res.</i> , 101(C1), 1295-1308.
RD.229	SST_CCI Uncertainty Characterisation Report
RD.234	Minnett, P. J., 1991: Consequences of sea surface temperature variability on the validation and applications of satellite measurements. <i>J. Geophys. Res.</i> , 96, 18,475-18,489.
RD.237	Kurihara, Y., Sakurai, T. and T. Kuragano, 2006. Global daily sea surface temperature analysis using data from satellite microwave radiometer, satellite infrared radiometer and in situ observations. <i>Weather Bull.</i> 73, 1-18.
RD.239	Roberts-Jones, J., Fiedler, E. K. and M. Martin, 2012. Daily, global, high-resolution SST and sea ice reanalysis for 1985-2007 using the OSTIA system. <i>J. Climate</i> 25, 6215-6232.
RD.242	Theocharus, E., E. Usadi and N.P. Fox, 2010: CEOS comparison of IR brightness temperature measurements in support of satellite validation. Part I: Laboratory and ocean surface temperature comparison of radiation thermometers, NPL report OP3, 136pp.
RD.243	Kennedy, J.J., R.O. Smith and N.A. Rayner, 2012: Using AATSR data to assess the quality of in situ sea-surface temperature observations for climate studies, <i>Remote Sensing of the Environment</i> , 116, 79-92.
RD.244	Reverdin G., Boutin J., Martin N., et al., 2010: Temperature Measurements from Surface Drifters, <i>J. Atmos. Ocean. Tech.</i> , 27, 1403-1409
RD.245	Emery, W., D. Baldwin, P. Schlüssel, and R. Reynolds, 2001: Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements, <i>Journal of Geophysical Research</i> , 106 (C2), 2387–2405, doi:10.1029/2000JC000246.
RD.246	O'Carroll, A.G., J.R. Eyre and R.W. Saunders, 2008: Three-way error analysis between AATSR, AMSR-E, and in situ sea surface temperature observations, <i>J. Atmos. Ocean. Tech.</i> , 25, 1197-1207, doi: 10.1175/2007JTECHO542.1
RD.247	Ullman D.S., Cornillon P.C. 2000. Evaluation of front detection methods for satellite-derived SST data using in situ observations. <i>J. Atmos. Oceanic Tech.</i> , 17(12), pp. 1667–1675
RD.258	Brasnett, B. 2012. A 20-year reanalysis of sea surface temperature, Report, CMC.

ID	Title
RD.259	SST_CCI System Specification Document
RD.263	Kantha L.H., and Clayson C.A., An improved mixed layer model for geophysical applications. J. Geophys. Res. Vol. 99 (C12), 25235–25266, 1994.
RD.264	Good, S.A., Martin, M.J., and N.A. Rayner, 2013. EN4: quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. Submitted to J. Geophys. Res.
RD.272	SST_CCI Product Validation Plan
RD.294	Roberts-Jones, J., Fiedler, E. K., M. Martin and A. McLaren, 2013. Improvements to the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. CCI Phase 1 (SST) Technical Report, SST_CCI-REP-UKMO-001, Issue C.
RD.305	SST_CCI Algorithm Theoretical Basis Document
RD.319	Kennedy, J. J., Rayner N.A., Millington, S.C. and M. Saunby, 2013. The Met Office Hadley Centre Sea Ice and Sea-Surface Temperature data set, version 2, part 2: Sea Surface Temperature analysis. In prep.
RD.326	Atkinson, C. P., N. A. Rayner, J. Roberts-Jones, and R. O. Smith (2013), Assessing the quality of sea surface temperature observations from drifting buoys and ships on a platform-by-platform basis, J. Geophys. Res. Oceans, 118, doi:10.1002/jgrc.20257
RD.329	SST_CCI System Verification Report
RD.330	SST_CCI Product Validation and Intercomparison Report
RD.331	SST_CCI Climate Assessment Report
RD.332	Woodruff, S. D., S. J. Worley, S. J. Lubker, Z. Ji, E. Freeman, D. I. Berry, P. Brohan, E. C. Kent, R. W. Reynolds, S. R. Smith and C. Wilkinson (2011), ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive, Int. J. Climatol. 31: 951-967, DOI:10.1002/joc.2103.
RD.333	Roberts-Jones, J., McLaren, A., and M. Martin, 2013. Estimating and assessing the impact of background error covariance parameters in the OSTIA system. In prep.
RD.334	Rayner, N. A., J. J. Kennedy, R. O. Smith and H. A. Titchner (2013) The Met Office Hadley Centre Sea Ice and Sea Surface Temperature data set, version 2, part 3: the combined analysis. In prep. for JGR Atmospheres
RD.335	Poulter, D. J. S., & Eastwood, S. (2008). Validation of the OSI SAF Metop SST product in polar regions. OSI-SAF report.
RD.336	Martin, M., Dash, P., Ignatov, A. et al., 2012. Group for high resolution sea surface temperature (GHRSSST) analysis fields inter-comparisons. Part 1: A GHRSSST multi-product ensemble (GMPE). Deep-Sea Research II 77-80, 21-30.
RD.340	Hoeyer, J.L., I. Karagali, G. Dybkjær, and R. Tonboe, 2012. Multi sensor validation and error characteristics of Arctic satellite sea surface temperature observations. Remote Sensing of Environment, 121, pp 335-346, doi:10.1016/j.rse.2012.01.013
RD.355	Oka, E. and Ando, K. (2004). Stability of Temperature and Conductivity Sensors of Argo Profiling Floats. Journal of Oceanography, 60, 2, 253-258.

The current version of each SST\_CCI project document is available via the SST CCI web pages at <http://www.esa-sst-cci.org/?q=documents#>.

## 1.4 Acronyms and abbreviations

The following acronyms and abbreviations have been used in this report with the meanings shown:

Acronym	Definition
ASR	Algorithm Selection Report
ATBD	Algorithm Theoretical Basis Document
ATLAS	Autonomous Temperature Line Acquisition System
ATSR	Along-Track Scanning Radiometer
AVHRR	Advanced Very High Resolution Radiometer
CAR	Climate Assessment Report
CCI	Climate Change Initiative
CEOS	Committee on Earth Observing Satellites
cf.	Compared With
CF	Climate Forecast
CIRIMS	Calibrated Infrared Radiometer In-situ Measurement System
CMUG	Climate Modelling User Group
DAR011	Department of Atmospheric Research radiometer number 11
DARD	Data Access Requirements Document
DMI	Danmarks Meteorologiske Institut
DBCP	Data Buoy Cooperation Panel
ECMWF	European Centre for Medium-Range Weather Forecasts
ECV	Essential Climate Variable
ESA	European Space Agency
GCOS	Global Climate Observing System
GHRSSST	Group for High-Resolution SST
GMPE	GHRSSST Multi Product Ensemble
GODAE	Global Ocean Data Assimilation Experiment
GT MBA	Global Tropical Moored Buoy array
GUM	Guide to Uncertainty of Measurement
HadSST	MOHC SST dataset
ICOADS	International Comprehensive Ocean-Atmosphere Data Set
IR	Infrared
ISAR	Infrared Sea Surface Autonomous Radiometer
JPL	Jet Propulsion Laboratory
L2	Level 2 product
L2P	L2 processed
L3	Level 3 product
L3U	L3 uncollated
L4	Level 4 product
M-AERI	Marine Atmospheric Emitted Radiance Interferometer
MetOp	Meteorological Operational (EUMETSAT)
MD	Match-up Dataset (single-sensor)
MMD	Multi-sensor Match-up Dataset
MMS	Multi-sensor Match-up System

<b>Acronym</b>	<b>Definition</b>
MOHC	Met Office Hadley Centre
NNR	Near-Nulling Radiometer
NOAA	National Oceanic and Atmospheric Administration
OSI-SAF	Ocean & Sea Ice Satellite Application Facility (EUMETSAT)
OSTIA	Operational Sea Surface Temperature and Sea Ice Analysis
PIRATA	Prediction and Research Moored Array in the Atlantic
PSD	Product Specification Document
PVIR	Product Validation and Intercomparison Report
PVP	Product Validation Plan
RAMA	Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction
RR	Round Robin
SISTeR	Scanning Infrared Sea Surface Temperature Radiometer
SoW	Statement of Work
SSD	System Specification Document
SST	Sea Surface Temperature
ST-VAL	Satellite SST Validation Technical Advisory Group (GHRSSST)
SVR	System Verification Report
TAO	Tropical Atmosphere Ocean
TRITON	
UCR	Uncertainty Characterisation Report
UoL	University of Leicester
UoE	University of Edinburgh
URD	User Requirements Document
VOS	Voluntary Observing Ship



## 2. DEFINITIONS

The following definitions are used throughout this document:

**Error:** result of a measurement minus a true value of the measurand. Generally, the “true” value of the error is not known.

**Uncertainty:** Is a parameter, associated with the result of a measurement that characterises the dispersion of the values that could reasonably be attributed to the measurand (given the measurement, in the light of our understanding of the sources of error in the measurement). Here, the parameter is the standard deviation of the dispersion, which is a confidence of 68% or ( $k=1$ ).

**Discrepancy:** The difference between the result and the validation value.

**(Relative) Bias:** The mean value of the discrepancy.

**Accuracy:** For the term “accuracy” there seems to be two definitions in common circulation. In RD.150, the Global Climate Observing System (GCOS) considers accuracy to be measured by “the bias or systematic error of the data, i.e., the difference between the short-term average measured value of a variable and the truth” where the average referred to has been sufficient to render the random uncertainty in the measured value negligible. In contrast, the definition from the Guide to Uncertainty of Measurement (GUM) [RD.191] is also used, whereby accuracy is “the closeness of agreement between the result of a measurement and a true value of a measurand” and therefore a measurement can be inaccurate either by virtue of a large systematic error or because it has a large random uncertainty. We find it useful to have a term available that distinguishes systematic and random uncertainty, and therefore in SST\_CCI documents accuracy refers to the estimated magnitude of the systematic error (true bias).

**Precision:** The difference between one result and the mean of several results obtained by the same method, i.e. reproducibility (includes non-systematic errors only).

**Calibration:** The process of quantitatively defining the system response to known, controlled system inputs

**Validation:** The process of assessing by independent means the quality of the data products (the results) derived from the system outputs.

**Skin Sea Surface Temperature (SST-skin):** The temperature measured by an infrared radiometer typically operating at wavelengths 3.7-12  $\mu\text{m}$  (chosen for consistency with the majority of infrared satellite measurements) that represents the temperature within the conductive diffusion-dominated sub-layer at a depth of ~10-20  $\mu\text{m}$ .

**Sub-Skin Sea Surface Temperature (SST-subskin):** The subskin temperature represents the temperature at the base of the conductive laminar sub-layer of the ocean surface.

**Depth Sea Surface Temperature (SST-depth):** Measurements of water temperature beneath the SSTsubskin, measured using a wide variety of platforms and sensors such as drifting buoys, vertical profiling floats, or deep thermistor chains at depths ranging from  $10^{-2}$  -  $10^3$  m. Here, the depth will usually be that associated with a drifting buoy (of order 20 cm) or a moored buoy (of order 1 m).

The SST\_CCI PVIR is written on the basis of these definitions.

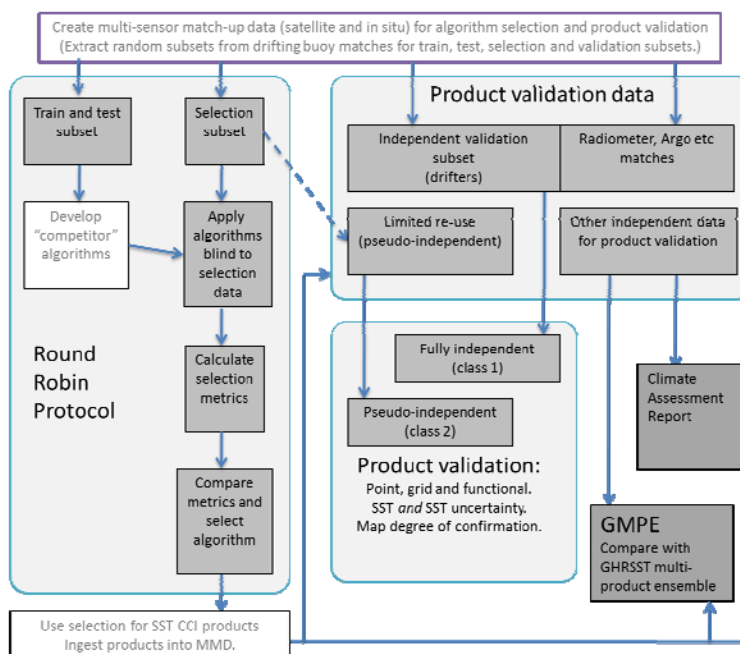


### 3. SUMMARY OF ACTIVITIES

The activities described in this document cover:

- Validation of SST\_CCI ATSR, AVHRR and analysis products, which were performed using independent using high quality SST measurements made in situ from a number of sources
- Intercomparison of the SST\_CCI analysis product against other such data within the GHRSSST GMPE.

The plan for these activities ensures rigour at all points, including independence of algorithm development from validation/assessment (both for data and people). It is inevitably rather complex, given several activities and multiple satellite and in situ data streams. A summary of the process of entire algorithm selection, product validation, intercomparison and climate assessment is shown schematically in Figure 3-1.



**Figure 3-1:** Flowchart indicating logical flow of algorithm selection, product validation inter-comparison and climate assessment for the SST\_CCI project. Activities and data sets specified in this document are in dark-grey boxes. The top box represents the multi-sensor match-up system which is a key source for data throughout. The arrows down from it represent the extraction of distinct subsets of data used for the activities that follow as indicated in the remainder of the diagram.

The process starts (top of figure) with the generation of the multi-sensor match-up database, which was the source of subsets of data used for both algorithm selection and product validation (Section 4). The SST\_CCI products will also be compared to other SST analyses (Section 6) and will undergo a climate assessment as summarised in the climate assessment report (CAR, RD.331).

### 3.1 Multi-sensor match-up database

A multi-sensor match-up dataset (MMD) is a set of temporal and spatial coincidences between multiple satellite datasets of both brightness temperatures and SST retrievals and time series of SST from in situ sensors. For the SST\_CCI project we have pre-matched all required in situ data to the set of satellite datasets required for the two different categories of output products (see Section 3.2). The in situ data comprises data from drifting buoys, the Global Tropical Moored Buoy Array (GT MBA), Argo floats and ship-borne radiometers. Further details on each in situ data type can be found in Section 4 and details on the source, coverage and availability of all datasets used within the SST\_CCI project are given in the SST\_CCI Data Access Requirements Document (DARD; RD.172).

It is important to note that each drifting buoy match-up was assigned to one of four categories:

1. Training: Data for empirical tuning of retrieval coefficients if required; in situ data for these match-ups were included in the RR dataset.
2. Testing: Data for evaluation of retrieval coefficients; in situ data for these match-ups were included in the RR dataset.
3. Selection: 'Blind' data for algorithm selection; in situ data for these match-ups were not included in the RR dataset.
4. Validation: For independent product validation; none of these match-ups were included in included in the RR dataset.

All other multi-sensor match-ups (GT MBA, Argo, radiometers) were assigned to the last category, validation, and are therefore fully independent of algorithm development. Drifting buoys are considered as being independent (category 4) or pseudo-independent (all categories) as the final choice of retrieval algorithms are not actually tied to drifters.

As part of the SST\_CCI activities all drifting buoy measurements underwent a series of quality control tests. The final subset of high quality drifting buoy validation match-ups, along with the GT MBA, Argo and ship-borne radiometers, comprise the reference dataset for SST\_CCI Phase I. Further details of the quality control tests and the final choice of reference data are given in Section 4.

### 3.2 SST\_CCI Products

Following an extensive user requirements review, which is summarised in the SST\_CCI User Requirements Document (URD; RD.171), the project has produced two categories of output products. These are:

1. Long term essential climate variable (ECV) products, where the priorities are for a long, stable climate record formed from two series of sensors. These products will be openly available.
2. Demonstration ECV product, based on wider use of the modern observing system to increase completeness and/or frequency of coverage. The primary purpose of these products is internal assessment regarding: experience with extended processing and the impact on analysis results of additional input data streams.

A summary of the output products that are addressed in this report is provided in Table 3-1. Further details on the content and format of each product are given in the PSD (RD.175).

In total there are twelve products validated and evaluated in this report:

- Ten satellite products
  - Long-term ATSR L3U (three products) and AVHRR L2P (seven products)
- Two analysis products
  - Long-term L4 analysis
  - Short-term demonstration L4 – microwave period

Category of product and description	Satellite sensors & data to be used	Level of data to be produced for each sensor (resolution/grid spacing)
<b><u>Long term ECV</u></b>  A long term, stable data record formed from data from the ATSR and AVHRR series of instruments. Will cover the period Aug 1991 to Dec 2010.	ATSR series (ATSR-1, ATSR-2, AATSR); Envisat format	L3U (0.05°)
	AVHRR series global area coverage (GAC) data	L2P (variable, ~4 km at centre of swath)
<b><u>Demonstration ECV</u></b>  A product to assess the impact of using a broader sample of the SST satellite observing system. Produced for a three month demonstration period only in CCI Phase I (June, July & August 2007).	AATSR; Envisat format (subset of long-term ECV)	L3U (0.05°)
	AVHRR series global area coverage (GAC) data (subset of long-term ECV)	L2P (variable, ~4 km at centre of swath)
	AMSR-E (L2P) (additional data stream cf. long-term ECV)	L2P (0.25°)

Table 3-1: Summary of SST\_CCI products.

The demonstration ECV will not be promoted to users as a record useful for climate given its intended purpose. It is assessed by direct comparison of the L4 to the long-term L4 record for the same period as part of the product intercomparison activities.

### 3.3 Uncertainties

A key development within the SST\_CCI project is the provision of enhanced uncertainty information for each pixel or cell in every SST\_CCI product. The enhanced uncertainty information will include estimates of uncertainty components that are uncorrelated between observations, correlated on synoptic spatio-temporal scales, and correlated on large scales. This facilitates a more realistic propagation of uncertainty from L2/L3 products to derivative products with coarser averaging. Details of the approach are available in the SST\_CCI Uncertainty Characterisation Report (RD.229). As the uncertainty information attached to SSTs constitutes part of the product it must be validated in its own right. Details of the uncertainty validation are included in Section 5.

**In all cases, we encourage users to exploit the uncertainty information provided within the SST\_CCI products and their assessment provided within this report for their particular data application.**

### 3.4 Independence of validation activities

It is important to note that the project has been scoped such that nearly all personnel involved with algorithm selection were not involved in product validation, inter-comparison or the climate assessment, and vice versa.

A summary of key personnel and their roles in the project relating to implementation, validation and assessment of the SST\_CCI products is given in Table 3-2.

Personnel	Algorithm Development	Algorithm Selection	Product Validation	Product Intercomparison	Climate Assessment
Merchant and team (UoE)	✓	✓			
Roquet and team (CMS)	✓				
Eastwood (MetNo)	✓				
Hoeyer (DMI)*	✓		✓		
Corlett (UoL)			✓		
McLaren and team (MO)				✓	
Rayner and team (MOHC)					✓

**Table 3-2:** Summary of personnel and their roles in SST\_CCI product implementation, validation and assessment. \*See main text regarding Hoeyer's distinct roles in development and validation

Hoeyer (DMI) contributed a tuned high latitude retrieval algorithm to the algorithm selection process but was not involved in the final selection process. In the end the algorithm was not selected and so independence of personnel in all steps has not been compromised. Moreover, the final validation and evaluation steps have still been carried out by other independent personnel.

### **3.5 Getting Endorsements**

This document has been written using the knowledge experience of the SST\_CCI project team, and on the basis of the best available methods and approaches from the scientific literature. We have sought endorsement of our methods through external peer review of the PVP (RD.272) and through submission of journal articles summarising our findings. Within the CCI programme the PVP was reviewed by the CCI Climate Modelling Users Group (CMUG) and by external review outside of the CCI programme by the GHRSSST Satellite SST Validation Technical Advisory Group (ST-VAL). A subset of PVP (RD.272) metrics has subsequently been adopted by the GHRSSST Climate Data Record Validation Technical Advisory Group (CDR-TAG) for the Climate Data Assessment Framework (CDAF).

### **3.6 Release of Products**

The SST\_CCI products shall be openly released (subject to any CCI data policy) as soon as this document (the PVIR) and the CAR (RD.331) are accepted by ESA.

## 4. PRODUCT VALIDATION

The SST\_CCI products have been validated against validation data that are fully independent (selected drifting buoys, Argo, GTMBA and radiometers) and pseudo independent (drifting buoys) (see Section 4.1.7 and Table 4-1 for definition and list of datasets). Uncertainties in the SST\_CCI products have been taken account of, along with known uncertainties in the independent reference data.

A key requirement in the SoW [RD.164] was for the final product and user assessment to be done by science team members who are not involved in the ECV production. Consequently, key staff from the lead groups involved in the validation and user assessment has had minimal involvement in algorithm development and selection, achieving the independence required (as summarised in Section 4.1.3).

In addition to this report, key findings will be published in the scientific literature in peer review journal articles. We expect papers to be published on (1) the validation of the L2P and L3U products, (2) the GMPE intercomparison, and (3) on the analysis uncertainty estimation and validation. Publication of peer reviewed journal articles is seen as the key step in ensuring scientific acceptance of the SST\_CCI outputs.

### 4.1 Introduction

#### 4.1.1 Definitions

We have adopted the CEOS definitions of validation and verification. Validation is defined by CEOS as the process of assessing, by independent means, the quality of the data products derived from the system outputs, and assess the fitness-for-purpose of the data products. Verification is defined by CEOS as the provision of objective evidence that a given data product fulfils specified requirements.

A list of the key definitions is provided in Section 2.

#### 4.1.2 Reference data

The product validation used reference dataset (pre-defined in the PVP, RD.272) including validation data constituting drifting buoys, the GTMBA, Argo floats and ship-borne radiometers. Details of the reference dataset for SST\_CCI product validation and its quality control procedures are given in Section 4.2.

#### 4.1.3 Rules and responsibilities for objective independent product validation

To ensure objective independent validation the following rules were adopted within the project:

- The overall validation was led by UoL (Corlett), who will also led the product validation

- The Met Office (McLaren and team) led the inter-comparison activities
- DMI (Hoyer) focussed on high latitude validation
- UoE (Bulgin) supported the uncertainty validation activities
- MOHC (Rayner and team) provided the reference dataset
- No other team members participated in product validation aside from the development of tools (Brockmann Consult)
- A set of in situ data was reserved solely for validation and was not used (previously) for algorithm selection or at any other time in the project (the reference dataset)

#### 4.1.4 Validation criteria

The ideal scenario for validation is for the reference measurement to be taken precisely at the time of the satellite overpass. Within the SST\_CCI project we have adopted the current GHRSSST limits such that the reference data are ideally within the satellite pixel within 2 hours of the satellite overpass as a minimum criterion. These limits are based on the current best estimates from the literature for the temporal resolution (Minnett, 1991; RD.234) and the need to validate the uncertainty on a single satellite pixel for the spatial resolution.

#### 4.1.5 Depth/time adjustments

To minimise uncertainties due to temporal matching a combined diurnal/skin-effect model was used to adjust the depth and time of the reference measurement to that of the satellite measurement. In the mean, this will reduce the uncertainty to  $\ll 0.1$  K for a statistically significant sample. The model is the same as that used within the processing chain to create depth SSTs at a standardised local time (ATBD, RD.305), and comprises a skin effect model (Fairall et al, RD.227) and warm layer model (the Kantha and Clayson turbulence closure, RD.263) driven by ECMWF surface winds and fluxes. Section 5.4 for skin; sub-skin to time Section 6.4

#### 4.1.6 Uncertainty verification

As previously mentioned, a key objective of the SST\_CCI project is to provide uncertainty information with each product and to validate both the SST and its associated uncertainty. This is in contrast to the traditional approach in satellite retrievals of SST of using validation to derive uncertainty information. Consequently, users are strongly encouraged to use the uncertainty information provided in the product and not to rely on comparisons to other datasets. To validate the uncertainties we will use the distributions of differences between the SST\_CCI products and the reference dataset and determine if these scale appropriately as a function of the product uncertainties.

In addition, we have provided maps to indicate the *degree of verification* that the validation provides taking into account the uncertainty and availability of the reference data.

The degree of verification maps are provided at 15° resolution for each SST\_CCI product and indicate where we have a very high, high, medium, low and very low degree of confirmation in the SST and its associated uncertainty information provided in the SST\_CCI products from product validation. For further details please see Section 5.6.

#### 4.1.7 Classes of validation

A requirement of the SST\_CCI project (SST\_CCI-UR-QUF-78; RD.171) is to validate the output products using independent reference data. However, this requirement must be offset against the need to validate each product that the SST\_CCI system produces. As the availability of independent data varies considerably over the years (and some data has been used for algorithm selection) the validation will use data on all available spatial and temporal scales. Therefore we define two classes of validation:

1. Independent data: Data not used in algorithm training, test or selection, and therefore both statistically independent and independent of the algorithm development and selection
  - Drifters (10% of all available from 2008 onwards)
  - GTMBA
  - Argo
  - Ship-borne radiometers
2. Pseudo-independent data: Use all drifter match-ups including those used in the algorithm development work. The selected SST algorithms are not tuned to drifting buoys, and in this case these matched data remain statistically independent of the SST CCI products, although not independent of the algorithm development and selection process.
  - Allows improved regional validation

Clearly the degree of verification associated with class 2 validation will not be the same as for class 1. Nevertheless the additional coverage will allow some additional confidence information to be provided, including for SST\_CCI L4, which (like L2 and L3 products, and unlike most L4 analyses) does not use the drifter data.

#### 4.1.8 Types of validation

A further approach to provide additional validation data is to consider the validation as being carried out for three types:

1. Type 1 - 'Point': These are single pixel comparisons to both class 1 and class 2 the reference dataset; the class 1 comparisons provide the highest quality validation and therefore can provide the highest degree of confidence.
2. Type 2 - 'Grid': These are comparisons to HadSST3, which potentially improves the match-up coverage (both temporally and spatially). Also, as this type of comparison uses 'average' in situ data there is likely to be a lower impact from outliers due to poor reference data.



3. Type 3 – ‘Functional’: This final type is needed in order to provide a degree of confidence everywhere, even areas where we have no reference measurements. For this we will look for comparable retrieval regimes stratified by, for example, TCWV. The final set of conditions can only be defined once the type 1 and type 2 analyses have been carried out in order to see what areas remain and what sensitivity each product has.

Within SST\_CCI Phase I only Type 1 comparisons have been used for this initial validation of SST\_CCI products as there is sufficient data coverage across the period analysed. There are no limitations caused by not performing Type 2 and 3 comparisons at this stage; some comparisons between SST\_CCI products and HadSST3 can be found in the CAR [RD.331].

#### 4.1.9 Analysis procedures

All SST\_CCI system outputs have been validated using both independent and pseudo independent point type validation data detailed in Section 4.1.8 noting the degree of independence detailed in Section 4.1.7. Discrepancies and uncertainties were derived using robust and non-robust statistical methods for each type of reference data, and where sufficient match-ups allow Uncertainties are provided for a confidence level of 68% (the “one-sigma” level). All validation was done using the total uncertainty as there are no uncertainty budgets for any of the reference data to allow a more detailed breakdown of the uncertainties. Time series of discrepancies and uncertainties are provided for each SST\_CCI dataset, as well as any dependence on auxiliary data in the MMD (e.g. wind speed), total column water vapour and satellite and solar zenith angles. Further details of the validation methodology are provided later in this section.

The results from the independent validation were compared to the products uncertainties to identify areas where they are self-consistent. All results will contribute to the degree of verification maps detailed in Section 5.6.

## 4.2 Reference dataset

### 4.2.1 Introduction

Validation is the “assessment by independent means of the quality and fitness for purpose” of the SST\_CCI products. This means, amongst other things, that the reference data should be independent of the SST\_CCI products, where possible. Where this is not possible, the following hierarchy of possible reference data will be adopted:

1. Independent in situ data
2. Other in situ data
3. Large scale comparisons with other satellite data
4. Large scale comparisons with historic data sets, climatologies

This section defines the reference data set to be used for validation of the SST\_CCI products, giving an overview of the data and an assessment of their quality, followed by an explanation of the rationale behind the choice of reference data.

When considering possible reference sources, consideration must be given to the nature of the SST being assessed. For satellite SST retrievals produced from infrared radiances, the SST is equivalent to the temperature at a depth of  $\sim 10 \mu\text{m}$  and is referred to as the skin SST; for satellite SSTs produced from microwave radiances, the SST is equivalent to the temperature at a depth of  $>100 \mu\text{m}$  and is a weighted average of the temperatures through the skin layer and into the sub-skin region beneath. The deviation between skin and sub-skin reduces to a mean bias of  $-0.17 \text{ K}$  when the surface wind speed is  $> \sim 6 \text{ ms}^{-1}$ , and so surface wind speed data is an essential component of any reference data set for satellite SST uncertainty determination and is provided in the MMD.

Ideally, the reference source for assessing the quality of the satellite data should be a measurement at a depth that is as close as possible to that provided by the satellite. Indeed, where possible, it should be the same as that provided by the satellite, which is currently achievable for infrared sensors using ship-borne radiometers, and potentially for microwave sensors using aircraft mounted radiometers (see for example <http://www.prosensing.com/Hurricane%20Wind%20Speed%20Radiometer.htm> as used by the NOAA National Hurricane Centre).

The current reference data set used by GHRSSST is that provided by surface drifting buoys. Although the uncertainty of this dataset is not traceable to the SI temperature standard, it has been chosen due to its significantly improved global coverage compared to other potential reference datasets. Other potential reference data include ship-based radiometers, moored buoys, and conventional ship measurements from engine room intakes or hull-mounted sensors; the GTMBA is usually considered separately from other moored buoys because they are in the open ocean and far from the coastal regions which often present particular difficulties for the accurate measurements of SST from space, and where most other moored buoys are deployed.

#### 4.2.2 Overview of data sources

Each reference data source is detailed in turn, with an assessment of their quality, sourced either from the literature or unpublished analysis by the project's Climate Research team.

For some data sources, uncertainty is divided into uncertainty arising from inter- and intra-platform errors. Inter-platform errors are random measurement errors, which are uncorrelated between different locations. Intra-platform errors are measurement errors which are correlated from location to location, because they persist as an individual drifting buoy or ship moves. Correlated intra-platform errors do not reduce as measurements are aggregated over space and time, whereas uncorrelated random inter-platform errors do.

There are three principal types of platform measuring SST in situ: ships, drifting buoys and moored buoys. In addition, Argo profiling floats provide useful numbers of high quality near surface measurements since 2000. Ships, buoys and Argo floats are identified by a unique call sign, or other identifier.

The sampling characteristics of these platform types are quite distinctive. Ships travel between ports, along shipping lanes, making regular observations, so the observations from a single ship can provide a representative sample for a large area along the shipping lane. Drifting buoys drift along with the prevailing surface currents, but they do not often travel far. They typically take hourly observations and provide dense sampling along a limited trajectory. Drifter deployments are designed to provide a fairly uniform coverage of the oceans, but there are places where they do not go. Similarly, Argo floats travel along with currents at depth and sample the ice-free oceans. Moored buoys take regular measurements at a fixed point.

In the early 1990s, Voluntary Observing Ships (VOS) provided the densest in situ measurements of SST. From around 1998, drifting buoys became more numerous. Argo and ship-born radiometer measurements have become available in any numbers only since 2000. Accordingly, our reference data set is heterogeneous in nature both in space and time (see for example Kennedy et al., 2012 (RD.243). Consequently, the validation of SST\_CCI products is somewhat challenging and requires us to use all available reference data sources and to properly consider the differences in depth, time and space between the various datasets.

#### 4.2.2.1 SST at approximately 0.2m depth from drifting buoys

##### 4.2.2.1.1 Background

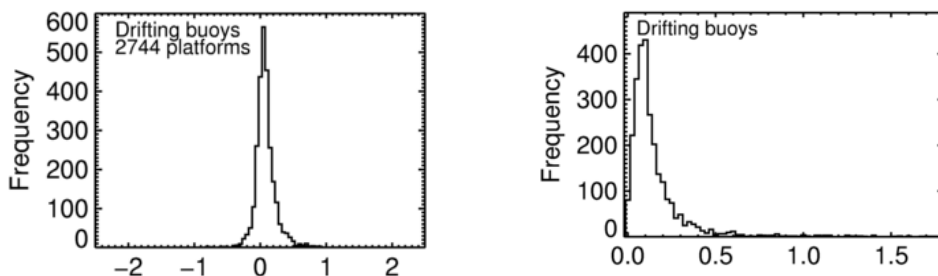
Drifting buoys consist of a surface float, approximately 30 cm in diameter, housing satellite communication and SST measurement equipment, along with a sub-surface sea-drogue spanning the upper 10 to 15 m of the water column, which allows the buoy to follow the integrated water movement over the depth of the drogue and the effect of surface wind and wave motion on the surface float (Lumpkin and Pazos, 2006; RD.058). The SST sensor is embedded in the underside of the buoy and measures at a depth of approximately 20 cm in calm seas. Movement of the buoy and the action of waves mean that the measurement is representative of the upper 1m of the water column (Lumpkin and Pazos, 2006; RD.058). The Global Drifter Program facilitates hourly global observations of SST, based on 15-minute averages of measurements. In June 2010 there were approximately 3000 buoys reporting hourly SST observations.

The major change in the network of drifting buoys since its inception has been a transition from a network containing a mixture of instrument designs (prior to 1993) to a standardisation of instrumentation post-1993 (Lumpkin and Pazos, 2006; RD.058). The effect of this change in instrumentation has not yet been assessed. Biases in the drifting buoy data are known to arise from a lack of maintenance of the buoys, leading to variations in the accuracy of their SST measurements (O'Carroll et al., 2008; RD.246). Since the buoys are not routinely recovered, and owing to a lack of independent SST data, the post-calibration of buoy measurements has not so far been possible (Emery et al., 2001; RD.245).

Since many retrieval algorithms utilising Advanced Very High Resolution Radiometer (AVHRR) measurements rely on SST measurements from drifting buoys to provide a "ground truth" for the regression-based retrievals, drifting buoys are not independent from these estimates.

##### 4.2.2.1.2 Accuracy

Kennedy et al (2012; RD.243) utilised coincident match ups between drifting buoy SST measurements and SST retrieved from Along Track Scanning Radiometer (ATSR) measurements (adjusted to sub-skin depth) for 2002-2007 to assess inter- and intra-drifter uncertainties (Figure 4-1).



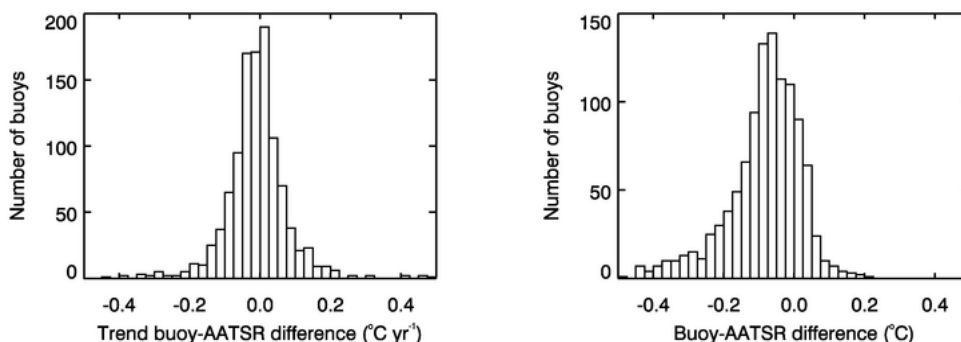
**Figure 4-1:** Distribution of inter-platform (left) and intra-platform (right) errors for drifting buoys between August 2002 and December 2007. Only platforms with more than 25 ATSR-drifter pairs are shown. (Figure adapted from Kennedy et al, 2012; RD.243.)

A range of intra- and inter-drifter errors was found. The inter-platform errors exhibited a very peaked distribution with standard error of about 0.29 K. The intra-platform errors displayed a long positive tail and the distribution is not easily summarised by one number.

Currently, uncertainties are not available for each drifter in the archive. On-going projects at the Met Office Hadley Centre and the University of Reading are seeking to address this issue.

#### 4.2.2.1.3 Stability

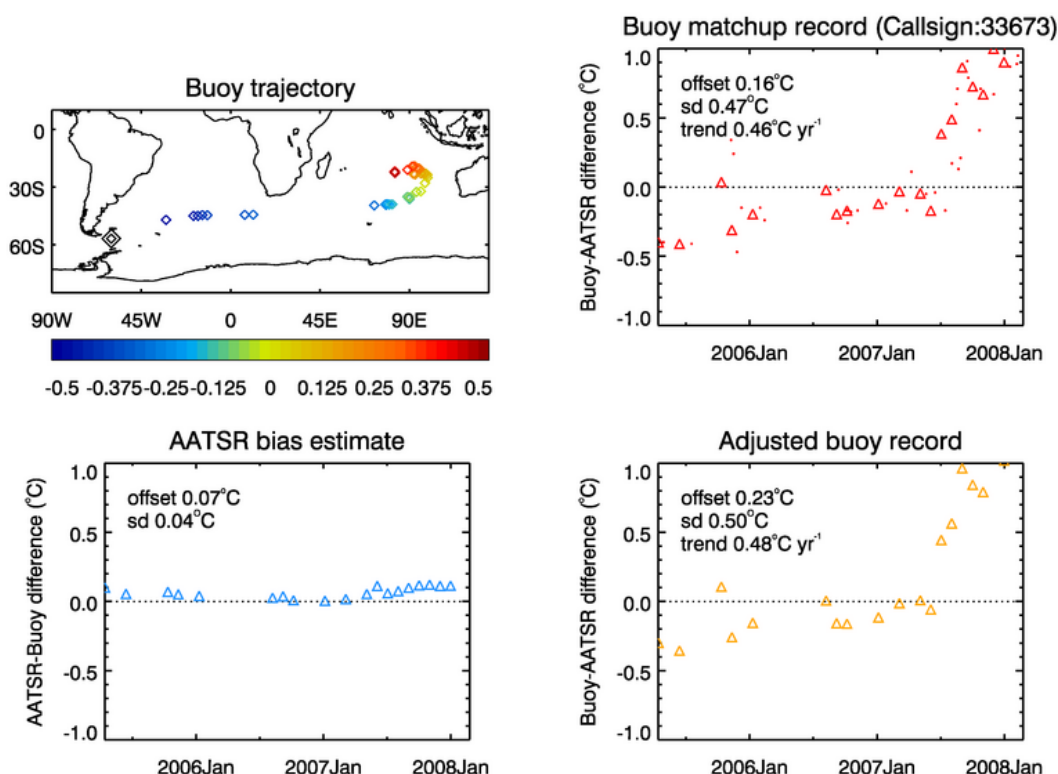
A recent study has examined differences between each of a pair of temperature sensors attached to a set of drifting buoys (Reverdin et al., 2010; RD.244). Drifting buoys were equipped with a standard thermistor, as deployed on the majority of Surface Velocity Program (SVP) drifters, and an additional high-quality platinum temperature probe, with the latter used to assess the accuracy of the former. The study by Reverdin et al. (2010; RD.244) revealed evidence of bias-offsets and a calibration drift in the thermistor-reported temperature for two drifting buoys (from a sample of 16) that were at sea for approximately one year. In regions sparsely sampled by the in-situ array, degradation of drifting buoy temperature sensors in this way could potentially lead to misleading validation results.



**Figure 4-2:** Comparisons between SST measured by drifting buoys over their lifetime and retrieved from AATSR measurements. Left: relative trends ( $^{\circ}\text{C}/\text{yr}$ ). Right: average differences ( $^{\circ}\text{C}$ ).

Work at the Met Office Hadley Centre (Atkinson et al., 2013, RD.326) examined the SST measured by drifting buoys over their lifetime, by comparison to Advanced ATSR matchups (as contained in a preliminary ATSR Reanalysis for Climate (ARC) data set). Once the matchup dataset had been created, unique drifting buoys were identified based on their WMO call signs. Each drifting buoy is assigned a WMO call sign for identification of the buoy on the GTS, but these call signs are often re-used after each buoy fails, with re-use typically occurring no faster than three months.

Relative trends between the SST as measured by each drifter through its lifetime and as retrieved from the AATSR measurements were calculated. The left-hand panel of Figure 4-2 shows the distribution of these trends for all the buoys examined. Some buoys exhibited large relative trends. Others show large constant offsets (right hand panel). Differences, along the track of the drifter, between the AATSR retrievals and drifters, i.e. biases in the AATSR data, were removed before analysis.



**Figure 4-3:** Calibration drift of buoy 33673, 2005-2007. Top left: trajectory of the buoy through its lifetime (about 2.5 years). Top right: difference between SST as measured by the buoy and as retrieved from coincident AATSR measurements. Bottom left: estimate of AATSR SST bias from comparison with other buoys. Bottom right: difference between SST as measured by the buoy and as retrieved from coincident AATSR measurements minus the AATSR bias.

Examining the distribution of per-buoy annual calibration drifts (left hand panel of Figure 10.2) we see an approximately normal distribution, with mean trend  $0.00^{\circ}\text{C yr}^{-1}$ . Fewer than 10% of buoys display trends exceeding  $\pm 0.1^{\circ}\text{C yr}^{-1}$ . These calibration drifts are less prevalent than average buoy offsets (right hand panel of Figure 4-2).

Some buoys seem reasonably stable, but then exhibit large SST biases in the period just before they stop reporting (Figure 4-3). Routine quality control of buoy data is performed by Data Buoy Cooperation Panel (DBCP) monitoring centres such that buoys displaying

large SST biases are removed from the GTS with a typical timescale of several weeks following the failure of the instrument.

Assessment of drifts, biases and root mean square errors in the calibration of individual buoys by reference to ATSR series retrievals and OSTIA reanalysis and operational data are continuing as part of the FP7 project ERA-CLIM. Periods where individual buoy data are found to be inaccurate are excluded from the SST\_CCI reference data set. Methods used in creating blacklists, such as those maintained by Météo France and the Met Office, are utilised to exclude erroneous measurements.

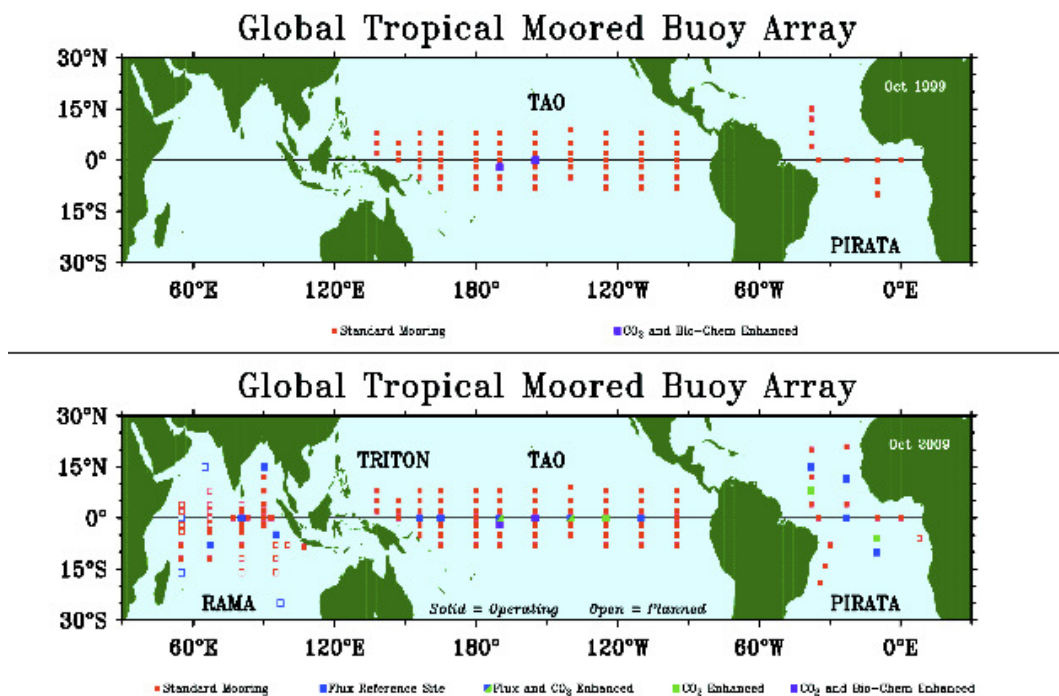
#### 4.2.2.2 SST at approximately 1 m depth from moored buoys

##### 4.2.2.2.1 Background

Moored buoys are normally relatively large and expensive platforms. Data are usually collected through one of Argos, Iridium, ORBCOMM, GOES or METEOSAT, transmitted in real-time and shared on the GTS of WMO. They are generally upgraded or serviced yearly. Many different designs exist for moored buoys depending on the ocean area. Moored buoys come in a wide variety of shapes and sizes, from over 12 m to the 1.5 m fixed buoys deployed in the North Sea. (<http://www.jcommops.org/dbcp/platforms/types.html>)

Since the 1980s, a moored buoy array has been built in all three tropical oceans. The Global Tropical Moored Buoy Array (GT MBA) comprises the Tropical Atmosphere Ocean/Triangle Trans-Ocean Buoy Network (TAO/TRITON) in the Pacific, the Prediction and Research Moored Array in the Tropical Atlantic (PIRATA), and the Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction (RAMA) in the Indian Ocean. Most of the buoys in the tropical arrays are the ATLAS mooring, developed in the 1980s, deployed in depths of up to 6000 metres.



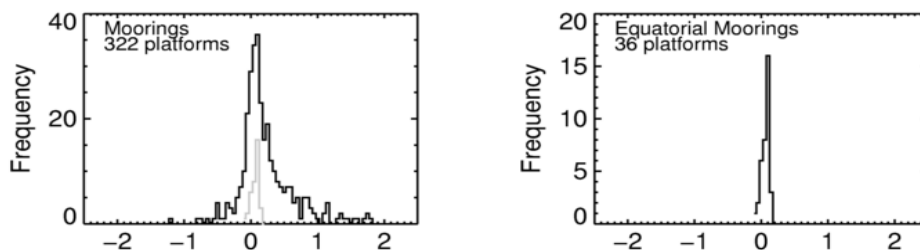


**Figure 4-4:** The evolving GTMBA, showing both existing and planned moorings (as of 2009). The upper panel shows the arrays as they existed in 1999; the lower panel shows the arrays in 2009 (solid circles) plus planned additions (open circles). (Taken from <http://www.atmos.washington.edu/~ackerman/GTMBA.pdf>)

In addition to the GTMBA, moorings are maintained off nations' coasts for weather forecasting purposes.

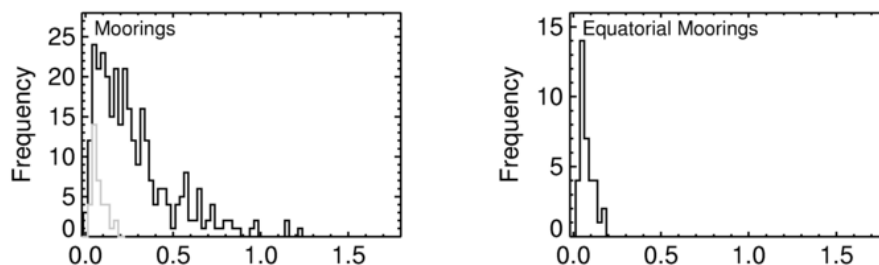
#### 4.2.2.2 Accuracy

Kennedy et al (2012; RD.243) utilised coincident match ups between moored buoy SST measurements and SST retrieved from ATSR measurements (adjusted to sub-skin depth) for 2002-2007 to assess inter- and intra-platform errors (Figure 4-5 and Figure 4-6).



**Figure 4-5:** Distributions of inter-platform errors for moorings (left, equatorial moorings are shown in grey) and equatorial moorings only (GTMBA, right) between August 2002 and December 2007. Only platforms with more than 25 ATSR-*in situ* pairs are shown. (Figure adapted from Kennedy et al, 2012; RD.243.)

Matches between some coastal moorings and the ATSRs can exhibit large differences (Figure 4-6). This is likely partly due to a mismatch of scales in these regions, where the SST is relatively variable compared to the tropics. However, it may also indicate problems with the buoys themselves; more investigation is needed.



**Figure 4-6:** Distributions of intra-platform errors for moorings (left, equatorial moorings are shown in grey) and equatorial moorings (GTMBA, right) between August 2002 and December 2007. Only platforms with more than 25 ATSR-*in situ* pairs are shown. (Figure adapted from Kennedy et al, 2012; RD.243.)

Currently, uncertainties are not available for each mooring.

#### 4.2.2.3 SST<sub>skin</sub> from shipborne radiometers

##### 4.2.2.3.1 Background

There are a number of infrared radiometers, designed to measure SST<sub>skin</sub> from a ship. Two provide particularly long records: the Marine-Atmospheric Emitted Radiance Interferometer (M-AERI, Minnett et al, 2001; RD.052) and the Infrared SST Autonomous Radiometer (ISAR, Donlon et al, 2008; RD.047).

The M-AERI has been measuring SST<sub>skin</sub> on board the Explorer of the Seas since 2000. It is an infrared spectroradiometer. The radiometric calibration of the M-AERI is accomplished using two internal blackbody cavities. The absolute accuracy of the M-AERI calibration is monitored by episodic use of a NIST-certified water bath blackbody calibration target. Residual errors in the retrieved temperature from the M-AERI measurements at temperatures characteristic of the sea surface are typically <0.03 K (Minnett et al. 2001; RD.052).

The ISAR is capable of measuring in situ sea surface skin temperature accurate to  $\pm 0.1$  K root mean squared error (Theocharus et al, 2010; RD.242) for deployment periods of up to 3 months. It uses two precision calibration blackbody cavities. Five ISAR instruments have been built and are in sustained use in the United States, China, and Europe (Donlon et al., 2008; RD.047).

Other radiometers have been used to measure SST<sub>skin</sub> from research vessels:

- the Scanning Infrared Sea Surface Temperature Radiometer (SISTeR), a radiometer with narrowband filters centred at 3.7, 10.8, and 12.0  $\mu\text{m}$ ;

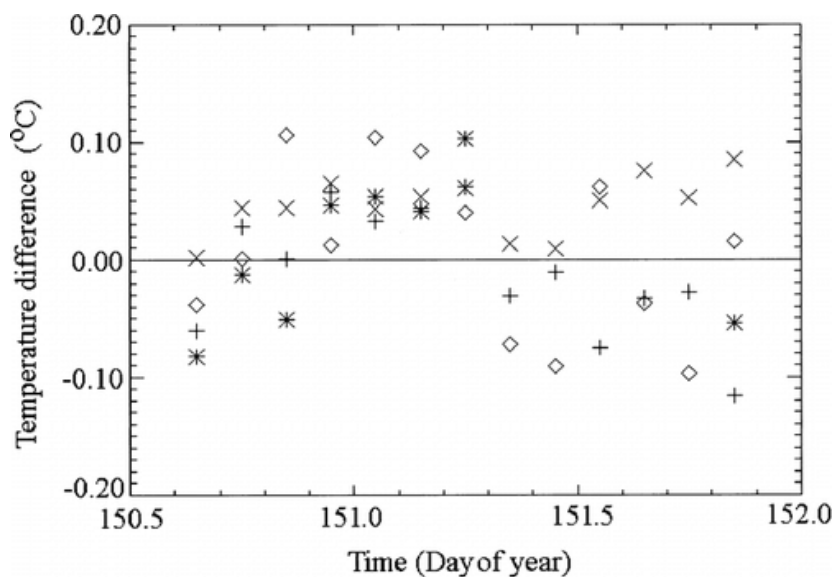


- the Jet Propulsion Laboratory (JPL) Near-Nulling Radiometer (JPL NNR), a self-calibrating sensor which detects radiation with wavelengths between 7.8 and 13.6  $\mu\text{m}$ ;
- the Calibrated Infrared In situ Measurement System (CIRIMS), with a design accuracy of  $\pm 0.1$  K, passing radiation with wavelengths 9.6–11.5  $\mu\text{m}$  and
- the DAR011 radiometer, a single-channel, self-calibrating, infrared radiometer passing radiation with wavelengths between 10.5 and 11.5  $\mu\text{m}$ .

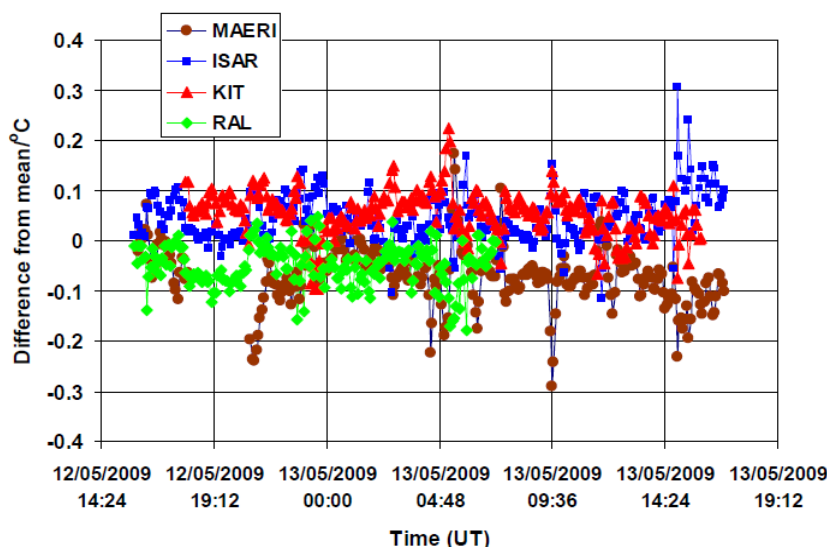
Ship radiometers do not provide global coverage. However, they are complementary to reference measurements at depth because they provide direct measurements of SST<sub>skin</sub>.

#### 4.2.2.3.2 Accuracy

Intercomparisons between the SST<sub>skin</sub> measured by various radiometers were carried out in 2001 (Barton et al, 2004; RD.050) and 2009 (Theocharus et al, 2010; RD.242). Both studies involved laboratory measurements against NIST or NPL standard blackbodies and day- and night-time measurements either at sea or of sea water. They both showed that the radiometers measure SST largely within  $\pm 0.1$  K of each other. Figure 4-7 is taken from Barton et al (2004; RD.050).



**Figure 4-7:** The differences between the M-AERI skin SST and those derived using the other radiometers averaged over the intercomparison period: ISAR-5, \*; SISTeR, x; JPL, and DAR011, + Reproduced from Barton et al (2004; RD.050), their Figure 5 and Figure 4-8 from Theocharus et al (2010; RD.242).



**Figure 4-8:** Difference of SST measured by M-AERI, ISAR, KIT and SISTeR from their mean. Taken from Theocharus et al (2010; RD.242), their Figure 3.11.9.

Currently, uncertainties are not available for each individual radiometer.

#### 4.2.2.4 Near-surface temperature measurements from Argo

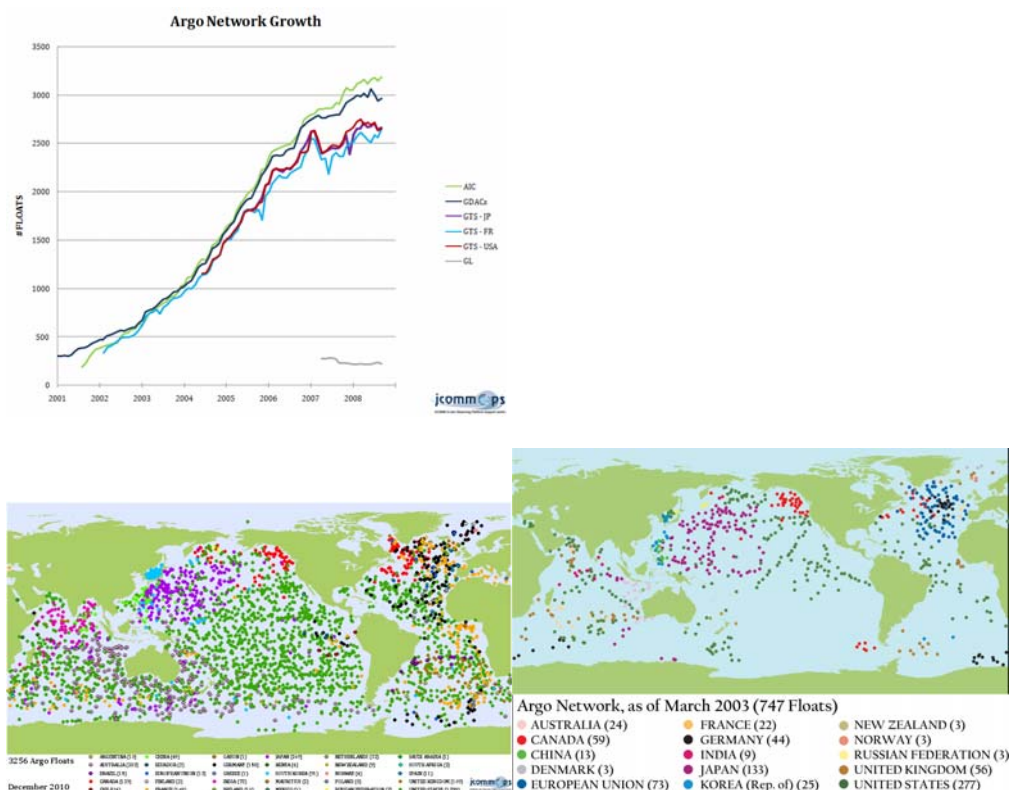
##### 4.2.2.4.1 Background

Argo is a global array of profiling floats measuring ocean temperature and salinity. From 1999 Argo data downloaded from the global data assembly centres (Coriolis or USGODAE) are included in the MOHC EN3 data set of quality controlled ocean temperature and salinity measurements (<http://www.metoffice.gov.uk/hadobs/en3/>). Argo data were collected and made freely available by the International Argo Project and the national initiatives that contribute to it (<http://www.argo.net>).

There are three models of profiling float used extensively in Argo. All work in a similar fashion: at typically 10-day intervals, the floats pump fluid into an external bladder and rise to the surface over about 6 hours while measuring temperature and salinity. Satellites determine the position of the floats when they surface, and receive the data. The bladder then deflates and the float returns to its original density and sinks to drift until the cycle is repeated.

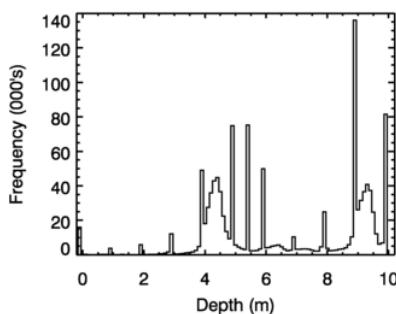
The array currently comprises over 3000 floats, which are distributed over the global oceans at an average 3-degree spacing. Floats have lifetimes of 4-5 years. The temperature data are reported to be accurate to a few millidegrees over the float lifetime (<http://www.argo.ucsd.edu/>).

Figure 4-9 shows how the coverage of the array has increased since 2001.



**Figure 4-9:** The evolution of the Argo array. Top: the number of active floats, 2001-2008. Bottom: snapshots of the Argo network in March 2003 (left) and December 2010 (right). Source: <http://wo.jcommops.org/cgi-bin/WebObjects/JCOMMOPS>.

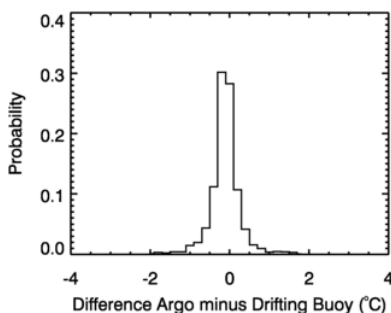
We will use only the near-surface measurements in our reference data set. These are available at various depths between the surface and 10 m, as shown in Figure 4-10.



**Figure 4-10:** Distribution of sampling depths over the upper 10m for Argo profiling floats, from the EN3 data set for the period 2000-2009.

#### 4.2.2.4.2 Accuracy

Coincident near-surface measurements from drifting buoys and Argo from 2000-2009 were examined (Figure 4-11). They have zero mean discrepancy and the distribution of differences is highly peaked, indicating that Argo near surface measurements are of comparable quality to those of drifting buoys.



**Figure 4-11:** Distribution of Argo-minus-drifting buoy differences from co-incident (within 10 km and 1 hour) observations (4-6 m, shallowest selected) 2000-2009. Image provided by Rob Smith (Met Office Hadley Centre)

Figure 4-11 summarises a comparison between Argo and drifting buoys, i.e. two measurements at depth (albeit at different depths). At all points in the SST\_CCI product validation, the different depths represented by the reference data and the SST\_CCI products being validated will be reconciled to ensure a like-for-like comparison.

However, it should be noted that some Argo data are subject to biases in reported pressures. These biases are usually less than 5db, but occasionally can be larger (> 20db, [http://www.argo.ucsd.edu/Argo\\_Data\\_and.html](http://www.argo.ucsd.edu/Argo_Data_and.html)). These bias errors are being removed by the reprocessing of historical Argo data at Regional Data Assembly Centres. Adjusted pressure data are stored in the PRES\_ADJUSTED variable, where this is available.

A subset of Argo floats cannot be corrected as the pressure bias was not transmitted by the floats. Within this subset, some will have a high probability of developing large biases. These floats are identified in the delayed-mode processing of Argo data and are flagged with higher pressure errors (20 db) in the PRES\_ADJUSTED\_ERROR variable.

Currently, uncertainties are not available for each Argo float.

#### 4.2.2.4.3 Stability

In addition to the pressure bias issues noted above, [http://www.argo.ucsd.edu/Argo\\_Data\\_and.html](http://www.argo.ucsd.edu/Argo_Data_and.html) cautions users that APEX profile data need corrections for a drift in their pressure sensors. The correction is estimated to be, on average, -2 dbar in 2003, decreasing to about 0 dbar in 2008 (due to improved sensor stability). However, a few older individual floats may have profiles with pressure offsets of over 10 dbar. Some APEX floats truncate any negative surface pressure drifts to zero. These floats, if their pressures drift towards negative values, have unknown pressure bias and are uncorrectable. Lists of WMO IDs of "uncorrectable" floats can be found at [http://www.marine.csiro.au/~cow074/quota/argo\\_offsets.htm](http://www.marine.csiro.au/~cow074/quota/argo_offsets.htm).

#### 4.2.3 Criteria for selection

As discussed in the previous section, the in situ SST observing array has evolved over the past few decades and is heterogeneous. By necessity then, our reference data set is also heterogeneous in space and time. We have used the hierarchy given in the ESA CCI

guideline document (RD.169, see Section 10.1) to help to determine our strategy for definition of the reference data set, i.e. what to include where and when.

We include only in situ measurements of SST in our reference data set because there are no independent satellite retrievals of SST whose record is sufficiently long or whose uncertainties are sufficiently well-characterised to help in times or locations of sparse in situ measurements (see Kennedy et al 2011a and b, RD.210 and RD.211, for example). Our User Requirements gathering exercise demonstrated that the users consulted were either in agreement with our proposed reference data set, or had no opinion (RD.171). Comparing to SST analyses, where these have been made globally complete by interpolation, might be an option were it not for the fact that such products usually incorporate ATSR or AVHRR retrievals and so provide no independence at all. Where we have few in situ measurements to create collocated match ups, we will compare to the gridded, uninterpolated HadSST3 data set and/or widen our area of comparison.

Practically, a reference data set needs to have stability, longevity and accessibility. As far as accessibility is concerned, all the data sources discussed above are freely available, at least for research purposes. All aforementioned data sources have records of at least a decade, some much more than this, and will continue to supply measurements into the future. Stability can be ensured through application of knowledge gained in the ARC project, or being developed in the ERA-CLIM project. The consequences of instability of measurement type can be circumvented using our knowledge of relative biases between measurement methods.

There is a requirement to demonstrate the stability, bias and accuracy of the SST\_CCI products on the 100km spatial scale (RD.171). To get an idea of where and when sufficient observations for meaningful comparison on this scale might be available, without actually performing matchups, we can examine gridded fields of available numbers of drifting buoy, tropical mooring and VOS measurements on a 1° latitude by longitude grid. In the figures that follow, grid boxes are coloured according to the measurement type if there are at least 30 measurements available in the grid box in the month displayed (from central limit theorem).

Argo and ship-borne radiometers will also be included in the reference data set, but their relative scarcity and recent availability mean they are neglected for the purposes of this demonstration. We select measurement types based on data quality (see previous sections for quantification of reference data quality), i.e. if there are > 30 drifting or moored buoy measurements available, we indicate that this would be the measurement type of choice in this location at this time. As mentioned above, keeping our assessment against the reference data set segregated by measurement method will allow us to exploit our knowledge of expected bias between the in situ and satellite measurements.

#### 4.2.4 Additional quality control

The ICOADS reference data used within the SST\_CCI Phase 1 is a blend of observations taken from ICOADS 2.5 (Woodruff et al., 2011; RD.332) and Met Office Hadley Centre QC flags. The QC flags provided have been produced by the HadISST2 QC system. The general QC procedures are described in Rayner et al. (2006; RD.72) and the high resolution background climatology and land-sea mask used by this system is described in Rayner et al. (2013; RD.334). This system carries out the following suite of checks: (i) observations are checked for a meaningful location, date and time and that they are not surrounded on all sides by land, (ii) each platform with an individual callsign is tracked to verify its reported position, speed and direction (those without a callsign or with a generic callsign, e.g. SHIP, are passed unchecked), (iii) each SST observation is checked that it is above the freezing point of seawater and within  $\pm 8^{\circ}\text{C}$  of the 1961-1990 background climatology interpolated to that day, (iv) each SST observations has a “buddy check”

applied which compares the value of an individual SST anomaly to the mean anomaly from neighbouring observations; individual observations differing too much from their neighbours are flagged as bad. The HadISST2 QC flags have been supplemented as follows:

1. Drifting buoy SST observations from ICOADS deck 715 have been blacklisted as investigation suggests they are of variable quality. (In ICOADS, a deck originally referred to a punched card deck, but is now used as the primary field to track ICOADS data collections).
2. Drifting buoy and ship SST observations have an additional QC flag set which follows the procedures described in Atkinson et al. (2013; RD.326). This flag is generated by tracking the quality of observations made by individual drifting buoys and ships over time using the Met Office Operational Sea surface Temperature and sea Ice Analysis (OSTIA) as a reference (a globally complete satellite based analysis). It differs from the SST checks described above in that observation quality is tracked over time to detect biases/instrument failures etc., rather than assessing observations individually. Drifting buoys observations are flagged where they are deemed to be too biased or too noisy, or a buoy is deemed to be out of water having run aground or been picked up. Ship observations are flagged when observations from a particular ship (identified by its callsign) are deemed unreliable (i.e. if a ship callsign is blacklisted all observations from this ship are flagged). In general, ship observations are of variable quality and this flag is intended to reduce ship observations to a higher quality subset.

#### 4.2.5 Content of Reference Dataset for Product Validation

The content of the reference dataset for product validation is given in Table 4-1. Estimates of uncertainty for each dataset are provided.

Data type	Time period	Coverage	Comment	Uncertainty	Reference
Ship-borne IR radiometers	2000 - 2010	Caribbean Sea; Bay of Biscay	Independent SSTskin	0.1 K	Barton et al., (2004; RD.050)
Argo floats	2000 - 2010	Global <sup>#</sup>	Independent SSTdepth	0.005 K	Oka and Ando et al. (2004; RD.355)
GT MBA	1991 - 2010	Tropics	Independent SSTdepth	0.1 K	Kennedy et al (2012; RD.243)
Drifting buoys	2008-2010* 1991-2010	Global <sup>#</sup>	Independent and pseudo-independent SSTdepth	0.2 K	O'Carroll et al (2008; RD.246)

**Table 4-1:** Content of SST\_CCI reference dataset for product validation

\* Independent drifting buoy data i.e. data not used in algorithm selection will only be available from 2008 onwards.

<sup>#</sup> Data are not truly “global” but cover majority of Earth’s oceans.



### 4.3 Validation of AVHRR products using the MMS

The first set of SST\_CCI data products validated were the L2P products from the AVHRR series. Seven different datasets were included:

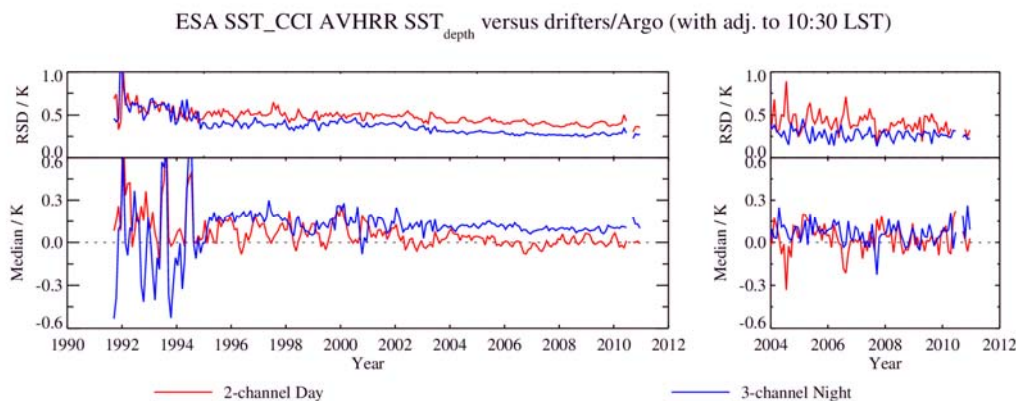
- AVHRR-MTA: Data from the AVHRR sensor on the METOP-A platform covering the period from 2007 to 2010.
- AVHRR-18: Data from the AVHRR sensor on the NOAA 18 platform covering the period from 2005 to 2010.
- AVHRR-17: Data from the AVHRR sensor on the NOAA 17 platform covering the period from 2002 to 2010.
- AVHRR-16: Data from the AVHRR sensor on the NOAA 16 platform covering the period from 2001 to 2006.
- AVHRR-15: Data from the AVHRR sensor on the NOAA 15 platform covering the period from 1999 to 2003.
- AVHRR-14: Data from the AVHRR sensor on the NOAA 14 platform covering the period from 1995 to 2001.
- AVHRR-12: Data from the AVHRR sensor on the NOAA 12 platform covering the period from 1991 to 1999.

All validation results were obtained from the MMS by ingesting SST\_CCI L2P products and extracting the full set of L2P fields for each match-up record. The data for each sensor was then separated and clear-sky only match-ups were output as a single MD per sensor for subsequent evaluation. The MD also included a set of depth/time adjustments as follows:

- DT1: An adjustment ranging from 0.0 to 2.5 K to adjust the in situ measurement to an equivalent SST<sub>skin</sub> measurement at the in situ measurement time, derived from the FKC diurnal/skin model described in Section 4.1.5.
- DT2: An adjustment ranging from 0.0 to 2.5 K to adjust the in situ measurement to an equivalent SST<sub>0.2</sub> measurement at the standardised satellite measurement time, derived from the FKC model described in Section 4.1.5.
- DT4: ranging from 0.0 to 2.0 K An adjustment to adjust the in situ measurement to an equivalent SST<sub>0.2</sub> measurement at the standardised satellite measurement time derived from the time history of in situ measurements stored in the MMS; these adjustments are only available for a subset of drifting buoy match-ups and are used as an independent check of the DT2 DT3 adjustment.

A detailed validation of each individual sensor was done and these results are presented in APPENDIX A. Here we consider the entire SST\_CCI AVHRR dataset as a whole.

A time series of all SST\_CCI AVHRR datasets compared to the pseudo-independent drifting buoy dataset as well as the independent Argo dataset (for the period it is available) is shown in Figure 4-12.



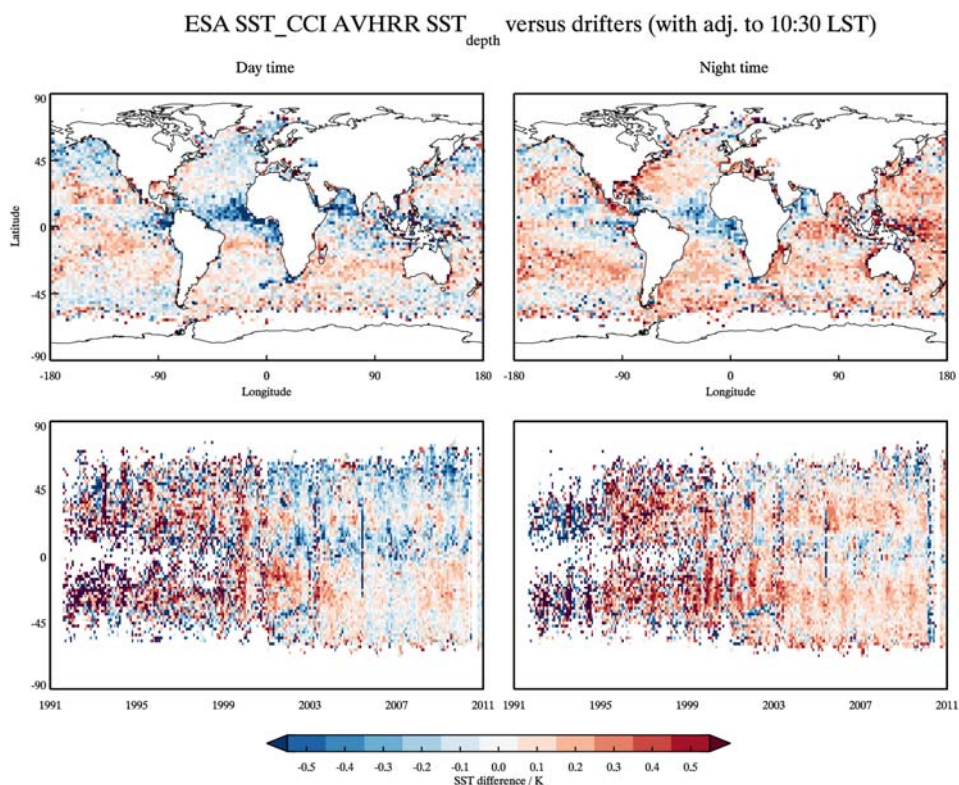
**Figure 4-12:** (Left) Time series of (lower) median discrepancy and (upper) robust standard deviation (RSD) for the SST\_CCI AVHRR mission compared to drifting buoys. Results are shown for daytime (red) and night time (blue) match-ups. Also, shown (right) is the equivalent time series for SST\_CCI AVHRR compared to Argo.

The time series in Figure 4-12 has two notable features. First, there are three clear regimes, 1991 to 1995 (NOAA-12), 1995 to 2002 (NOAA-14, NOAA-15 & NOAA-16) and 2004 to 2010 (adding NOAA-17, NOAA-18 and METOP-A). These three periods do correspond to different AVHRR sensors and so the quality of data is clearly linked to the individual sensor performance and characterisation. However, the final regime does not seem to correspond exactly to instrument changes and so may also be driven by changes in the drifter network or the auxiliary data used in the retrieval.

The data in the first regime are relatively noisy (higher RSD) and have periodic fluctuations in relative bias (high range of monthly median differences). This is a period of low numbers of drifter measurements, so part of what is observed is likely to come from non-uniform match-up distributions. However, the dominant origin is likely the unstable calibration of the AVHRR sensor on NOAA-12 during the period 1991 to 1994 (see Section A.7). The second regime has a clear notable annual cycle in the day time results. The third period is much more stable (no clear annual cycle). There is a clear day/night difference, with night time data showing a warm relative bias for the latter two regimes. The comparisons to Argo provide similar results although there is arguably better agreement between day time and night time data (however the results are much noisier).

The spatial distribution of the discrepancies for the SST\_CCI AVHRR mission is shown in Figure 4-13, which includes the latitude/longitude variation and time/latitude variation for both daytime and nighttime.





**Figure 4-13:** (Upper) Latitude/longitude variation of the median discrepancy for the AVHRR mission compared to drifting buoys for (left) daytime and (right) nighttime and (Lower) time/latitude variation of the same statistical measure. The drifting buoy measurement has been adjusted to the satellite measurement time using the FKC model as described in the main text. Each cell has at least 30 match-ups.

In Figure 4-13 we again see the day/night bias with the night time being warmer than the daytime for the later missions. Also, we see a notable increase in the noise prior to 2000 at the resolution of the plot as a result of the lower number of drifters. Indeed, prior to 1999 there are notable gaps in the drifter coverage, particularly in the tropics. A further feature in Figure 4-13 are the cold biases observed in the regions known to be susceptible to tropospheric dust aerosol such as the eastern Atlantic and western Indian oceans.

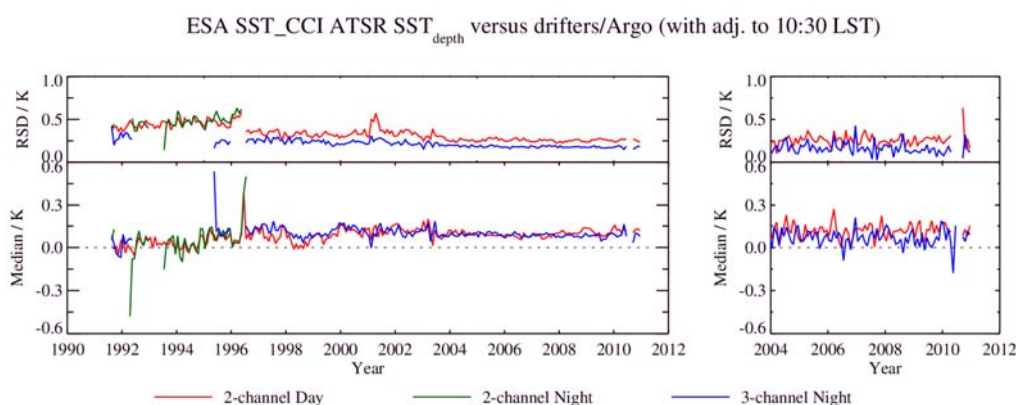
#### 4.4 Validation of ATSR Products using the MMS

The second set of SST\_CCI data products validated were the L3U products from the ATSR series. Three different datasets were included:

- AATSR: Data from the AATSR sensor on the ENVISAT platform covering the period from 2002 to 2010.
- ATSR-2: Data from the ATSR sensor on the ERS-2 platform covering the period from 1995 to 2003.
- ATSR-1: Data from the ATSR sensor on the ERS-1 platform covering the period from 1991 to 1997.

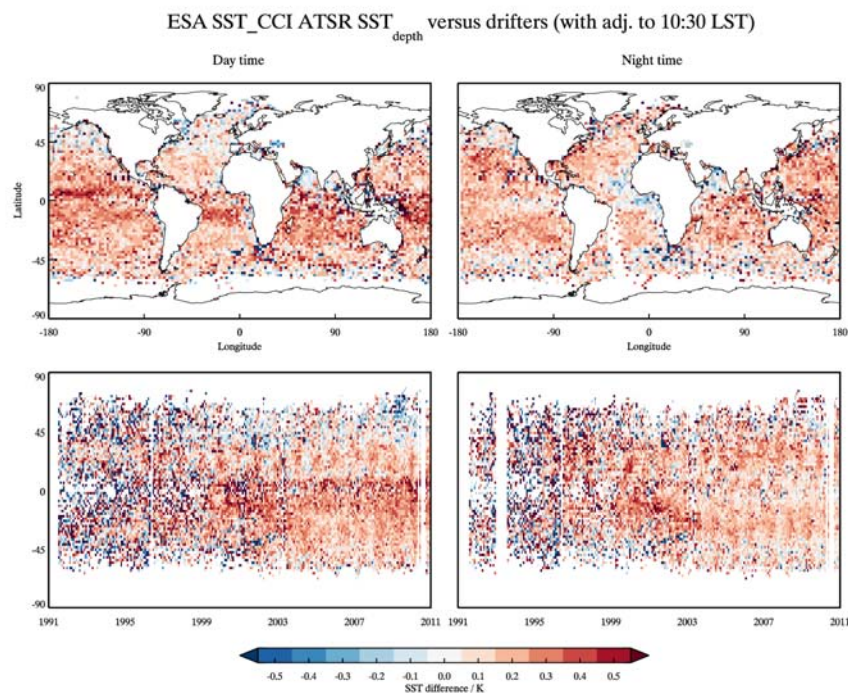
The validation procedures were as for the AVHRR validation in that all validation results were obtained from the MMS by ingesting SST\_CCI L3U products and extracting the full set of L3U fields for each match-up record. The data for each sensor was then separated and clear-sky only match-ups were output as a single MD per sensor for subsequent evaluation. The MD also included a set of diurnal/skin adjustments as described in Section 4.1.5. A detailed validation of each individual sensor was done and these results are presented in APPENDIX B. Here we consider the entire SST\_CCI ATSR dataset as a whole.

A time series of all SST\_CCI ATSR datasets compared to the pseudo-independent drifting buoy dataset as well as the independent Argo dataset (for the period it is available) is shown in Figure 4-14.



**Figure 4-14:** (Left) Time series of (lower) median discrepancy and (upper) robust standard deviation (RSD) for the SST\_CCI ATSR mission compared to drifting buoys. Results are shown for daytime (red) and night time (blue) match-ups. Also, shown (right) is the equivalent time series for SST\_CCI AVHRR compared to Argo.

The time series in Figure 4-14 has three regimes as seen for AVHRR with the first two regimes being identified by the changeover between ATSR-1 and ATSR-2 but the second and third regime being identified by the an unknown occurrence towards the end of 2003. Globally there is good agreement between day time and night time relative biases for the ATSRs for the drifter match-ups, with a clear difference in the robust standard deviation between day time and night time: this is expected due to the increased retrieval noise of a 2-channel retrieval compared to a 3-channel retrieval, as well as to differences in the cloud masking between day and night (as a result of different spectral channels being used). The Argo comparisons, like for AVHRR, show a difference to the drifters in that the night time bias may be slightly cooler than that observed for drifter comparisons. The spatial distribution of the discrepancies for the SST\_CCI AVHRR mission is shown in Figure 4-15, which includes the latitude/longitude variation and time/latitude variation for both daytime and nighttime.



**Figure 4-15:** (Upper) Latitude/longitude variation of the median discrepancy for the ATSR mission compared to drifting buoys for (left) daytime and (right) nighttime, and (lower) time/latitude variation of the same statistical measure. The drifting buoy measurement has been adjusted to the satellite measurement time using the FKC model as described in the main text.

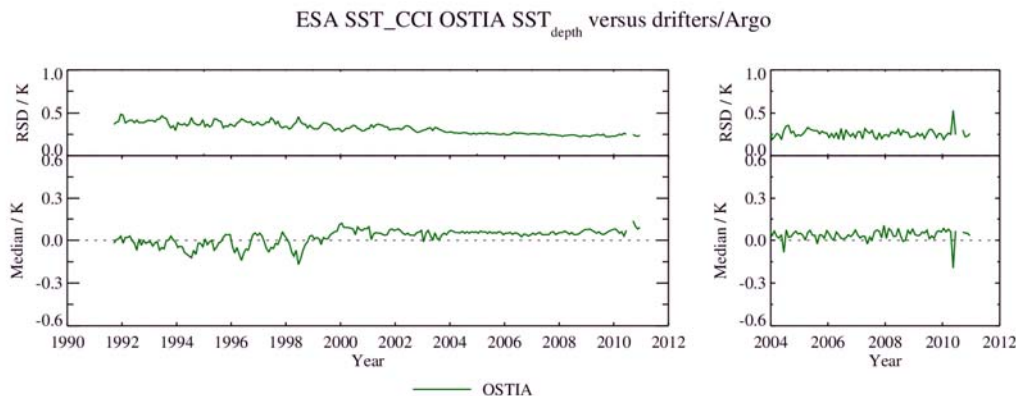
The maps in Figure 4-15 show a general warm bias and confirm the agreement between day and night results seen in the time series analysis. As for AVHRR there is evidence of residual effects of tropospheric mineral dust in the Atlantic and NW Indian oceans. Other notable features are warm biases in regions of predominantly cirrus clouds in the daytime, and a cool day time bias in high northern latitudes, particularly in the Pacific. A further observation is the apparently better “global coverage” of drifting buoy match-ups in the early 90’s compared to AVHRR, the reason for which is unclear and may be due to the method of building the MMS from pre-existing MDs for SST\_CCI Phase I. (We cannot guarantee the same procedures were used to generate the MD files used in Phase I for ATSR and for AVHRR so in SST\_CCI Phase II the MMS will be built from individual in situ datasets and not from pre-existing MDs so there will be better consistency between sensors as there will be derived using the same methodology.)

## 4.5 Validation of analysis products using the MMS

The final set of SST\_CCI data products validated were those from the SST\_CCI analysis covering the period from 1991 to 2010. The SST\_CCI analysis products have been validated twice, once using the MMS (as described in this section) and using an independent match-up processing system specifically for the analysis products as described in Section 4.6.

The validation procedure for the MMS analysis was as for the AVHRR and ATSR validation in that all validation results were obtained from the MMS by ingesting SST\_CCI analysis products and extracting the full set of analysis fields for each match-up record; all match-up records were used including those that are flagged as cloudy in the AVHRR and ATSR products. All data were then output as a single MD file for subsequent evaluation. The OSTIA MD also included a set of diurnal/skin adjustments as described in Section 4.1.5 in an attempt to minimise the depth and time difference contributions to the uncertainty budget. However, it soon became clear that the analysis output was not always a “daily average” and was indeed dominated by day time sampling from the AVHRR and ATSR data (due to differences between the day time and night time cloud masking). Consequently, it was not possible to set a reference time for the analysis outputs for an adjustment to be calculated. Here, we consider equivalent comparisons to those presented in Section 4.3 and Section 4.4. Additional analysis validation results using the MMS are provided in APPENDIX C.

A time series of the SST\_CCI analysis dataset compared to the pseudo-independent drifting buoy dataset as well as the independent Argo dataset (for the period it is available) is shown in Figure 4-16.

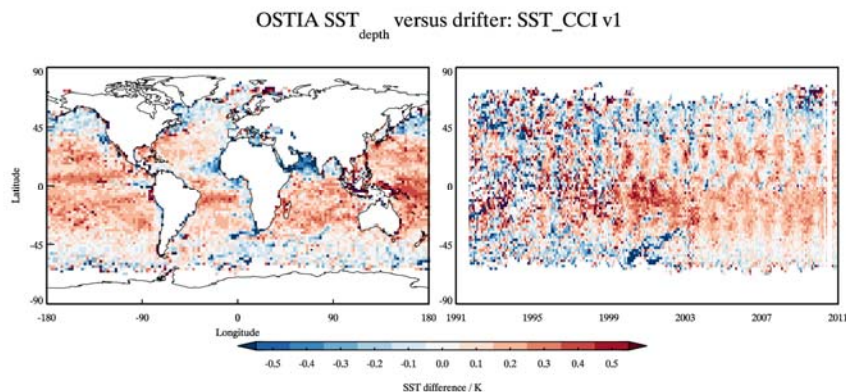


**Figure 4-16:** (Left) Time series of (lower) median discrepancy and (upper) robust standard deviation (RSD) for SST\_CCI analysis compared to drifting buoys. Also, shown (right) is the equivalent time series for SST\_CCI analysis compared to Argo.

The time series in Figure 4-16 is a combination of the AVHRR and ATSR datasets and does not exactly show the same three regimes observed for the satellite datasets although the highly stable system from 2004 onwards is clearly present. Prior to 2004 the notable cycling originating from the AVHRR record is seen. However, there is no evidence of the highly noisy comparisons for the early AVHRR mission (see Figure 4-12) due to the bias correction to ATSR in the analysis system.



The spatial distribution of the discrepancies for the SST\_CCI analysis is shown in Figure 4-17, which includes the latitude/longitude variation and time/latitude variation for both daytime and nighttime.



**Figure 4-17:** (Left) Latitude/longitude variation of the median discrepancy for the SST\_CCI analysis compared to drifting buoys and (right) the time/latitude variation of the same statistical measure.

The map in Figure 4-17 confirms the generally warm bias globally expected given the input data streams. Suspected tropospheric aerosol effects causing negative relative bias in the NW Indian Ocean are stronger than in either input dataset. The time latitude plot shows the high variability in the 1999 to 2004 period is mainly in the tropics and particularly the Southern Hemisphere. Similar features were seen in both the ATSR and AVHRR equivalent plots, which may indicate a common origin or error in the reference dataset. Despite there being a minimal annual cycle in the time-series median (Figure 4-16) after 2004, in the time/latitude plot it is clear that this arises from opposite annual cycles of relative bias in the northern and southern hemispheres during this period.

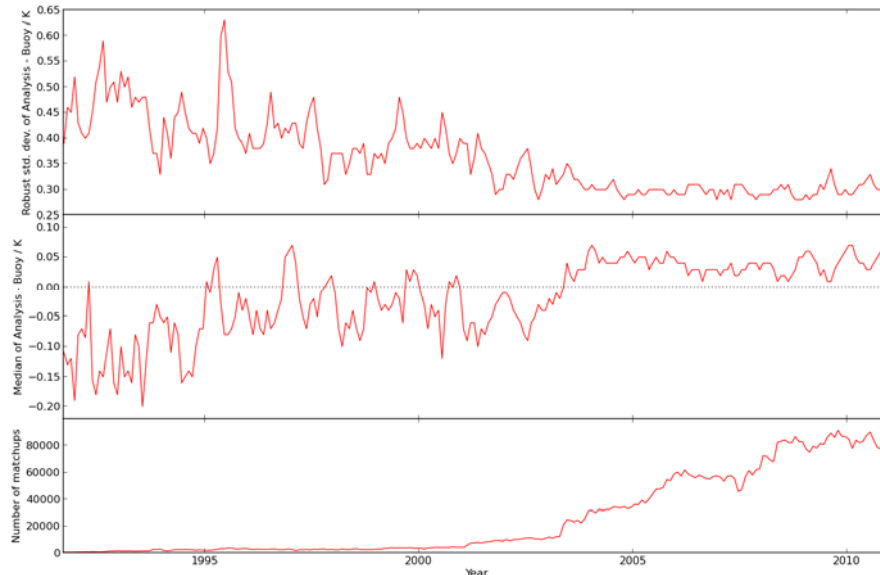
Two further notable features in Figure 4-17 are the cold biases below 45 °S between 2000 and 2001 and the cold biases above 50 °N in 2009 and 2010. These features are both visible in the input AVHRR and ATSR data but have lower magnitudes. The reason for the apparent amplification in the analysis is unclear but may be due to the negatively biased results being persisted in the analysis more often than positively biased results for the extra cloudy matches used in the validation.

## 4.6 Validation of SST\_CCI analysis using re-matching

Validation of SST\_CCI analysis uncertainty estimates was conducted using drifting buoy observations from ICOADS that did not fail the quality control procedures of Rayner et al. (2006; RD.72) and Atkinson et al (2013; RD.326). Drifting buoy observations were chosen as they cover the full length of the ESA SST\_CCI long term product and their nominal depth of measurement matches the depth of the SST\_CCI data. Although the level 4 analyses approximate the daily average SST, they are formed from SSTs adjusted to correspond to 10.30 am or pm local time. Therefore, buoy observations made within 30 minutes of these times were selected. For each match-up the difference between the analysis and the drifting buoy was analysed.

### 4.6.1 The long term product

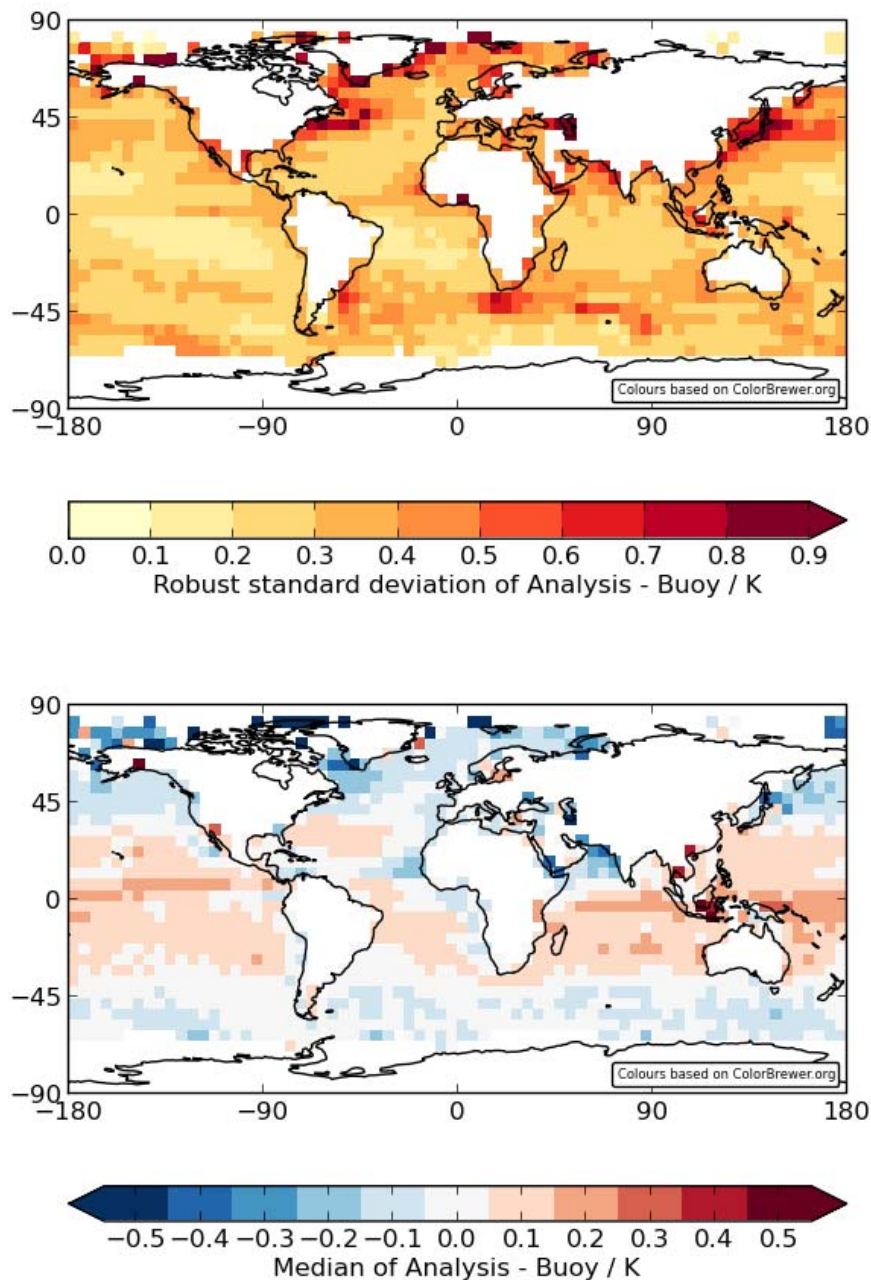
Figure 4-18: shows the median and robust standard deviation of the difference between buoys and analyses aggregated for each month of the long term product. In the top panel, a decrease over time in the standard deviations is seen. A clear change (of the order 0.2 K) is observed in the mean difference between the analysis and buoys (bottom). However, both these results could be associated with changing spatial sampling from the buoy sampling (see below for examples). In the most recent years of the SST\_CCI data the robust standard deviation is typically approximately 0.3 K. From 2004 – 2005 the median differences are fairly constant. There is evidence of a step change in the mean differences from about 0.05 K to 0.03 K in 2006. After 2008 there appears to be a seasonal cycle in the mean differences which was not evident in the preceding years.



**Figure 4-18:** Median and robust standard deviation of the difference between the analyses and buoy data for each month of the long term product.

The spatial distributions of the median differences and robust standard deviations are shown in Figure 4-19. As might be expected, the standard deviations are larger in more variable regions of the ocean, because of mismatch-up effects. There is a very clear latitude dependent distribution to the median differences, which is generally positive at low latitudes and negative at high latitudes. Exceptions are the Arabian Sea and the

eastern low-latitude Atlantic, which both exhibit negative differences. In the earlier years the spatial coverage of the buoys is poor and low latitudes are very weakly sampled (Figure 4-20). This changing coverage is likely – at least in part – to be the cause of the variation over time of the mean differences shown in Figure 4-18.



**Figure 4-19:** Median and robust standard deviation of the difference between the analyses and buoy data in 5° grid cells calculated using data from the full period of the long term product.

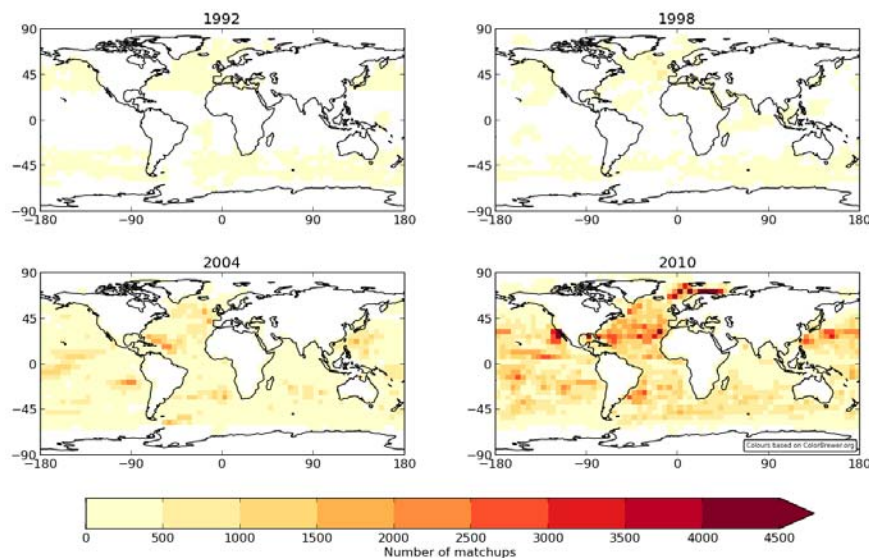
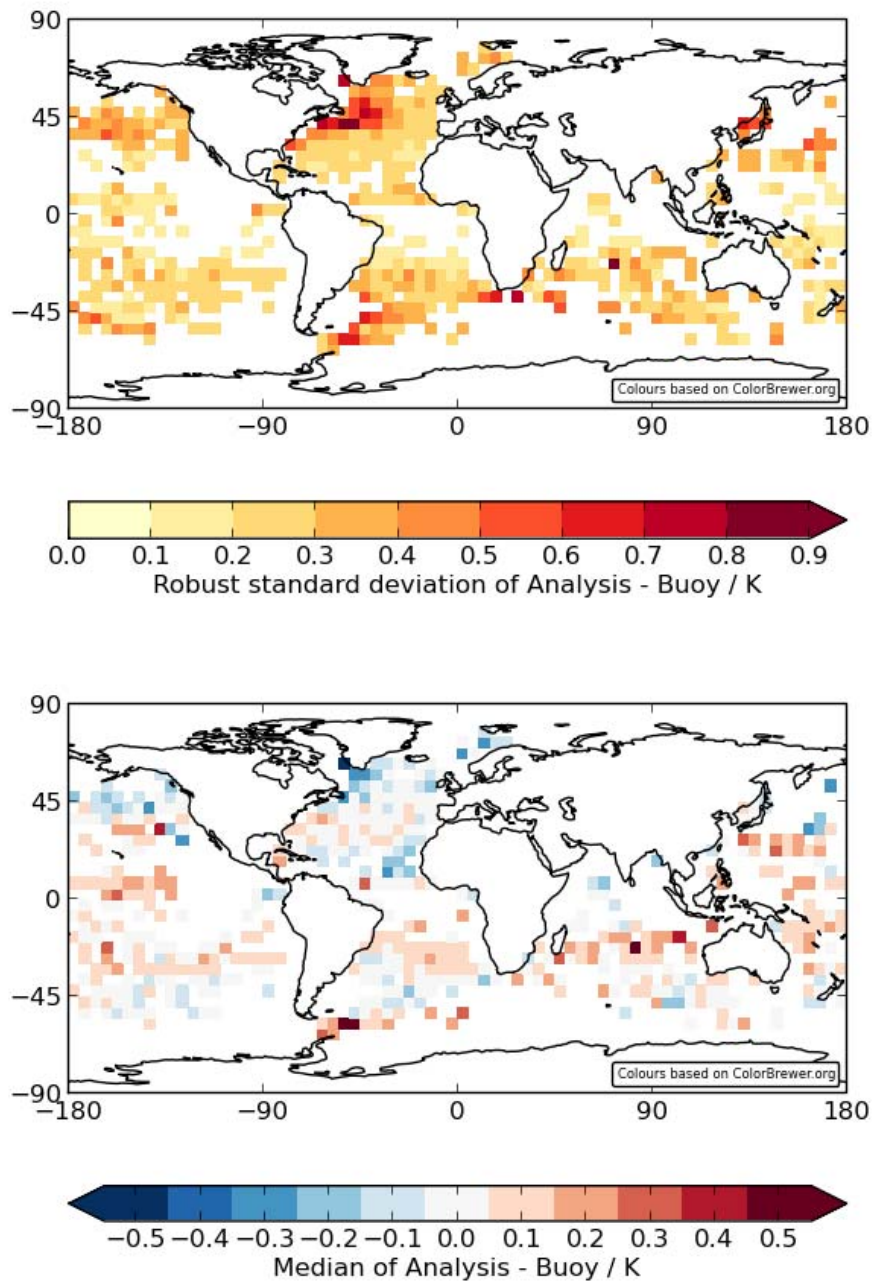


Figure 4-20: Number of matchups between analyses and buoys during 1992, 1998, 2004 and 2010.

#### 4.6.2 The demonstration product

The spatial distribution of the robust standard deviations and median differences for the first period of the demonstration product (June – August 2007) is shown in Figure 4-21. Although the low number of matchups has led to many gaps in the plots, it is possible to see some similar features to those seen in the equivalent plots in the long term product. Aggregated across the globe and all three months, the robust standard deviation is marginally lower than for the long term product (0.30 K compared to 0.31 K for the long term product over the same three months). Over all the matchups the median difference is 0.04 K compared to 0.02 K for the long term product during the same period.





**Figure 4-21:** As Figure 4-19 but showing results for the first period of the demonstration product (June – August 2007).

## 4.7 High-latitude validation using the MMS

Several of the traditional global SST algorithms have elevated errors in the high latitudes, compared to the low- and mid-latitudes (Hoeyer, et al, 2012 [RD.340]; Poulter and Eastwood, 2008 [RD.335]) with significant seasonal variations in the performance. This section therefore is aimed at validating the CCI SST products for the high latitudes. The Arctic Ocean is in this context understood as all observations at latitudes 60 °N and Northwards. The Southern Ocean is defined as the region southwards from 50 °S.

The match-ups were selected as the satellite versus in situ observation that was within 2 hours and 15 km from each other. This check was introduced to discard a few erroneous match-up pairs very far apart. Only match-up pairs labeled “good” (“matchup.valid” = 0), with satellite quality flags (“sat.quality\_level”) of 5 have been used and a gross error check has been performed to discard all match-ups pair where in situ or satellite observations were lower than -3 °C or higher than 36 °C. Match-ups from ice filled waters have been discarded, except in section 4.7.3.4 where the distance to ice error statistics is considered.

All results reported here use drifting buoy match-ups available in the categories: “testing”, “selection” and “validation” to maximize the number of match-ups and at the same time provide independency of the algorithm development.

The satellite SST field use throughout this section is the 20 cm SST estimate (“sat.sea\_surface\_temperature\_depth”), which is at the same depth as the drifting buoy observations.

### 4.7.1 Spatial and temporal coverage of match-ups

The spatial and temporal coverage of the match-ups are shown in the figures below. In the Arctic, most of the match-ups are located in the Nordic Seas and the Barents Sea, whereas the Southern Ocean match-ups are more uniformly distributed for all longitudes. Both high latitude regions have very few match-ups in the areas with seasonal ice cover.

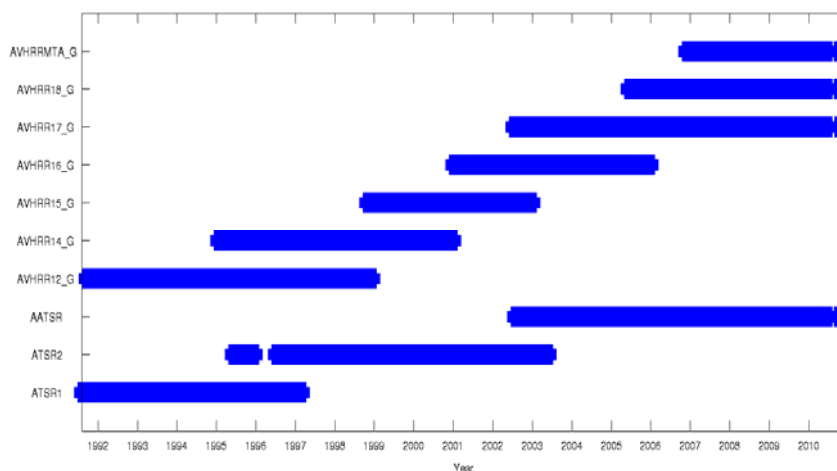


Figure 4-22: Timeline of satellite vs. in situ match-ups for the various satellite products

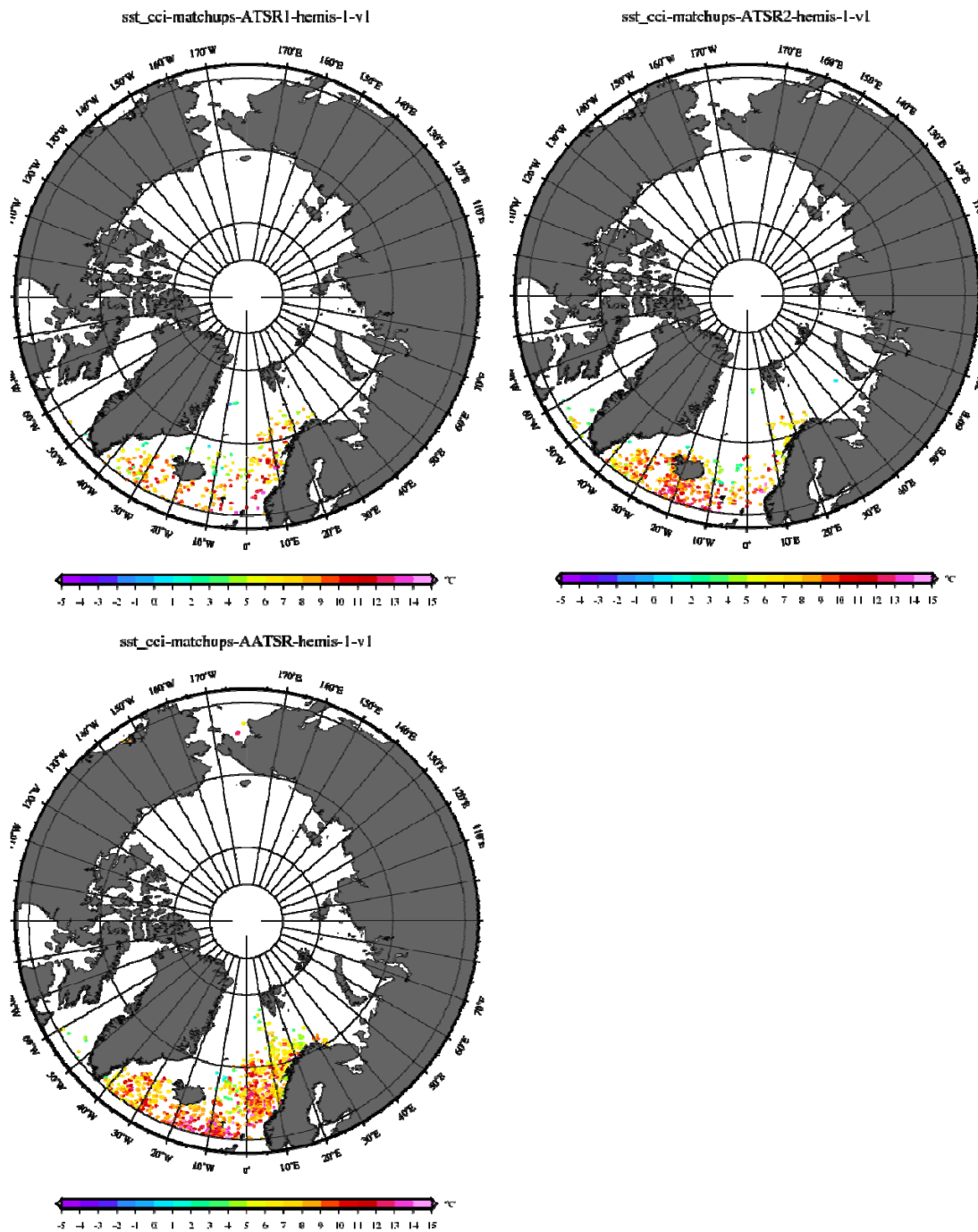
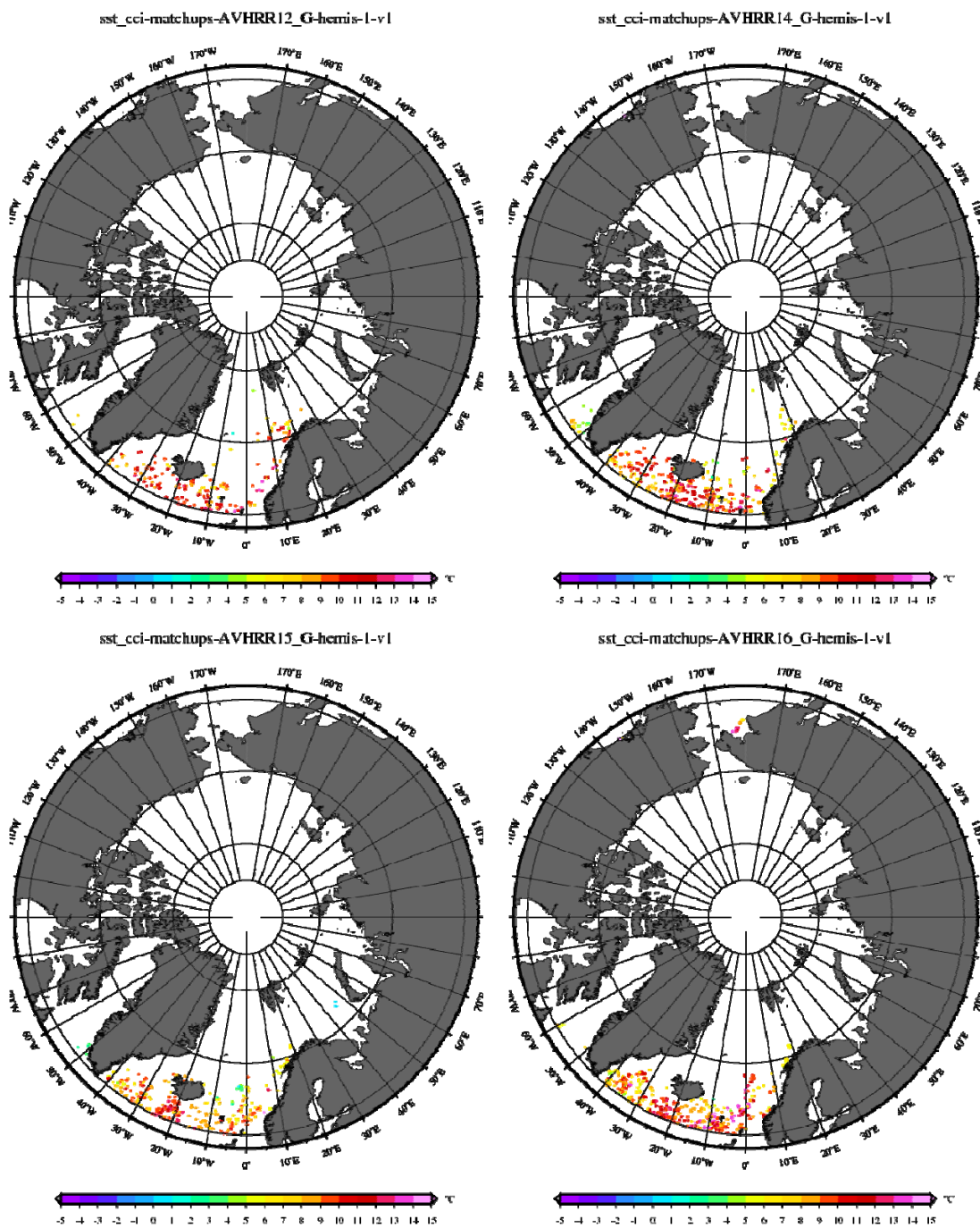


Figure 4-23: Spatial coverage of the Arctic satellite – in situ match-ups for the individual ATSR L3U products. The colors indicate the in situ SSTs.





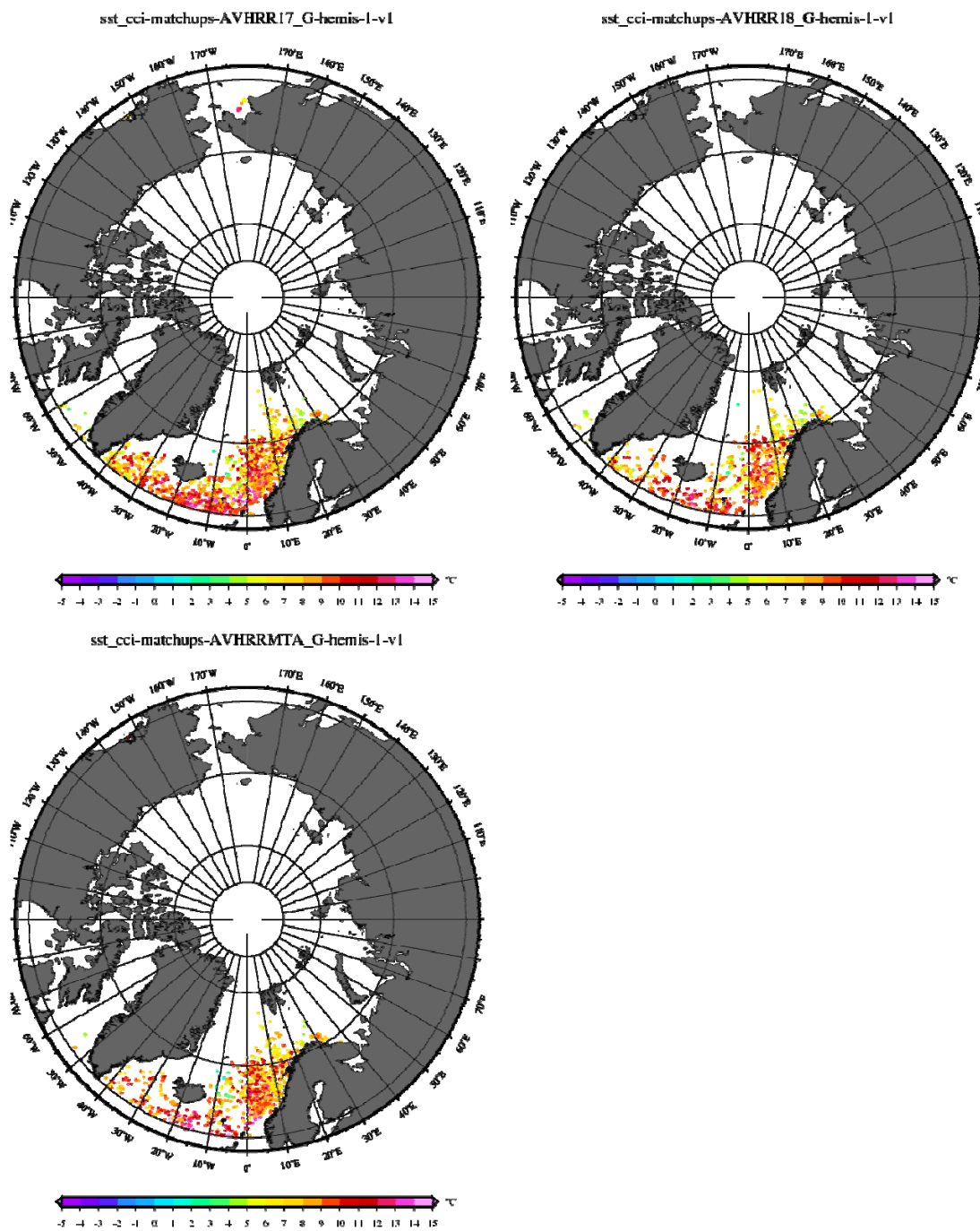


Figure 4-24: Spatial coverage of the Arctic satellite – in situ match-ups for the individual AVHRR L2P products. The colors indicate the in situ SSTs.

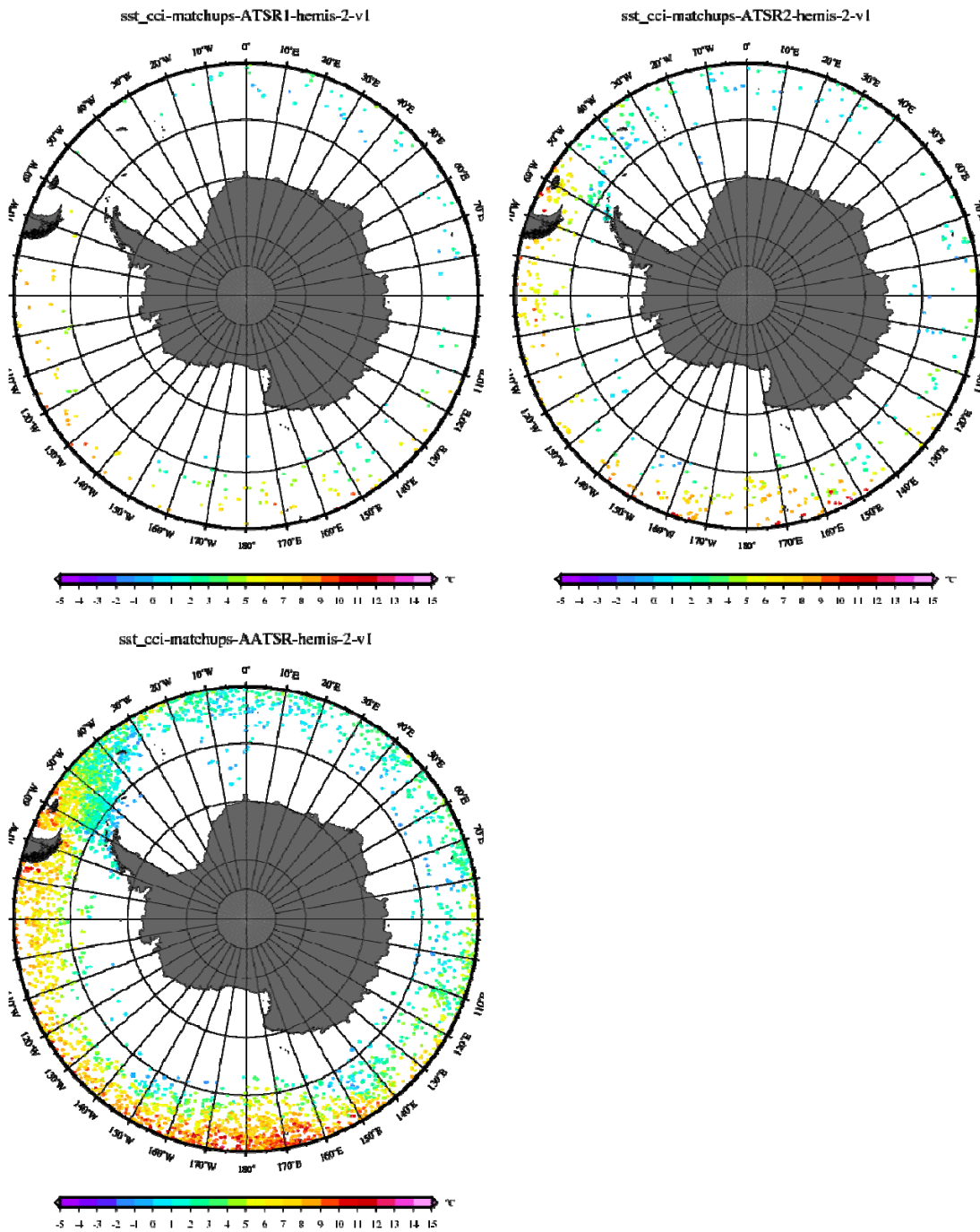
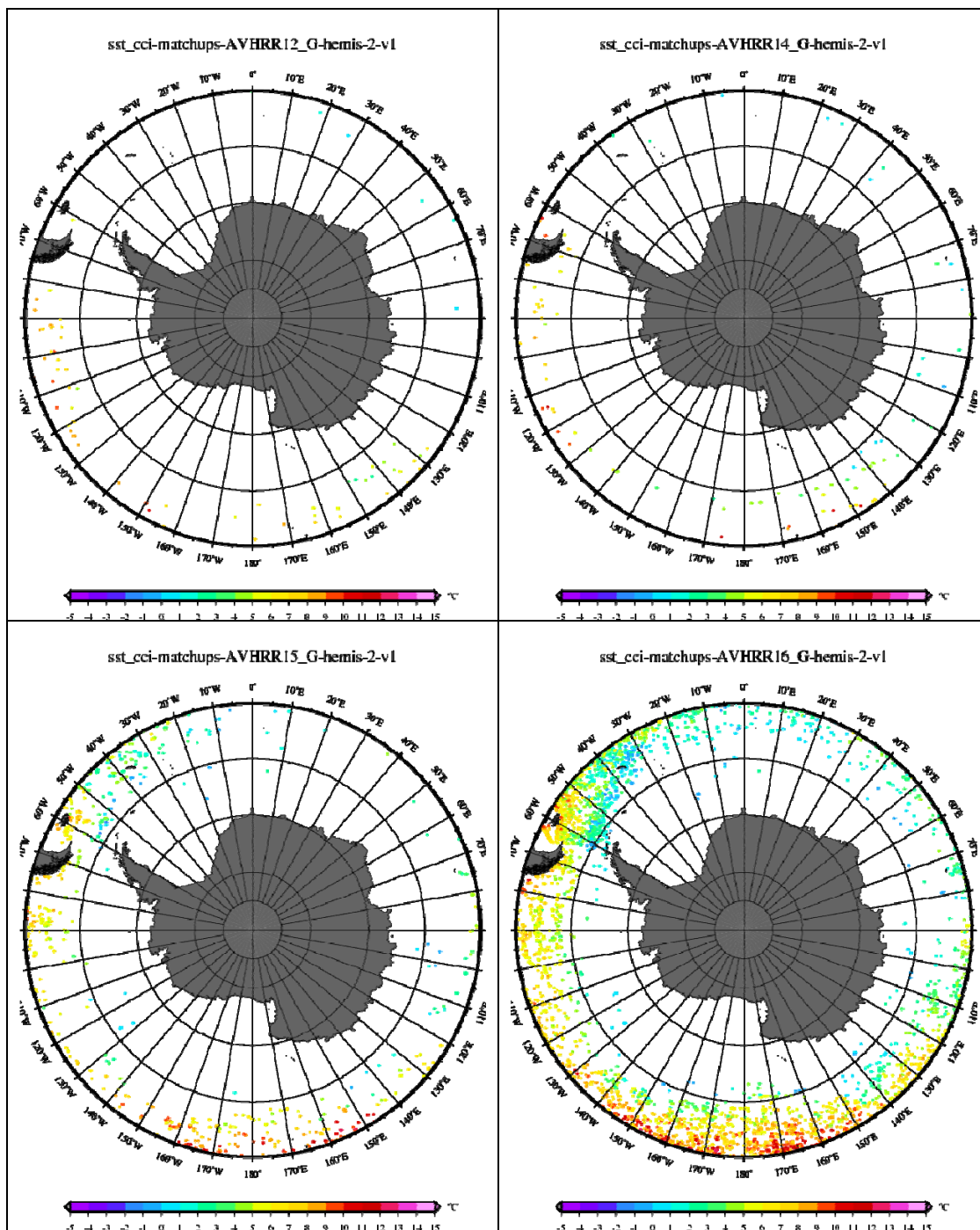


Figure 4-25: Spatial coverage of the Southern Ocean satellite – in situ match-ups for the individual ATSR L3U products. The colors indicate the in situ SSTs.





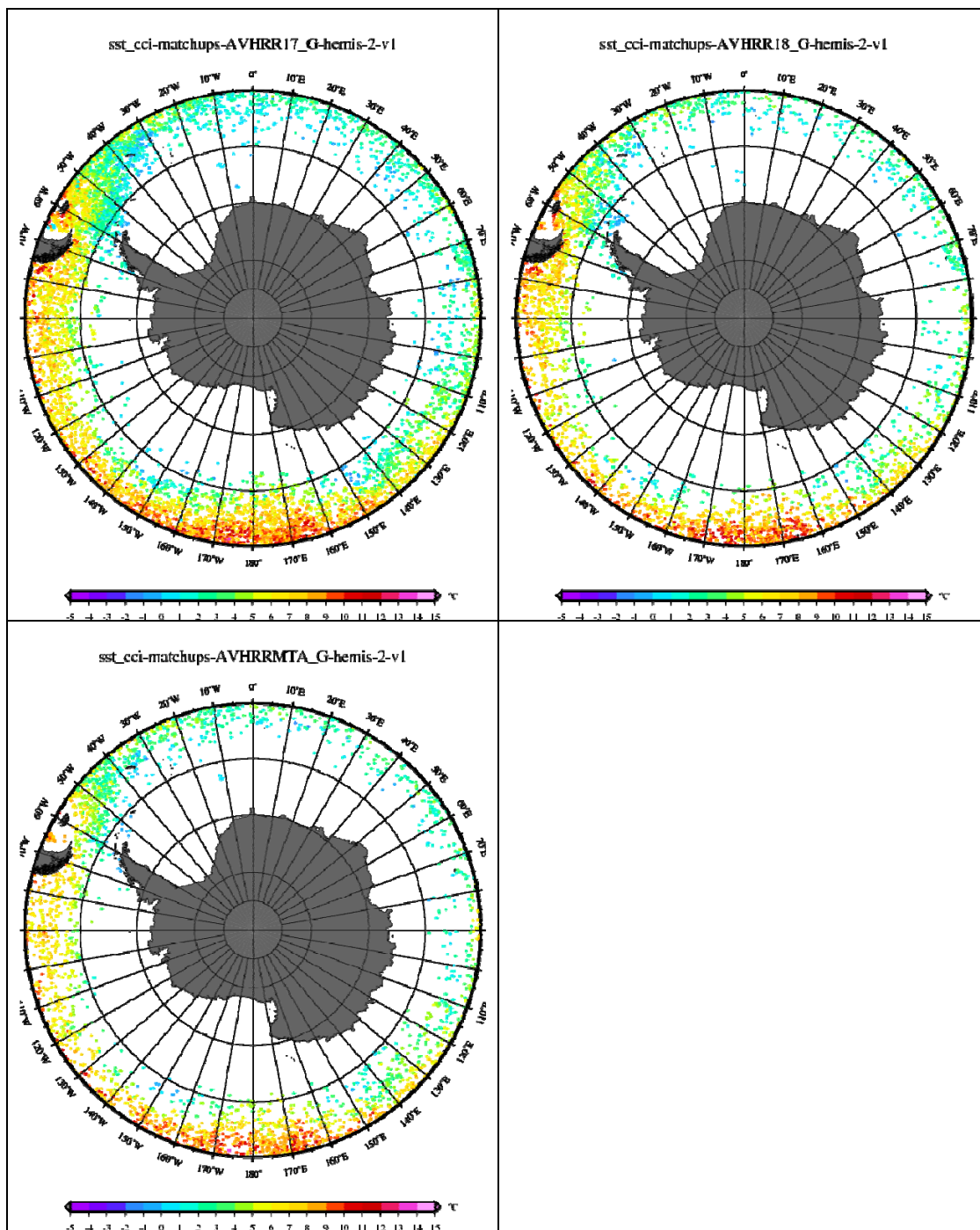
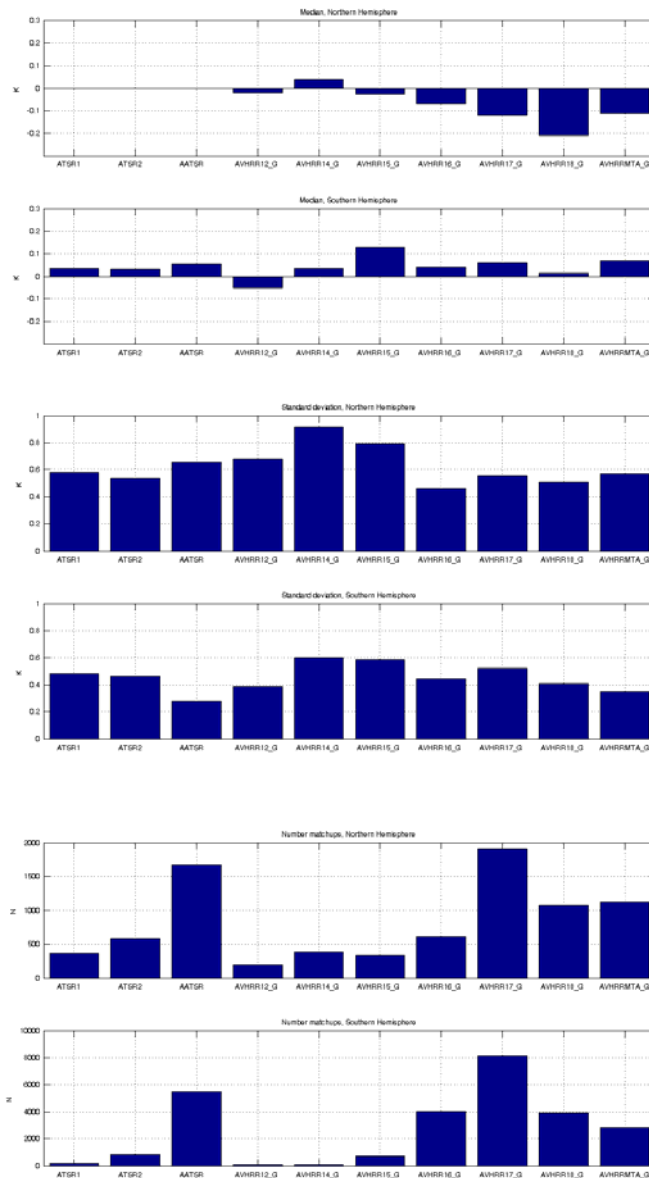


Figure 4-26: Spatial coverage of the Southern Ocean satellite – in situ match-ups for the individual AVHRR L2P products. The colors indicate the in situ SSTs.

#### 4.7.2 General error numbers

The overall statistics for each of the products are shown in figure 4-27 for the Arctic and the Southern Ocean. The numbers are median and standard deviation.

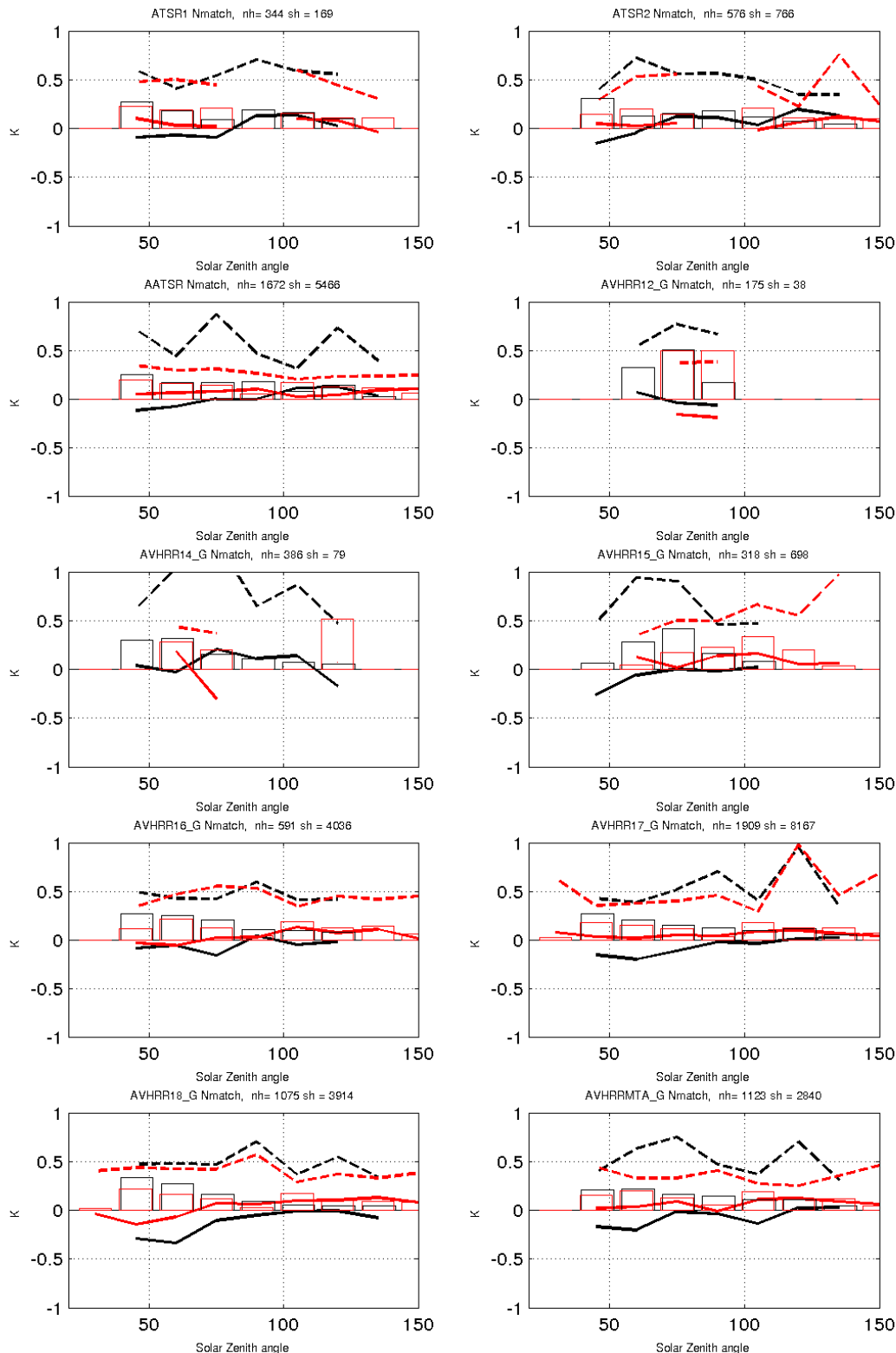


**Figure 4-27:** Median (top), standard deviation (middle) and number of match-ups (lower) for the satellite – in situ match-ups for all satellite products for Arctic Ocean and Southern Ocean.

### 4.7.3 Detailed validation

Very large seasonal variations are seen in the atmospheric and oceanographic conditions for the high latitudes. Infrared satellite observations are affected by the atmospheric and oceanographic state and a detailed validation is therefore carried out to assess the performance of the satellite products as a function of: solar zenith angle, total column water vapor, Julian day during the year and as a distance to ice edge. The results are shown in figures 4-28 to 4-31. All numbers are based upon at least 15 match-ups before a calculation was performed.

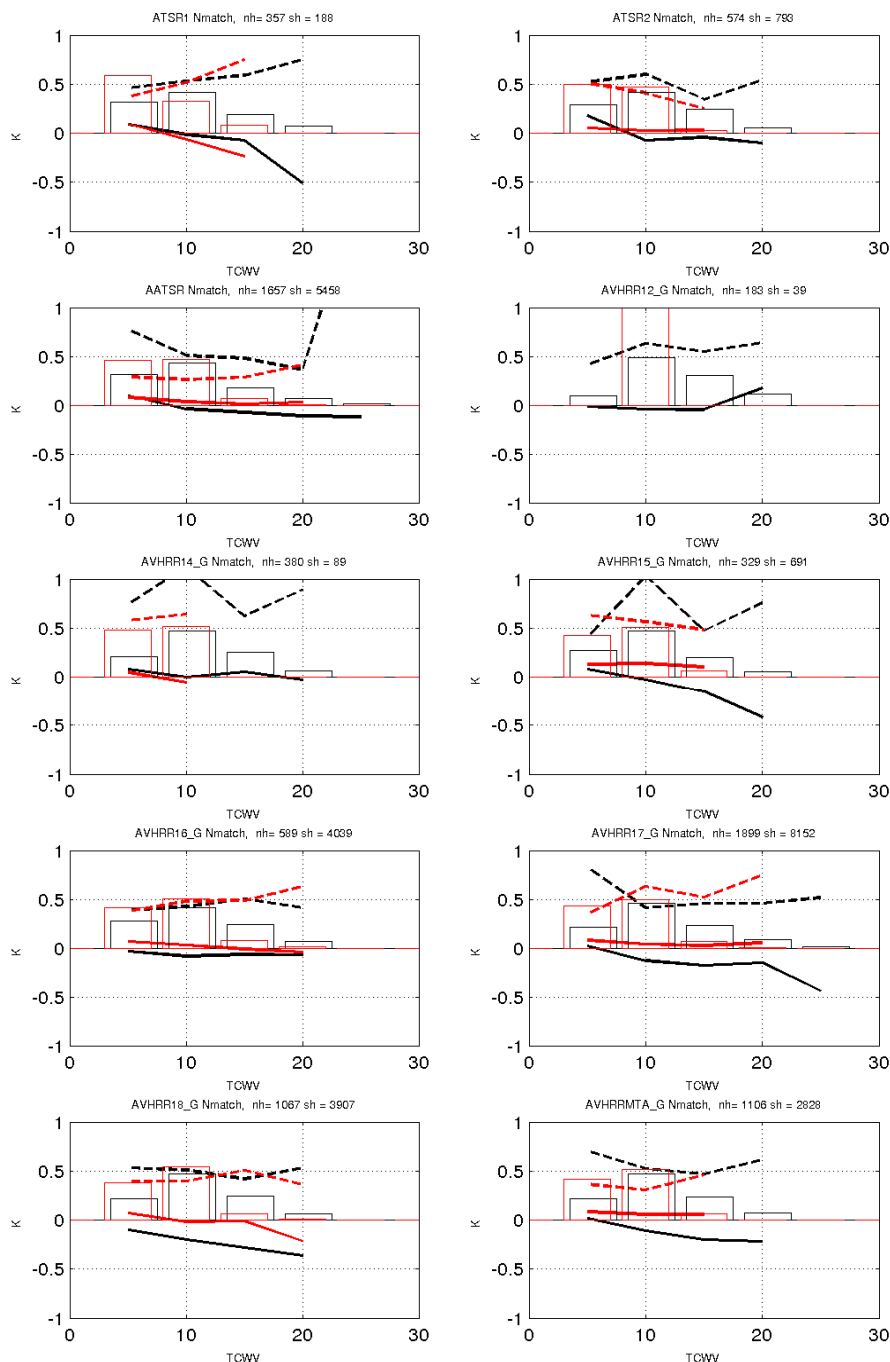
### 4.7.3.1 Solar zenith angle



**Figure 4-28:** Satellite performance with respect to solar zenith angle. Black lines indicate Arctic Ocean and the red lines show results from the Southern Ocean. The thick solid lines show median and dashed lines show the standard deviation. The bars with thin lines

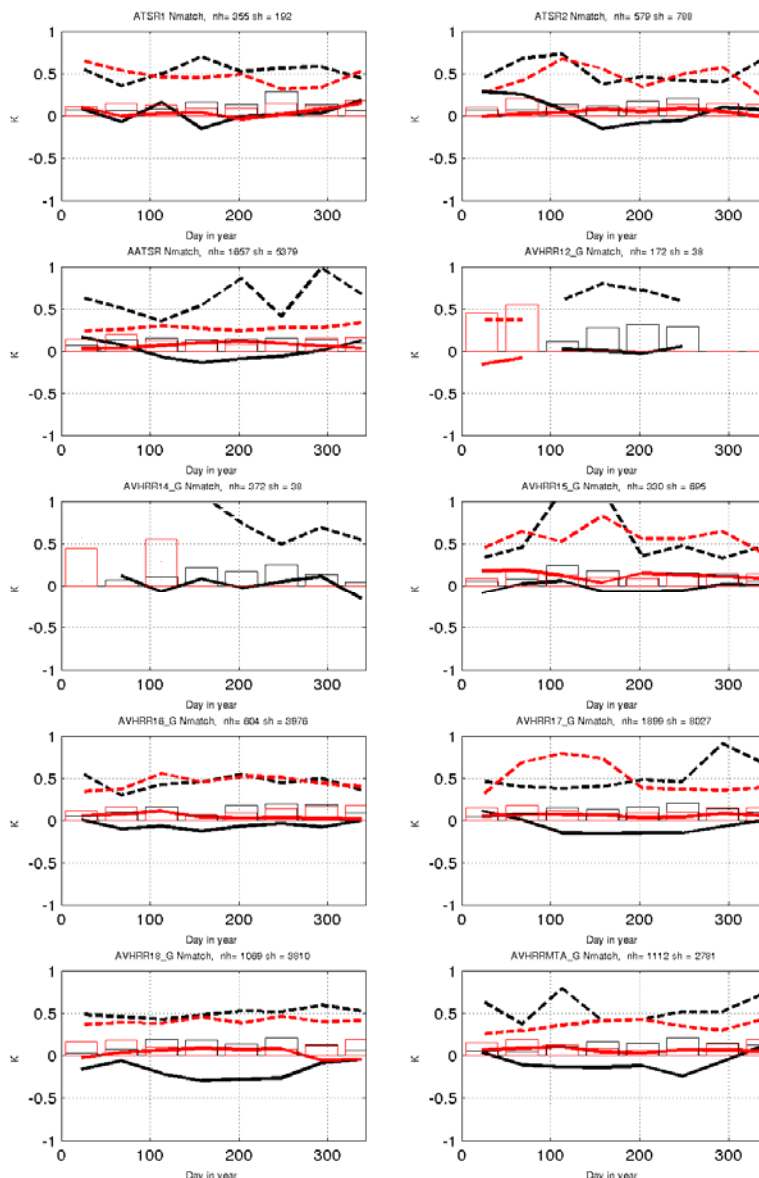
show the normalized number of matches. The total number of match-ups for a given hemisphere is given in the title for each figure.

### 4.7.3.2 Total column water vapour



**Figure 4-29:** Satellite performance with respect to the total water vapor content ( $\text{kg/m}^2$ ). Black lines indicate Arctic Ocean and the red lines show results from the Southern Ocean. The thick solid lines show the median and dashed lines show the standard deviation. The bars with thin lines show the normalized number of matches. The total number of match-ups for a given hemisphere is given in the title for each figure.

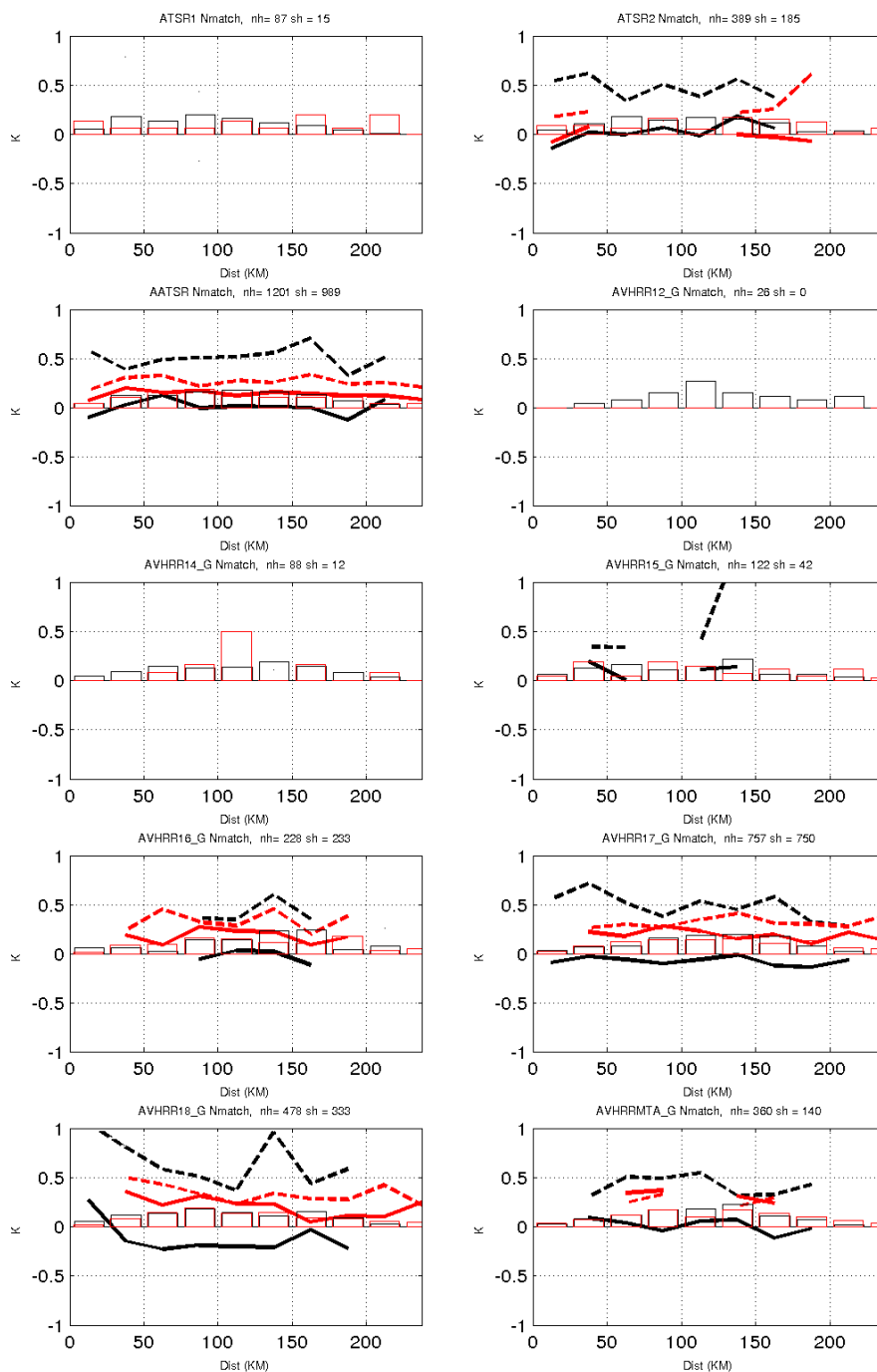
### 4.7.3.3 Day in year



**Figure 4-30:** Satellite performance with respect to the day in the year. Black lines indicate Arctic Ocean and the red lines show results from the Southern Ocean. The thick solid lines show the median and dashed lines show the standard deviation. The bars with thin lines show the normalized number of matches. The total number of match-ups for a given hemisphere is given in the title for each figure.

### 4.7.3.4 Distance to Ice

The available in situ and satellite observations in the vicinity of the ice edge is low compare to other regions of the high latitudes. All categories of match-ups (“training”, “testing”, “validation” “selection” ) have thus been used to product the results in figure 4-31 to increase the number of match-ups in this region.



**Figure 4-31:** Satellite performance with respect to the distance from the sea ice (defined here as sea ice concentration > 15%). Black lines indicate Arctic Ocean and the red lines show results from the Southern Ocean. The thick solid lines show the median and dashed lines show the standard deviation. The bars with thin lines show the normalized number of matches. The total number of match-ups for a given hemisphere is given in the title for each figure. A minimum number of 20 match-ups is required to calculate statistics.

#### 4.7.4 Summary

The summary of the high latitude validation is given in bullet form below

- Overall validation showed a better performance of all the products in the Southern Ocean, compared to the Arctic Ocean.
- ATSRs: Close to zero bias in Arctic, small positive in Southern Ocean
- AVHRRs: Significant negative biases for late AVHRRs (NOAA 16,17,18, METOP-A), largest for AVHRR 18\_G with a median of -0.21°C
- CCI AVHRR + AATSR showed seasonal bias variations in the Arctic with a cold summer bias.
- Cold Arctic biases are also found for humid atmospheric conditions and low solar zenith angles (daytime).
- No seasonal signals in the validation results can be identified in the Southern ocean and the dependence upon solar zenith angle and total water vapor is significantly smaller than in the Arctic Ocean.
- No significant trends are seen in the validation statistics close to the ice edge, however the limited number of match-up prevent a reliable conclusion for most of the products.

#### 4.8 Summary of validation results

The SST\_CCI long-term products have been validated against both independent and pseudo-independent reference data. The following conclusions are drawn from the evidence presented here for:

- SST\_CCI long-term L2P AVHRR
  - Regional biases of several 10ths are calculated
  - Day time data generally cooler than night
  - Strong degree of consistency between later sensors; earlier sensors markedly more variable
- SST\_CCI long-term L3U ATSR
  - Regional biases of few 10ths (generally somewhat larger than ARC, except for ATSR-1)
  - Day time data generally warmer than night
  - Discrepancy between day and night coverage larger than expected (indicates issue with day time cloud mask)
- SST\_CCI long-term L4 analysis
  - Skewed towards day time (c.f. day/night coverage in L2P/L3U)



- Biases comparable to those seen in L3U data (due to ATSR bias correction).

## 5. VALIDATION AND VERIFICATION OF SST\_CCI UNCERTAINTY ESTIMATES

### 5.1 Introduction

A key aim of the SST\_CCI project is to provide a pixel level standard uncertainty for all products. A further aim is to verify the uncertainties using independent measurements. In this section we attempt to provide an assessment of the product standard uncertainties using match-ups to drifting buoys. Here the drifting buoys are not totally independent as some match-ups were used in the algorithm selection process. However, the SST\_CCI products are not tied to drifting buoys in any way so we use all drifter match-ups from the MMS as a pseudo-independent dataset.

#### 5.1.1 Uncertainty validation for AVHRR and ATSR data using the MMS

The approach to uncertainty validation using the MMS is to compare the robust standard deviation (RSD) of the discrepancy between the SST\_CCI and drifter measurements as a function of the measurement uncertainty as provided in the SST\_CCI products. In an ideal case the standard deviation of the differences between the satellite SST and a reference SST would scale as a function of the satellite uncertainty, i.e.

$$\sigma_{sat-ref} = \sigma_{sat}$$

However, the reference data has its own uncertainties to consider, as discussed in Section 4.2. Consequently the standard deviation of the differences between the satellite SST and a reference SST are really a combination of both the uncertainty in the satellite SST and the uncertainty in the reference SST, i.e.

$$\sigma_{sat-ref} = \sqrt{\sigma_{sat}^2 + \sigma_{ref}^2}$$

As such, at low satellite uncertainties the standard deviation of the differences is dominated by the uncertainty in the reference data. As you move to higher satellite uncertainties the satellite uncertainty will then dominate as the reference uncertainty becomes a less significant contribution to the total uncertainty. In reality there are other terms to consider relating to:

- The difference in spatial sampling (a point reference measurement versus a satellite pixel);
- The difference in depth of the measurements;
- The difference in time of the measurements.

This approach also considers uncertainty due to environmental effects related to the homogeneity of a region/process. For example, validation in a region dominated by fronts at low wind speed the first term (spatial sampling) will be systematic for any one single match-up. However, as the number of match-ups increases the uncertainty will reduce by  $1/N^1$  as you sample the variability at multiple locations. Consequently, the effect is considered to be a pseudo-random term and not a systematic term. Likewise, in an area of strong solar radiation and low wind speed the second term (difference in depth) would

be systematic for any one match-up. Here, we attempt to reduce these latter three terms to  $\ll 0.1$  K in the mean through the use of a depth/time adjustment, large number of match-ups (to reduce pseudo-random terms) and through like versus like ( $SST_{\text{skin}}$  versus  $SST_{\text{skin}}$  or  $SST_{\text{depth}}$  versus  $SST_{\text{depth}}$ ) comparisons. For that reason, these terms are neglected in our uncertainty validation.

For the  $\sigma_{\text{ref}}$  term we have used:

- 0.20 K for drifters
- 0.10 K for GTMBA
- 0.01 K for Argo
- 0.10 K for ship-borne radiometers.

These values are currently the best estimate of the dataset uncertainty (as summarised in Table 4-1). None of the reference data comes with an uncertainty per measurement.

### 5.1.2 Uncertainty validation using re-matching

Validation of SST\_CCI analysis uncertainty estimates was conducted using drifting buoy observations from ICOADS that did not fail the quality control procedures of Rayner et al. (2006; RD.72) and Atkinson et al (2013; RD.326). Drifting buoy observations were chosen as they cover the full length of the ESA SST CCI long term product and their nominal depth of measurement matches the depth of the SST CCI data. Although the level 4 analyses approximate the daily average SST, they are formed from SSTs adjusted to correspond to 10.30 am or pm local time. Therefore, buoy observations made within 30 minutes of these times were selected.

For each drifting buoy observation, the closest analysis SST and the corresponding uncertainty estimate were extracted. An uncertainty in the drifting buoy value was added to the analysis uncertainty estimate. The difference between the buoy and analysis SSTs was divided by the combined uncertainty estimate to give a normalised quantity. The difference between a single buoy measurement and the analysis in that location is the combination of any systematic offset that exists and the random errors in the buoy measurement and analysis, which are spatially variable. By dividing by the uncertainty estimate the spatial variability in the uncertainty is removed, making it easier to calculate statistics for data aggregated globally. The standard deviation of all these normalised values should be unity if the level 4 and buoy uncertainties are correct. If the uncertainty estimates are under-estimated or over-estimated the standard deviation will be greater or less than unity respectively.

The buoy observations will be different to the analysis due to the mismatch in observation time and location and because of any errors in the buoy data. These issues will cause the standard deviation to be greater than unity even if the analysis uncertainties are correct if they are not accounted for. A wide range of uncertainty estimates have been reported for drifting buoys. These range from 0.12 K to 0.67 K (Kennedy 2013, RD.328). The impact of making different assumptions of drifting buoy uncertainty is shown in the results.

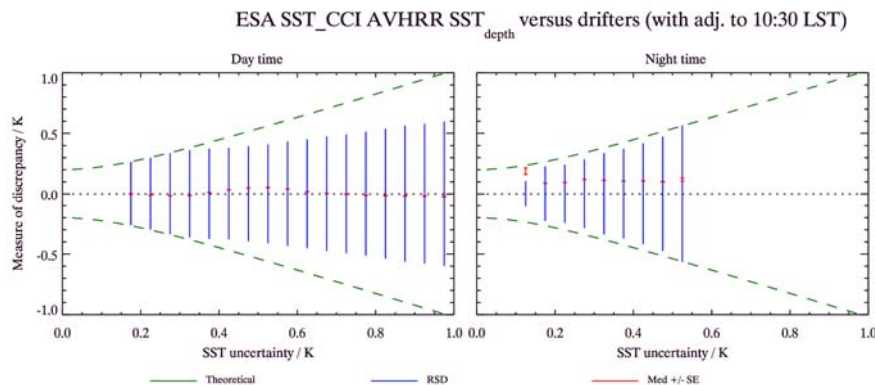
As standard deviation is a non-robust statistic, a robust measure of this quantity is used instead. This was calculated by finding the range either side of the median that encompasses 68.3% of the data. This is equivalent to one standard deviation if the distribution of quantities is Gaussian. Similarly, the range that encompasses 95.4% of the data gives a robust estimate of two standard deviations. At least 100 matchups were required for an estimate of the standard deviation to be calculated.

## 5.2 Results for SST\_CCI AVHRR products using the MMS

The results for the uncertainty validation of the entire AVHRR mission are shown in Figure 5-1. For the daytime match-ups the spread of uncertainties are from ~ 0.15 K to 1.0 K with the difference between the theoretical and measured RSD values increasing towards higher AVHRR uncertainties. This result indicates there is some discrimination between uncertainty levels but those product uncertainties > 0.4 K are generally over estimated as the measured RSD is lower than the theoretical RSD value. Indeed the degree of over estimation increases towards higher uncertainties such that uncertainties around 1.0 K in magnitude are over estimated by roughly 50%.

For night time match-ups the spread of uncertainties is lower (from 0.1 K to 0.55 K) and excellent discrimination is seen with the measured RSD tracking the theoretical RSD values across the full range.

It is noted that the standard error is fairly consistent across all uncertainties with a notable night time bias of roughly 0.1 K.



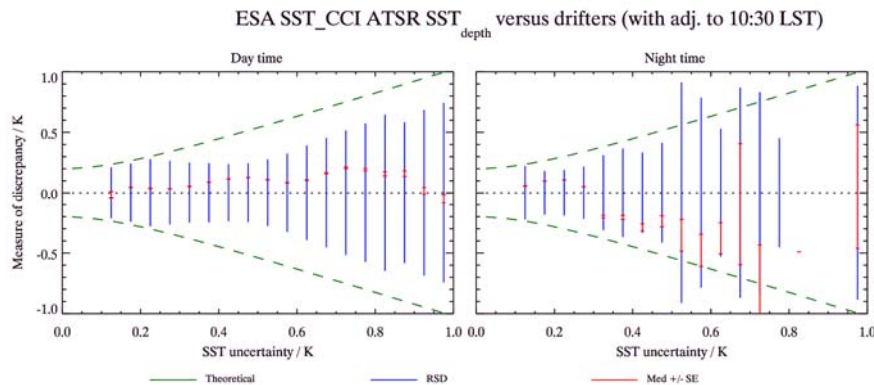
**Figure 5-1:** Plot of SST\_CCI AVHRR product uncertainty against the robust standard deviation of the discrepancies between SST\_CCI AVHRR and drifting buoys for (left) daytime and (right) night time match-ups. The green lines indicated the theoretical dispersion of uncertainties assuming an average drifter buoy measurement uncertainty of 0.2 K. The blue lines indicated the measured dispersion for each uncertainty level. The red lines indicate the standard error for each uncertainty level and also provide an indication of the number of match-ups.

## 5.3 Results for SST\_CCI ATSR products using the MMS

The results for the uncertainty validation for the whole ATSR mission are show in Figure 5-2. For the daytime match-ups the spread of uncertainties is again ranges from ~0.15 K to 1.0 K and the difference between the theoretical and measured RSD values generally increases towards higher ATSR uncertainties with some cyclic behaviour. As for AVHRR, it would seem that the ATSR day time uncertainties are generally over estimated.

For night time match-ups the spread of uncertainties is larger than observed for AVHRR and does not cover the full range contiguously. The discrimination is very good up to uncertainties of 0.5 K but then varies somewhat. However, it is also clear from the spread in the standard error observed for night time uncertainties > 0.5 K that a very low number of match-ups is causing this variability.

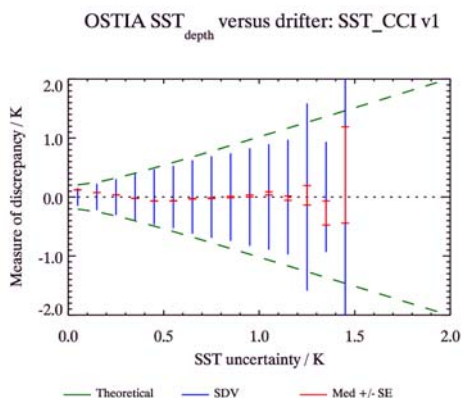
Unlike the AVHRR case, the standard error for ATSR fluctuates across the range, seeming to increase in day time towards higher uncertainties; at nighttime there is a statistically significant step change at 0.3 K where uncertainties below this value have a warm bias and uncertainties above this value have a cold bias.



**Figure 5-2:** Plot of SST\_CCI ATSR product uncertainty against the robust standard deviation of the discrepancies between SST\_CCI ATSR and drifting buoys for (left) daytime and (right) night time match-ups. The green lines indicated the theoretical dispersion of uncertainties assuming an average drifter buoy measurement uncertainty of 0.2 K. The blue lines indicated the measured dispersion for each uncertainty level. The red lines indicate the standard error for each uncertainty level and also provide an indication of the number of match-ups.

## 5.4 Results for SST\_CCI analysis products using the MMS

The results for the uncertainty validation for the SST\_CCI analysis dataset are shown in Figure 5-3. The spread of uncertainties is from ~0.05 K to 1.5 K and the agreement between the theoretical and measured RSD values is excellent across the full range of uncertainties. Some divergence is seen for uncertainties above 1.2 K but the increase in spread of the standard error indicates a low number of match-ups at these levels.



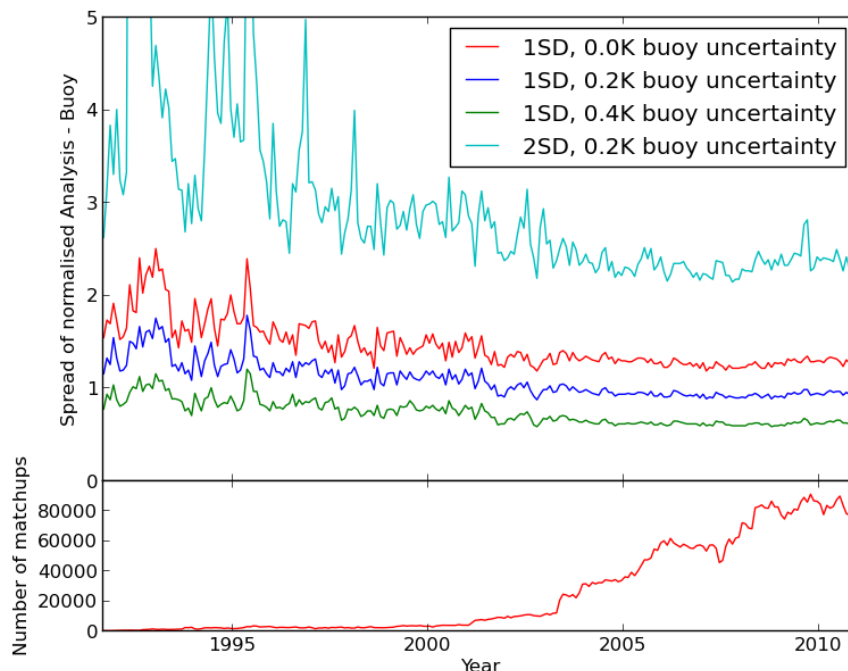
**Figure 5-3:** Plot of SST\_CCI analysis product uncertainty against the robust standard deviation of the discrepancies between SST\_CCI analysis and drifting buoys. The green lines indicated the theoretical dispersion of uncertainties assuming an average drifter buoy measurement uncertainty of 0.2 K. The blue lines indicated the measured dispersion for each uncertainty level. The red lines indicate the standard error for each uncertainty level and also provide an indication of the number of match-ups.

The standard error is reasonably consistent although there is a slightly linearity between a warm bias at uncertainties of 0.05 K to a slight cold bias at uncertainties of 0.6 K.

## 5.5 Results for SST\_CCI analysis products using re-matching

### 5.5.1 The long term product

Figure 5-4 shows the robust estimate of one or two standard deviations of normalised analysis minus buoy differences for each month of the long term product when different sizes for the drifting buoy uncertainties are assumed. Early in the record the number of buoy observations recorded at the required times was limited, leading to more noisy results than in later years. A decrease in the robust standard deviations is apparent over the first half of the record but the quantities are steady after the mid-2000s. Estimates of one standard deviation are shown for drifting buoy uncertainties of 0.0 K (red line) 0.2 K (blue) and 0.4 K (green). Dependent on which buoy uncertainty estimate is used, the analysis uncertainty estimates may be said to be underestimated, of the correct size, or overestimated. If buoy uncertainties are of order 0.2 K during the years 2000, which fits with more recent estimates, the implication is that the analysis uncertainty estimates are about right, since the ratio is close to 1. These estimates of buoy uncertainties are at the lower end of the possible range. If larger values than these were used it would suggest that the analysis uncertainties are overestimated. However, larger values are perhaps unlikely because of the high level of quality control performed on the measurements (Kennedy 2013, RD.328). Overall, it can only be concluded from these results that the analysis uncertainty estimates give a reasonable estimate of the 68% confidence range of the data.

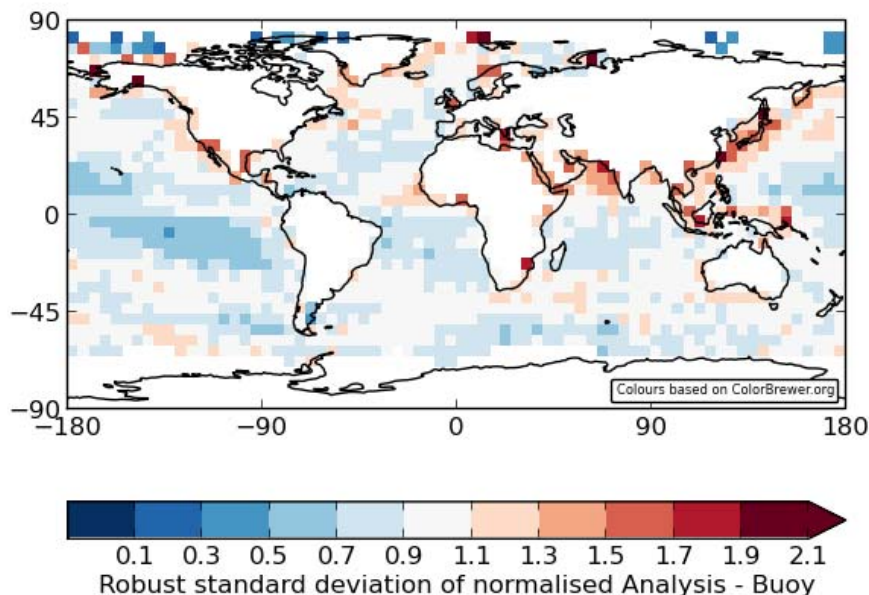


**Figure 5-4:** Robust measures of one or two standard deviation(s) of the difference between the analyses and buoy data divided by the analysis uncertainty estimates for each month of the long term product under the assumption of a range of sizes for the drifting buoy uncertainties. Lower panel: number of matches used.



Also shown in Figure 5-4 is a comparison of the robust estimates of one (blue line) and two (cyan) standard deviations when the drifting buoy uncertainty is assumed to be 0.2 K. In the late-2000s the estimate of one standard deviation is approximately one, which indicates that the uncertainty estimates are of the correct size. However, the estimate of two standard deviations is greater than two, which - in contradiction to the result for one standard deviation - indicates an underestimate of the analysis uncertainties. This occurs because the distribution of normalised buoy minus drifter data is non-normal, with excess data in the tails of the distribution.

The spatial distribution of the standard deviations is shown in Figure 5-5. A number of spatially coherent features are apparent. The values are low in the Pacific and high in the north Indian Ocean and near some coastlines such as in the north-west Pacific. However, there should be some caution in interpreting these results in high variability regions. The statistics do not incorporate estimates of the mismatch in observing time and location between the buoy and the analysis, which will be largest where variability is high.

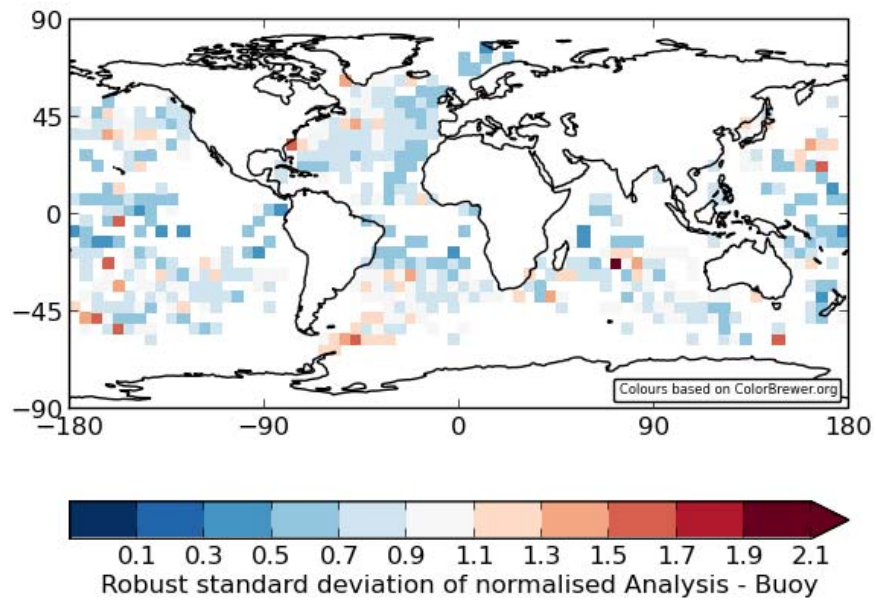


**Figure 5-5:** Standard deviation of the difference between the analyses and buoy data divided by the analysis uncertainty estimates for in 5° grid cells calculated using data from the full period of the long term product.

### 5.5.2 The demonstration product

The spatial distribution of the robust standard deviations for the first period of the demonstration product (June – August 2007) is shown in Figure 5-6. The statistics are similar to those for the long term product and similar spatial patterns are observed, although there are many gaps in the plot due to the limited amount of matchups.





**Figure 5-6:** As Figure 5-5 but showing results for the first period of the demonstration product (June – August 2007).

## 5.6 Uncertainty verification using the MMS

As part of the aim to encourage users to use the uncertainties given in the SST\_CCI products we attempt here to provide verification maps to indicate where we have independently verified the product uncertainties are of the right order of magnitude. This section contains a first attempt at such maps in order to solicit feedback from users as to their usefulness.

The maps are generated by calculating the % difference between the calculated and theoretical RSD of the differences between the SST\_CCI datasets and drifting buoys. The comparisons are carried out across the full range of uncertainties at each 15 degrees of latitude and longitude (taking into account uncertainties in the drifting buoy data). The median % difference for each latitude/longitude cell is then scaled to give an indication of verification according to:

- Very high – uncertainties are confirmed to be within 20% of their quoted values
- High – uncertainties are confirmed to be within 20% - 40% of their quoted values
- Medium – uncertainties are confirmed to be within 40% - 60% of their quoted values
- Low – uncertainties are confirmed to be within 60% - 80% of their quoted values
- Very low – uncertainties are confirmed to be within 80% - 100% of their quoted values

We also include a “not verifiable” category where it has not been possible to independently verify the product uncertainties using the reference dataset. It is important to recognise that it does not mean that the product uncertainties should not be used in these cases, just that we cannot confirm their magnitude independently.

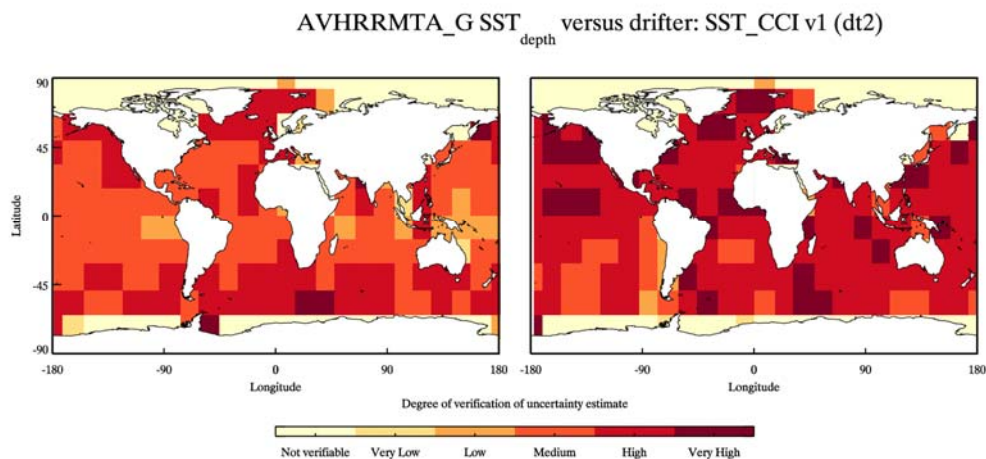
The first version presented here has the following limitations, which will be updated in the next validation report following feedback from users on the usefulness of these types of plots:

- The distribution of sea ice has not been factored in so verification results in polar regions are going to be lower than they actually are
- Currently the maps do not distinguish between cases where
  - The degree of verification is low due to a low number of match-ups from cases where
  - The degree of the verification is low due to the measured uncertainties being over-estimated or under-estimated.
- A method for implementing Type 3 or functional match-ups has not yet been implemented. NB: A functional match-up is an attempt to transfer knowledge from one region to another. For example, to look for match-ups with similar TCWV, SZA, AOD etc. and create a dummy location (no in situ). It is ‘reasonable’ to assume the uncertainty model is correlated between such locations.

Nevertheless the maps will provide users with a unique assessment of the product uncertainty quality and will allow them to scale the product uncertainties should they wish to do so in regions of low verification.

### 5.6.1 SST\_CCI AVHRR example verification map

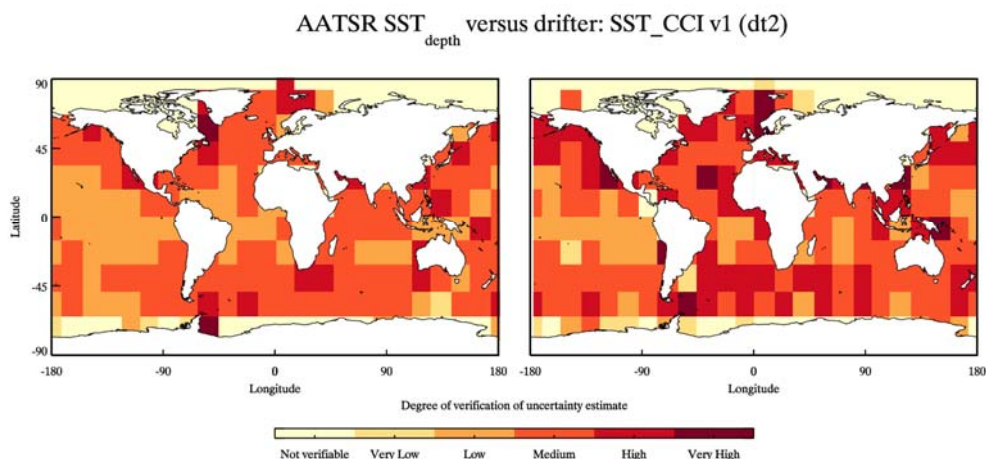
The uncertainty verification results for the AVHRR-MTA sensor are shown in Figure 5-7 for day time and night time products. The coverage is generally good with very few unverified regions. On average the uncertainties are of medium to high quality with nighttime uncertainties being better than daytime.



**Figure 5-7:** Verification maps for day time (left) and night time (right) AVHRR MTA SST<sub>depth</sub> uncertainties assessed using drifting buoy SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). This plot shows the degree to which the SST CCI product uncertainties can be verified using independent reference data. It should not be taken as an indication of SST CCI product data quality and is intended to help the user interpret their own results from applying product uncertainties in their analysis

### 5.6.2 SST\_CCI ATSR example verification map

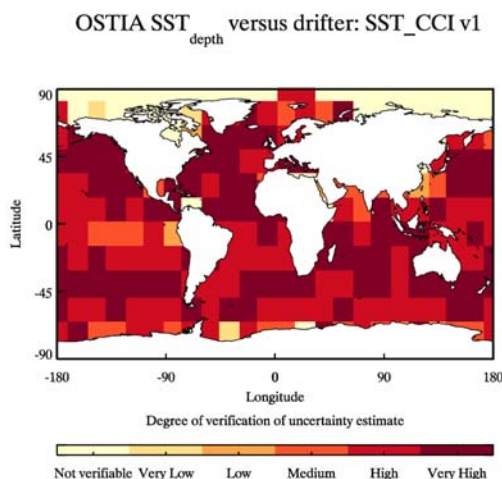
The uncertainty verification results for the AATSR sensor are shown in Figure 5-8 for day time and night time products. The coverage is generally good with very few unverified regions. On average the uncertainties are considered to have low/medium verification compared to the reference dataset; night time uncertainties are notably better than day time. This is a direct result of the notable differences between the measured and theoretical RSD values presented in Section 5.3.



**Figure 5-8:** Verification maps for day time (left) and night time (right) AATSR SST<sub>depth</sub> uncertainties assessed using drifting buoy SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). This plot shows the degree to which the SST CCI product uncertainties can be verified using independent reference data. It should not be taken as an indication of SST CCI product data quality and is intended to help the user interpret their own results from applying product uncertainties in their analysis

### 5.6.3 SST\_CCI analysis example verification map

The uncertainty verification results for the SST\_CCI L4 dataset are shown in Figure 5-9. The coverage is very good with very few unverified regions. On average the uncertainties are of high quality compared to the reference dataset and in general regions of medium and low quality occur in areas that contain low numbers of drifting buoys.



**Figure 5-9:** Verification maps for OSTIA SST<sub>depth</sub> uncertainties assessed using drifting buoy SST<sub>depth</sub>. This plot shows the degree to which the SST CCI product uncertainties can be verified using independent reference data. It should not be taken as an indication of SST CCI product data quality and is intended to help the user interpret their own results from applying product uncertainties in their analysis.

## 6. PRODUCT INTERCOMPARISON

### 6.1 Introduction

The aim of this study is to compare the new SST\_CCI analysis with existing long-term L4 SST reanalyses. The intercomparison will be achieved using the GMPE (Group for High Resolution SST (GHRSSST) multi-product ensemble) system, as described in Martin et al. (2012; RD.336). In the GMPE system, L4 SST products with a minimum spatial resolution of 0.25 degrees are regridded to the GMPE grid. This is a regular latitude-longitude grid with 0.25 degree spacing. An ensemble median SST (referred to here as the “GMPE median”) and standard deviation at each point is calculated. The production of a median SST using all the datasets provides a new SST product which potentially has smaller errors than any of the component reanalyses, as was found for the NRT version of the GMPE system (Martin et al., 2012; RD.336). All land-sea-ice points are set to missing data. The updated sea ice field for each day is taken here from the SST\_CCI analysis. When comparing the reanalyses and the GMPE median to Argo data the full resolution reanalyses are interpolated to the observation locations.

### 6.2 Intercomparison datasets

Only L4 (global, gap-free, gridded) SST reanalyses with a minimum of 10 years of data were considered. In total, 6 reanalysis datasets including OSTIA CCI were used. OSTIA CCI is the only dataset not to use in situ data as an input, and is based on satellite data only. Although the datasets are all “SST” products, they are valid for different depths. OSTIA CCI uses input data specifically corrected to 20 cm. Input data are also corrected to 1030 hrs and 2230 hrs local time, producing an estimate of the daily mean temperature at this depth. This is the only reanalysis which uses methods to try to produce data valid for a specific depth and local time. The HadISST2 dataset is also valid for a nominal depth of 20 cm. The OSTIA v1.0 and MGDSSST reanalyses are foundation temperatures, which mean they are approximately pre-dawn temperatures, without the effects of diurnal warming. AVHRR-OI is a mean in the sense that all data are used, but an actual daily mean temperature is not necessarily produced. The satellite data used in the CMC reanalysis is referenced to ship and buoy data with a typical depth of 1 m, although no particular method was applied to the reanalysis to make it valid for a particular depth. All of these datasets use optimal interpolation analysis methods. Detailed descriptions of these reanalyses and the methods used to generate them are provided in the references given in table 1.

In addition to the reanalyses shown in Table 6-1, a short period of the OSTIA CCI reanalysis (June, July, and August 2007) was rerun with the addition of data from AMSR-E and TMI microwave instruments. This demonstration product, demo 1, was also compared to the other reanalyses in the long-term GMPE system (section 6.3.2).

	OSTIA CCI	OSTIA v1.0	CMC	HadISST2 (realisation 396)	MGDSST	AVHRR-OI
<b>Time period</b>	1991-2010	1985-2007	1991-2011	1899-2007	1982-2011	1981-present
<b>AVHRR</b>	NOAA12-19 [CCI]	Pathfinder (1985-2007)	NOAA16-19 (2001-2011) [NAVO]; MetOp-A (2007-2011) [NAVO]	Pathfinder (1981-2006)	Pathfinder (1982-2006); NOAA17-19 (2007-2011) [NESDIS]; MetOp-A (2010-2011) [NESDIS]	Pathfinder (1985-2005); NOAA-unspecified (2006-present) [NAVO]
<b>ATSR-series</b>	ATSR-1,2, AATSR [CCI]	ATSR-1,2, AATSR [NEODC]	ATSR-1,2, AATSR [ESA]	ATSR-1,2, AATSR [ARC]	None	None
<b>AMSR-E</b>	None	None	2002-2011 [REMSS]	None	2003-2011 [JAXA]	None
<b>TMI</b>	None	None	1998-2002 [REMSS]	None	None	None
<b>WindSat</b>	None	None	2003-2011 [REMSS]	None	2011 [JAXA]	None
<b>In situ</b>	None	ICOADS	ICOADS; GTS (final 5 years)	ICOADS	GTS	ICOADS
<b>Resolution</b>	1/20°	1/20°	1/5°	1/4°	1/4°	1/4°
<b>SST depth</b>	Daily mean at 20 cm depth	Foundation	1 m (referenced to ship and buoy data)	20 cm	Foundation	Mean
<b>Reference</b>		Roberts-Jones et al. (2012)	Brasnett (2012)	Kennedy et al. (2013)	Kurihara et al. (2006)	Reynolds et al. (2007)

Table 6-1: Input datasets to long-term GMPE

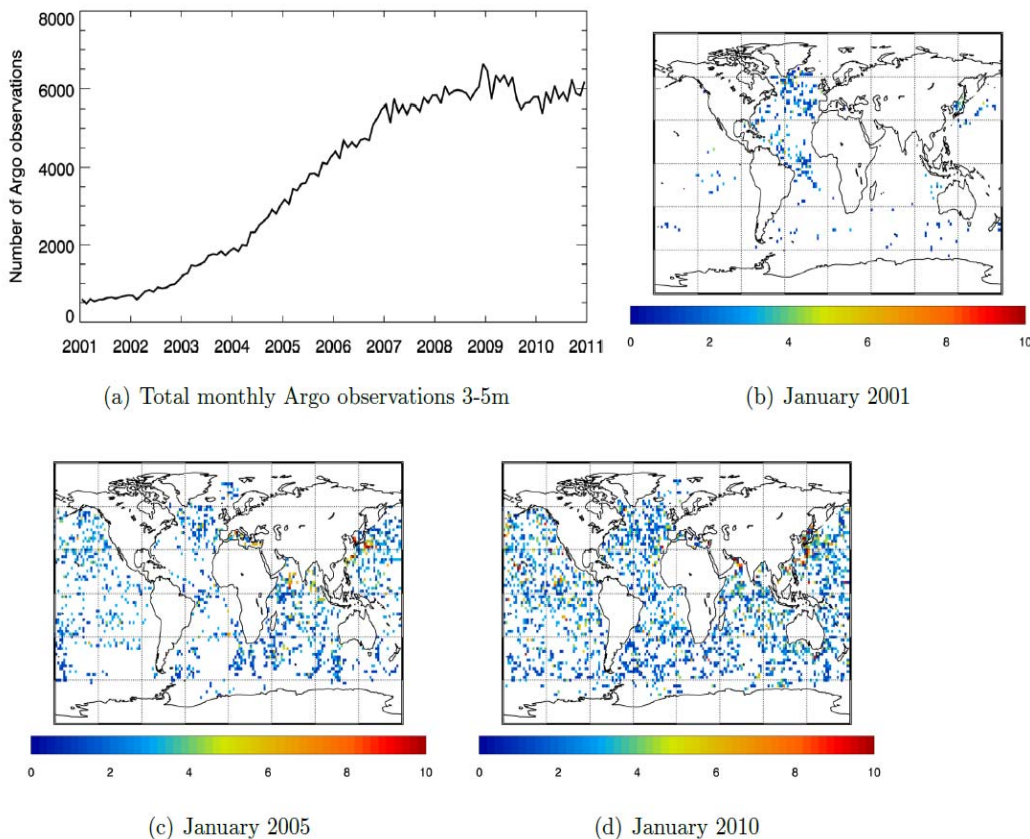
## 6.3 Inter-comparison of reanalyses

### 6.3.1 Validation of 10-year time series against independent data from Argo

None of the contributing L4 products uses Argo data, making this dataset suitable for independent validation. Near-surface (3-5 m depth) Argo measurements have been shown to provide a good estimate of foundation SST using a triple collocation of Argo data, surface drifters and AATSR satellite data (Merchant and Corlett, 2010, pers. comm., see Martin et al., 2012; RD.336). However, as mentioned in section 6.2, the various



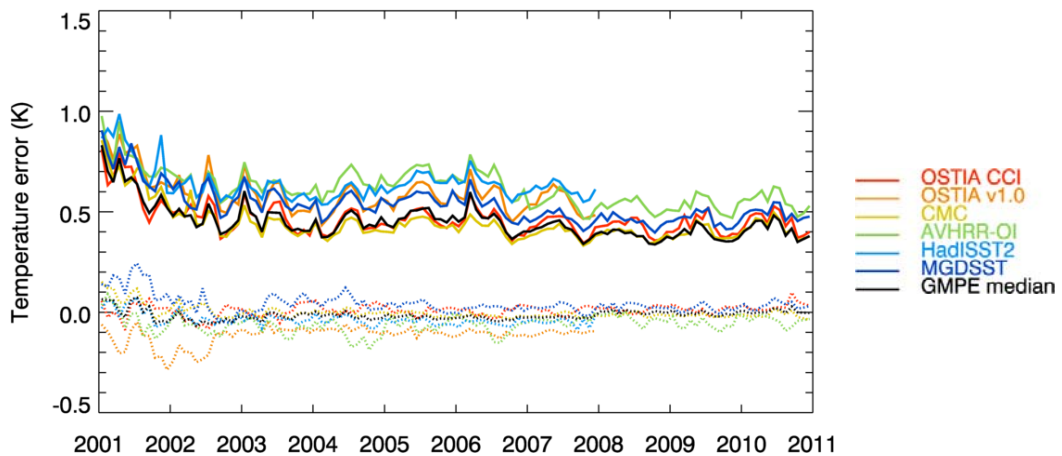
reanalyses are intended to be valid at different depths (Table 6-1). We would therefore expect to find biases compared to the Argo foundation temperature and this should be taken into account when comparing the following results. For example, the OSTIA CCI reanalysis, a daily mean at 20 cm depth, would be expected to be warmer than a foundation temperature.



**Figure 6-1:** (a) Monthly global total number of Argo measurements between 3m and 5m depth and (b)-(d) number of measurements for example months on 2x2 degree grid for beginning, middle and end of time period.

Argo observations have been extracted from the EN4 dataset (Good et al., 2013; RD.264). The observations have undergone quality control procedures to remove suspect observations as described in Good et al. (2013; RD.264). Argo observation times are distributed throughout the day. Argo data are available from the year 2000 but few observations from this time mean the comparisons have been performed here only for 2001 onwards. Figure 6-1 shows a time series of the total monthly global number of Argo observations used for the comparisons and shows the spatial distribution of these observations for January 2001, January 2006 and January 2010 as example months. Figure 6-1 demonstrates the maturing of the Argo dataset by 2007 and the spreading of the observation network to (almost) cover the global ocean.

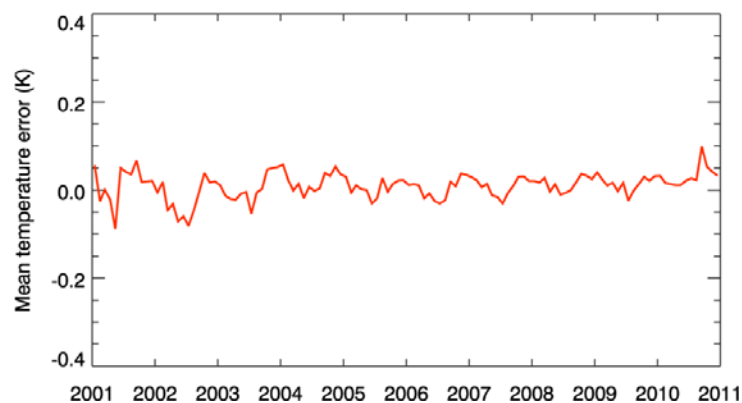




**Figure 6-2:** Global monthly standard deviation (solid line) and mean error (dashed line) compared to Argo between 3-5 m depth, for all reanalyses used in long-term GMPE, and GMPE median. Mean error is reanalysis-minus-Argo.

Figure 6-2 shows a monthly time series of standard deviation and mean error for each of the reanalyses compared to Argo, for the globe. Figure 6-2 demonstrates the effect of a reduced match-up data volume on the statistics. The statistics level out towards the end of 2002 as the number of Argo observations increases.

The reanalyses tend to group together in terms of global standard deviation (Figure 6-2). CMC, OSTIA CCI and the GMPE median have the lowest standard deviations over the period 2001-2010, MGDSSST and OSTIA v1.0 are in the middle, and AVHRR-OI and HadISST2 have the largest global standard deviations.



**Figure 6-3:** Global mean error (reanalysis-minus-Argo) for OSTIA CCI reanalysis only

Figure 6-3 illustrates there is a seasonal cycle in the global bias of the OSTIA CCI reanalysis compared to Argo (reanalysis-minus-Argo) which is not seen in the other reanalyses. This is of the order 0.15 K at the beginning of the period (2001), decreasing to around 0.10 K after 2003. Closer inspection reveals this is related to regional temperature cycling, which is seen throughout the time series (Figure 6-8 and Figure 6-9). This will be discussed further in section 6.4.

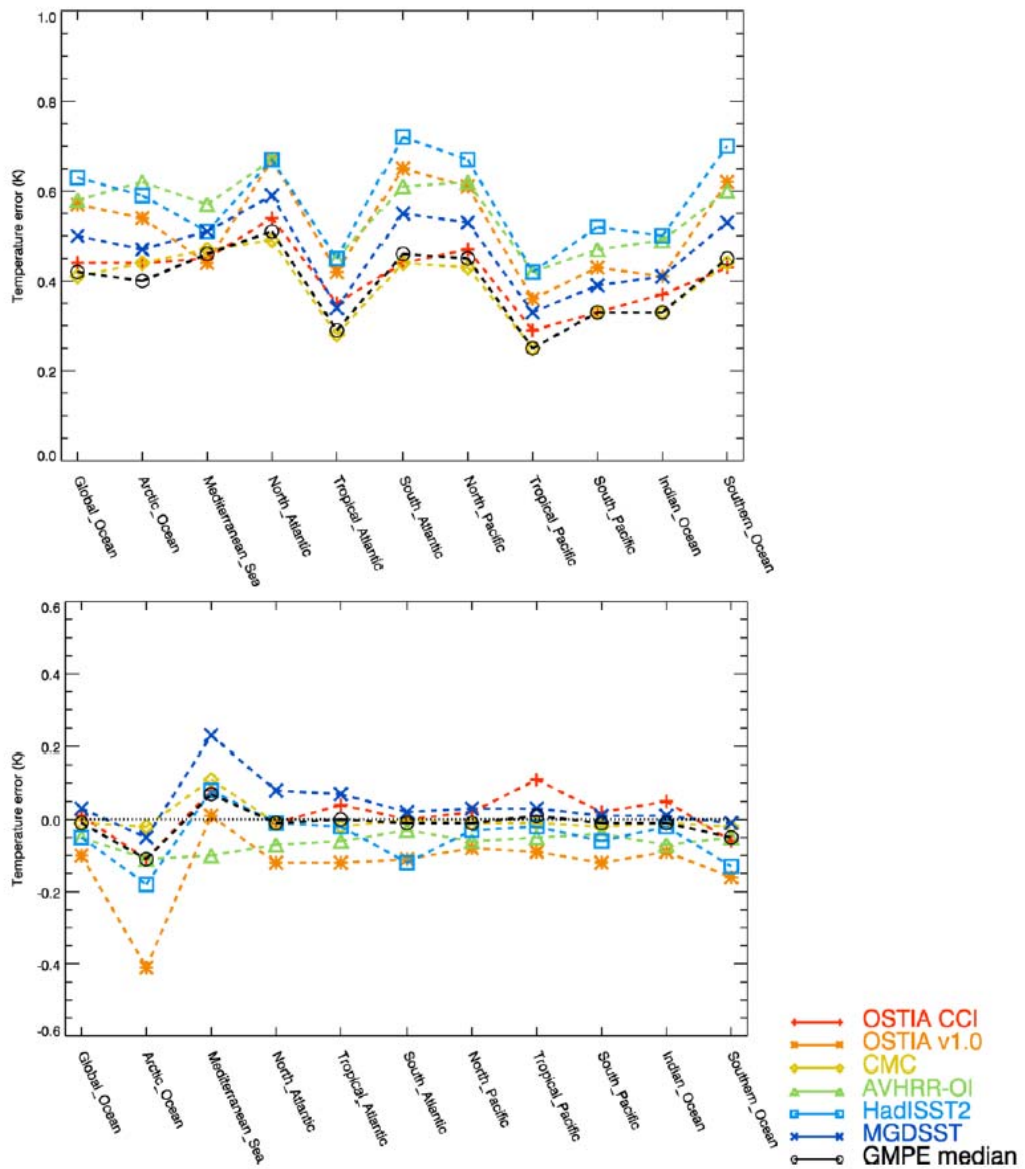
Analysis	STD	Mean Error	Num Argo Obs
OSTIA CCI	0.44	0.01	448245
OSTIA v1.0	0.57	-0.10	233615
CMC	0.41	-0.01	448244
AVHRR-OI	0.58	-0.05	448247
HadISST2	0.61	-0.04	230692
MGDSST	0.49	0.03	448230
GMPE median	0.41	-0.01	446505

**Table 6-2:** Global standard deviation (K), mean error (K) (reanalysis-minus-Argo) and number of Argo observations, 2001-2010. Note OSTIA v1.0 and HadISST2 finish in 2007.

Table 6-2 shows the global mean statistics compared to Argo for each of the reanalyses over the time period 2001-2010 (or 2007 for HadISST2 and OSTIA v1.0). The reanalysis with the smallest mean standard deviation is CMC, which at 0.41 K has the same global mean standard deviation as the GMPE median. This is a surprising and impressive result, given that the GMPE median was found to be better than all its component analyses in the NRT version of GMPE (Martin et al., 2012; RD.336). CMC is the only contributing reanalysis to use microwave data for the whole of the time period we are considering (Table 6-1).

OSTIA CCI also performs well, with a small standard deviation compared to other reanalyses (Table 6-2). This is particularly impressive given that it is purely a satellite product, unlike the other reanalyses which also use in situ observations. OSTIA CCI (which uses improved input data and an upgraded version of the OSTIA system) is clearly an improvement over the OSTIA v1.0 reanalysis. This will be discussed further in section 6.3.3.

The GMPE median, CMC and OSTIA CCI datasets all have the lowest magnitude global biases. However, regional biases should also be examined, as these can average out in a global value as presented here.

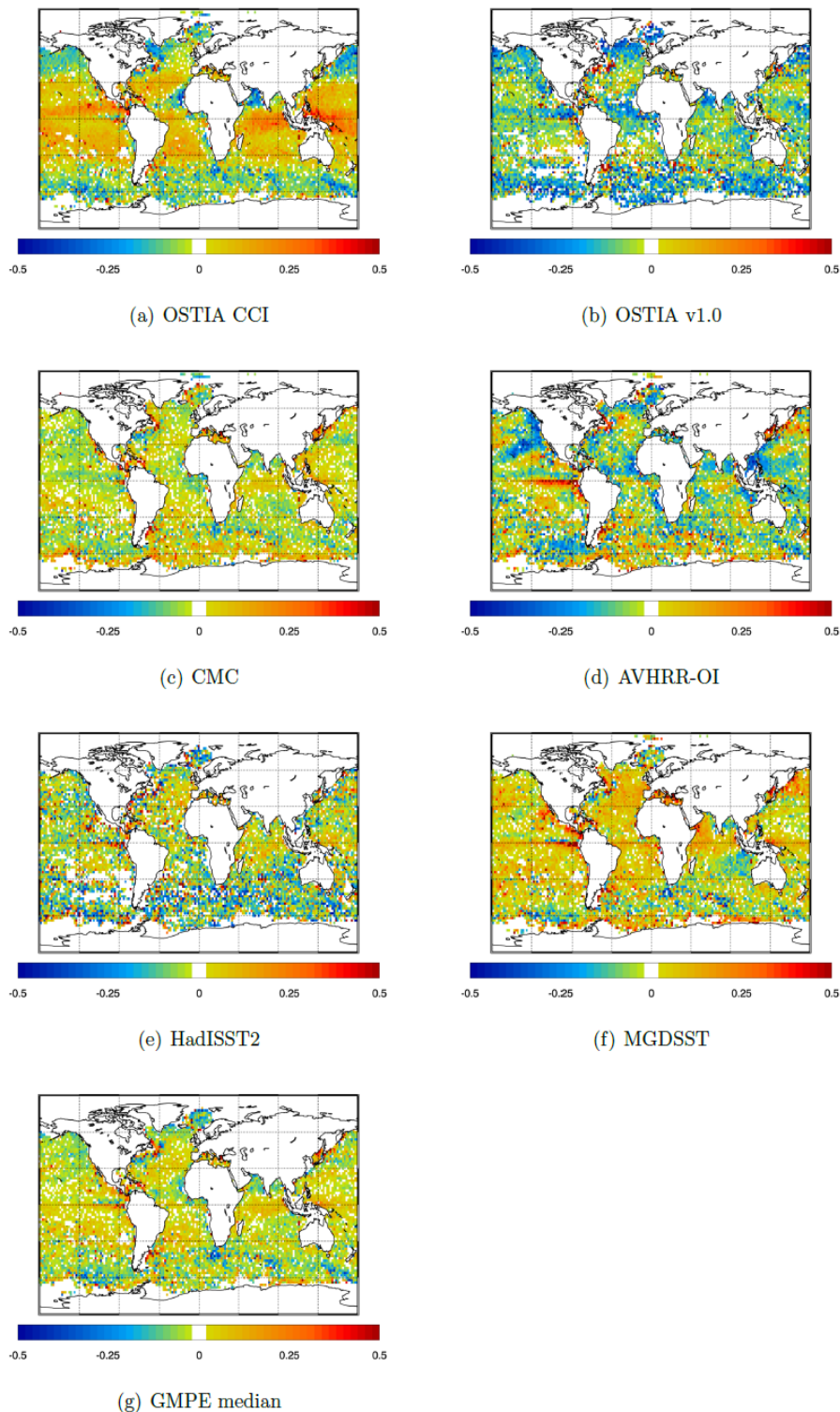


**Figure 6-4:** Standard deviation (top) and mean error (bottom) for ocean regions, 2001-2010, compared to Argo between 3-5 m depth for each reanalysis and GMPE median. Mean error is reanalysis-minus-Argo.

Figure 6-4 shows the statistics for the 2001-2010 period split into ocean region (using the MyOcean definitions). This demonstrates that some reanalyses perform better than others in certain regions. In particular, although OSTIA CCI performs well in the Northern and Southern mid-latitudes, the bias in the Tropics is larger, and exceeds 0.1 K in the Tropical Pacific. However, this bias in the Tropics is similar in magnitude to that found for some other reanalyses in the Tropics. OSTIA CCI, CMC and the GMPE median all perform well regionally in terms of the RMS error (Figure 6-4).

The robustness of the statistics shown in Figure 6-4 is dependent on the number of observations available in particular regions. For example, the Arctic and the Mediterranean have fewer observations than other ocean areas, but over the whole

period 2001-2010 there should be enough observations to be able to draw sensible conclusions.

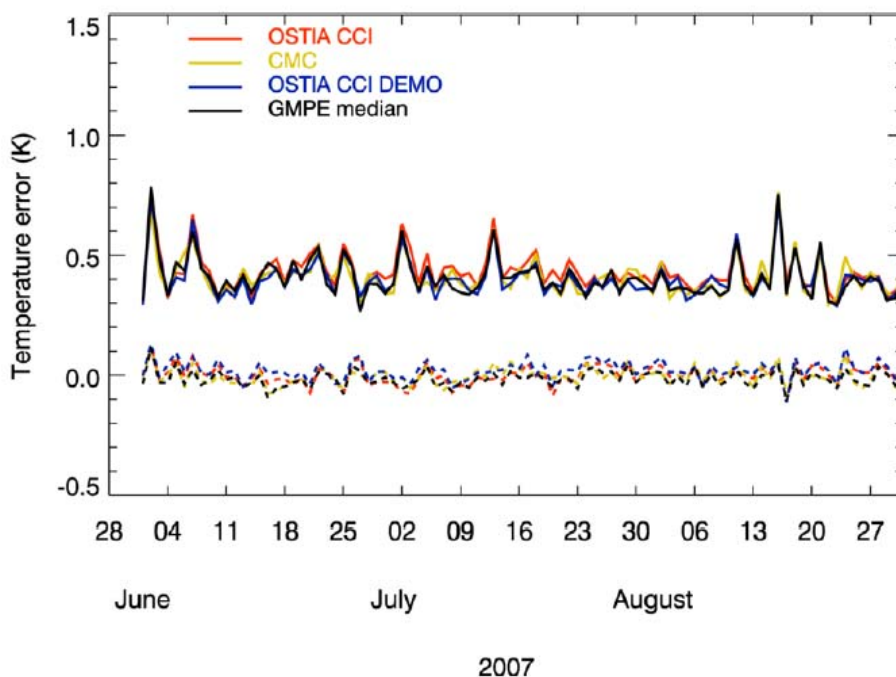


**Figure 6-5:** Mean error (reanalysis-minus-Argo) on 2x2 degree grid, for 2001-2010, for each reanalysis and GMPE median.

Figure 6-5 shows spatial plots of the mean error (reanalysis-minus-Argo) for each reanalysis, and the GMPE median. The GMPE median and CMC reanalysis have the smallest global bias (Table 6-2) and a small bias is consistent over the different regions of the globe (Figure 6-4, Figure 6-5c and f). Although the global average of the mean error for OSTIA CCI is small (Table 6-2) the error in the Tropics is up to  $\sim 0.25$  K (Figure 6-5a), showing a positive bias in the reanalysis for this region, i.e. the analysis is too warm. This is also seen in the average regional statistics given in Figure 6-4 and reflects biases in the input data (e.g. Figure 4-15).

### 6.3.2 Effect of including microwave data in OSTIA CCI reanalysis

The period June to August 2007 (JJA 2007) was rerun using the OSTIA CCI system with the addition of microwave data from the AMSR-E and TMI instruments, to produce the demo 1 product. Microwave data has the advantage of providing surface temperature measurements without being affected by clouds (provided precipitation is not heavy) and is thus able to provide a surface SST where infra-red instruments cannot. However, microwave data is known to be less accurate and of lower resolution than infra-red, and the large footprint of microwave instruments means there is a lack of data around coasts. There is therefore a trade-off between potential improvements to the analysis from more data and potential degradation because of input data errors.



**Figure 6-6:** Daily global standard deviation (solid line) and mean error (dashed line) for four reanalyses for JJA 2007, compared to Argo 3-5 m depth. Mean error is reanalysis-minus-Argo.

Figure 6-6 shows a time series of global statistics for JJA 2007 against Argo data, comparing the OSTIA CCI reanalysis with the OSTIA CCI demonstration product including microwave data. CMC and the GMPE median are also shown, as these datasets produced the best results on comparison with independent Argo data (section 3.1). Regional statistics for this three-month period are given in Table 6-3.



Region	OSTIA CCI		OSTIA CCI demo		CMC		GMPE median	
	STD	Mean Error	STD	Mean Error	STD	Mean Error	STD	Mean Error
Global	0.44	0.00	0.41	0.02	0.42	-0.01	0.42	-0.02
N Atlantic	0.57	0.04	0.52	0.07	0.52	0.07	0.54	0.05
Tr Atlantic	0.23	0.10	0.22	0.10	0.20	0.00	0.19	0.01
S Atlantic	0.33	-0.02	0.32	0.00	0.36	-0.06	0.34	-0.06
N Pacific	0.49	-0.07	0.44	-0.04	0.42	0.01	0.42	-0.03
Tr Pacific	0.38	-0.03	0.34	-0.02	0.31	-0.02	0.31	-0.04
S Pacific	0.34	0.01	0.33	0.02	0.36	-0.03	0.34	-0.04
Indian Ocean	0.32	0.05	0.31	0.06	0.30	-0.03	0.30	0.02
Southern Ocean	0.40	-0.07	0.41	-0.09	0.43	-0.06	0.41	-0.09

**Table 6-3:** Statistics comparison JJA 2007, reanalysis-minus-Argo.

The inclusion of microwave data in the OSTIA CCI reanalysis improves the global standard deviation consistently over this time period, making it slightly better than both CMC and the GMPE median (Figure 6-6, Table 6-3). However, it does introduce a more positive bias in the analysis compared to results for the long-term OSTIA CCI product (0.02 K compared to 0.00 K respectively) for this period.

Table 6-3 also shows the mean statistics for JJA 2007 for various areas of the global oceans, defined using the MyOcean regions. Results for the Arctic and Mediterranean regions are not given here due to the small amount of data available in these areas over a three month period. An improved standard deviation is seen in all ocean regions (except the Southern Ocean) for the demo 1 product compared to the long-term OSTIA CCI reanalysis. In most regions the bias is more positive for demo 1 compared to OSTIA CCI, and sometimes increases the bias in magnitude. However, we do expect a positive bias when comparing the OSTIA CCI product with the Argo foundation temperature (section 6.3.1).

Figure 6-7 shows spatial plots of the mean error to Argo for JJA 2007 for OSTIA CCI, demo 1 (OSTIA CCI demo), CMC and the GMPE median. OSTIA CCI demo (Figure 6-7b) also shows a positive bias in the Tropics, as seen in the long-term OSTIA CCI reanalysis (Figure 6-7a and Figure 6-5a), albeit relatively modest for this particular period.

Therefore it seems inclusion of microwave data improves the precision of the OSTIA CCI reanalysis, likely by providing data in cloudy regions, but is generally unable to correct biases. The microwave data are bias-corrected to AATSR in the OSTIA system, and therefore are not able to correct any biases in the reanalysis caused by the AATSR input data.

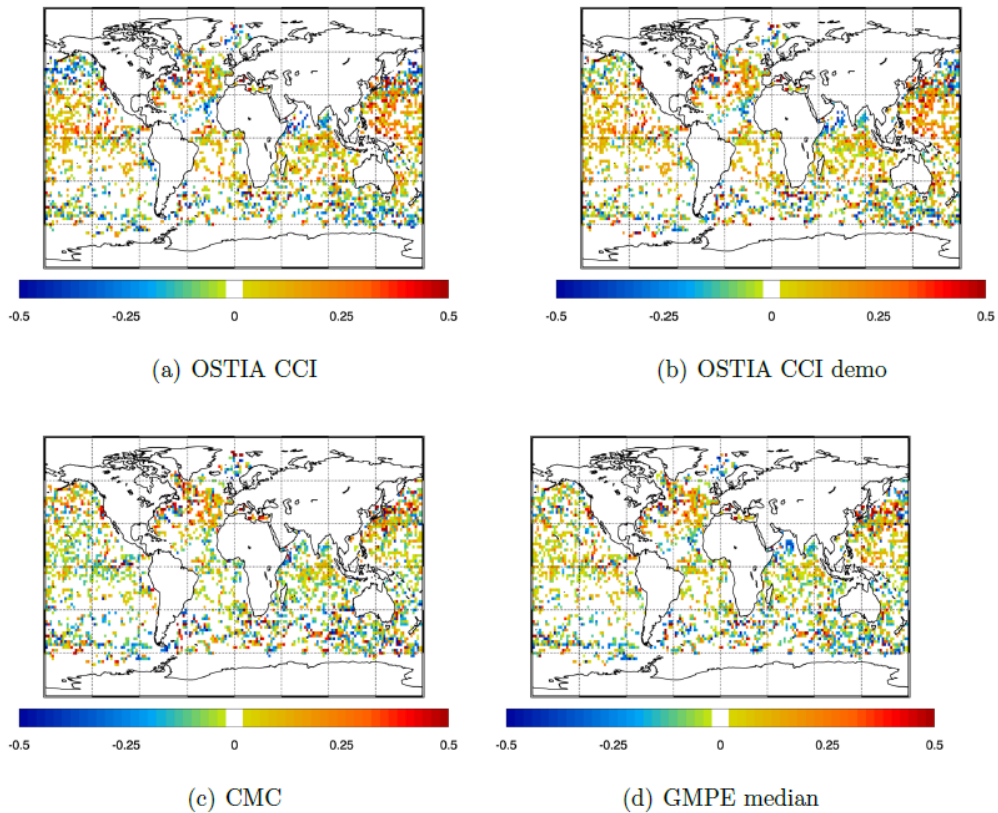


Figure 6-7: Mean error (reanalysis-minus-Argo) on 2x2 degree grid, for JJA 2007.

### 6.3.3 Comparison of OSTIA CCI to OSTIA v1.0

The OSTIA CCI reanalysis was produced using new input data and an updated version of the system used to produce the OSTIA v1.0 reanalysis (Roberts-Jones et al., 2012 [RD.239]; Roberts-Jones et al., 2013 [RD.294]). In all regions, the standard deviation is improved (reduced in magnitude) for the new OSTIA CCI reanalysis compared to the OSTIA v1.0 reanalysis (Table 6-4). The same is true of the bias in all regions, apart from the Tropical Pacific. The issue in the Tropics indicates reflects bias in the OSTIA CCI input data in this region, as mentioned previously in section 6.3.1.

For comparison, the GMPE median and CMC statistics are also shown in Table 6-4. The values for OSTIA CCI are much closer to the statistics for the GMPE median and CMC than are those for OSTIA v1.0, demonstrating the newer OSTIA reanalysis product is now in line with the best-performing SST reanalyses.



Region	OSTIA CCI		OSTIA v1.0		GMPE median		CMC	
	STD	Mean Error	STD	Mean Error	STD	Mean Error	STD	Mean Error
Global	0.44	0.01	0.57	-0.10	0.42	-0.01	0.41	-0.01
N Atlantic	0.54	-0.01	0.67	-0.12	0.51	-0.01	0.49	-0.01
Tr Atlantic	0.35	0.04	0.42	-0.12	0.29	0.00	0.28	-0.02
S Atlantic	0.44	0.00	0.65	-0.11	0.46	-0.01	0.44	0.00
N Pacific	0.47	0.02	0.61	-0.08	0.45	-0.01	0.43	-0.01
Tr Pacific	0.29	0.11	0.36	-0.09	0.25	0.01	0.25	-0.01
S Pacific	0.33	0.02	0.43	-0.12	0.33	-0.01	0.33	-0.02
Indian Ocean	0.37	0.05	0.41	-0.09	0.33	-0.01	0.33	-0.01
Southern Ocean	0.43	-0.06	0.62	-0.16	0.45	-0.05	0.44	-0.02

**Table 6-4:** Statistics comparison 2001-2010, reanalysis-minus-Argo.

## 6.4 Intercomparison using GMPE data

### 6.4.1 Reanalysis anomaly to GMPE median

No data independent to the reanalyses are available prior to the Argo dataset used from 2001. In order to gain some insight into the relative accuracy of the contributing reanalyses throughout the whole period 1991-2010, comparisons of the anomaly of each reanalysis to the GMPE median were made. These are displayed on Hovmuller plots (Figure 6-8), where the monthly anomaly by latitude has been calculated on a 2x2 degree grid for each reanalysis.

Figure 6-8a shows the difference to the GMPE median of the OSTIA CCI reanalysis. Several features are immediately obvious. There is a distinct seasonal cold bias at around 50N which is consistent throughout the whole time period. This was found to begin in spring (March, April, May) and deepen in summer (June, July, August) (Figure 6-9). This could indicate an issue with the input data. As this is a seasonal feature, the cold bias does not show up as strongly in figure 5a as the more persistent warm bias does in the Tropics. This Tropical bias is also seen in Figure 6-8a, and does have a seasonal component although this is smaller than is seen at 50N.

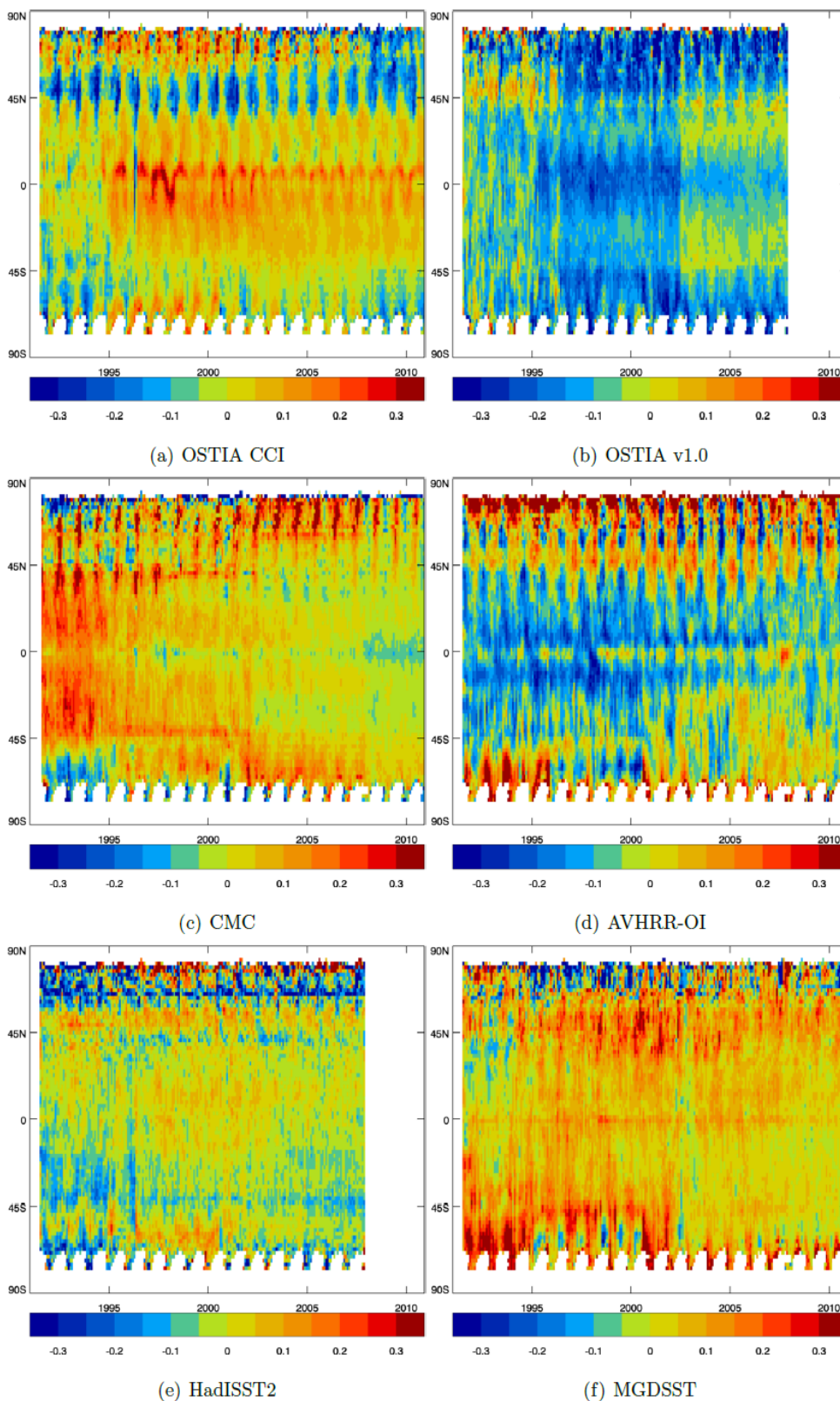
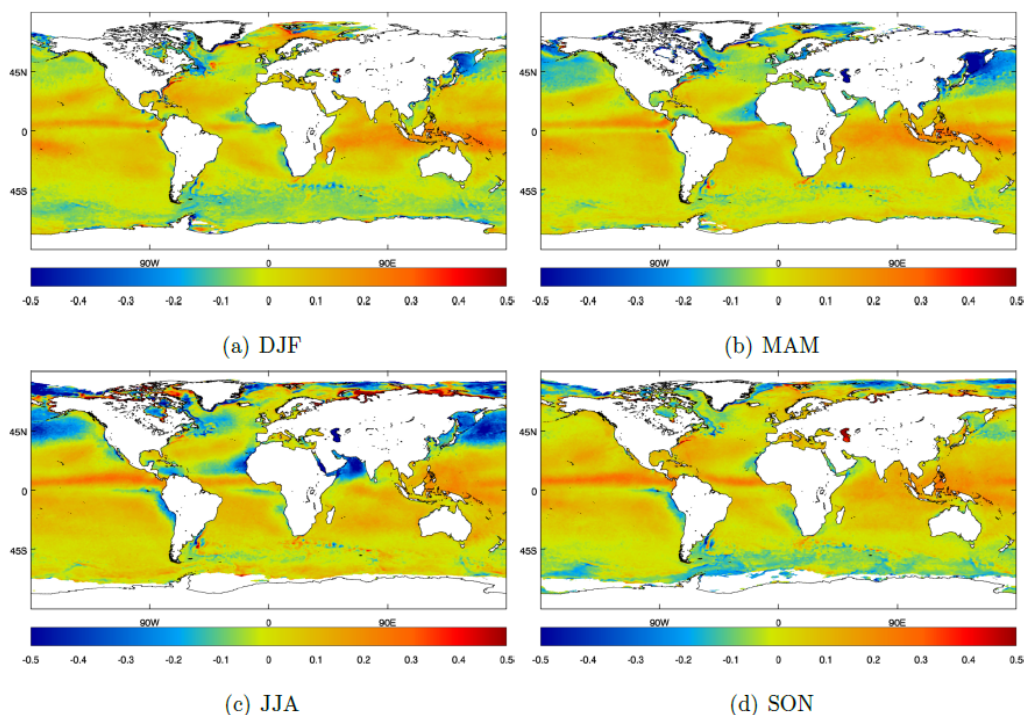


Figure 6-8: Monthly anomalies to GMPE median by latitude of the contributing reanalyses, on 2x2 degree grid.



**Figure 6-9:** Mean seasonal 1991-2010 anomaly to GMPE median for OSTIA CCI reanalysis.

The magnitude of the warm Tropical bias varies in the time periods when the different instruments of the ATSR series are used:

- ATSR-1: 1991/08 – 1995/05, 1996/01 – 1996/06
- ATSR-2: 1995/06 – 1995/12, 1996/07 – 2002/07
- AATSR: 2002/07 – 2010/12

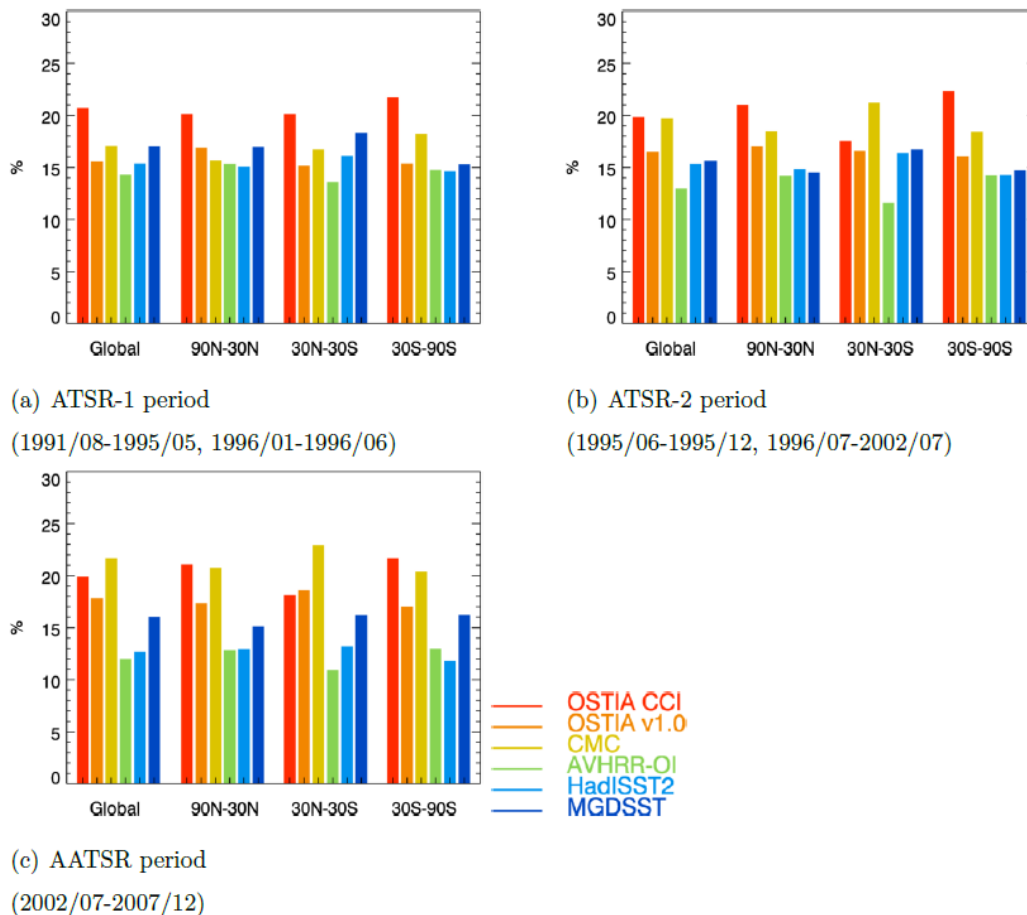
The Tropical warm bias is strongest in the ATSR-2 period, weaker in the AATSR period and does not appear in the ATSR-1 period (Figure 6-8a). There is a brief period of 7 months when ATSR-1 returns after a 7-month switch to ATSR-2. During this time a distinct cold anomaly appears in the Tropics. This occurs from mid-May 1996 to early June 1996.

The CMC reanalysis shows a warm bias in the Tropics (extending to 45N and S) to the GMPE median in the ATSR-1 period only (Figure 6-8c). The bias for the ATSR-2 period is smaller in magnitude than in the ATSR-1 period, and the AATSR period shows improvement again.

OSTIA v1.0 shows three distinct periods of bias to the GMPE median corresponding to the use of ATSR series data in this reanalysis (Figure 6-8b). For this product, the ATSR-1 data was not used in the ATSR-2 gap. Although the magnitude of the difference in bias is small (around 0.05 K) it is distinct and seen at all latitudes. The AATSR period for OSTIA v1.0 shows the most organised pattern of anomalies with distinct latitude bands, compared to the ATSR-1 and ATSR-2 periods. MGSST and AVHRR-OI do not include data from the ATSR series of instruments so do not show these same patterns (Figure 6-8d, Figure 6-8f). They both however show an improvement in the bias to the GMPE median in the Southern Hemisphere towards the end of the time series. Although it uses ATSR data, HadISST2 does not show distinct boundaries for the ATSR periods.

### 6.4.2 Analysis contribution to the median

Figure 6-10 is a summary of the contribution of each analysis to the GMPE median for the three periods of the ATSR series, for various latitude bands. Two of the reanalyses finish in 2007 (HadISST2 and OSTIA v1.0) so results are given up to and including that year for the AATSR period.



**Figure 6-10:** Percentage contribution of grid points to GMPE median for each reanalysis, for the three periods of the different ATSR-series instruments. AATSR only shown to 2007 as OSTIA v1.0 and HadISST2 reanalyses end in this year. It is assumed ATSR-1 is used in ATSR-2 gap but this is not necessarily the case for all reanalyses.

Figure 6-10a shows statistics for the ATSR-1 period 1991/08-1995/05 and 1996/01-1996/06. In this period, for the Northern and Southern latitudes beyond the Tropics (90N-30N and 30S-90S), and the Tropics themselves (30N-30S), the OSTIA CCI reanalysis makes the largest number of contributions to the median. These are wide latitude bands, so the seasonal temperature cycling centred on 50N in OSTIA CCI (section 4.1) does not affect these statistics. Figure 6-10b shows statistics for the ATSR-2 period 1995/06-1995/12 and 1996/07-2002/07. In the Northern and Southern latitudes, OSTIA CCI still has the largest percentage of contributions to the median, but in the Tropics, where the OSTIA CCI bias is poorer than for other regions (section 3.1) the contribution to the GMPE median is smaller, and CMC has the highest percentage of contributions. Figure 6-10c shows statistics of the reanalyses' contributions to the GMPE median for the AATSR period 2002/07-2007/12. In the Tropics, OSTIA CCI now has only the third highest contribution to the median, behind CMC and OSTIA v1.0, with MGDSST not far

behind OSTIA CCI. In the Northern and Southern latitudes OSTIA CCI still has the largest number of contributions to the median, but CMC is very close. In conclusion, OSTIA CCI performs well in terms of grid point contributions to the GMPE median. However, in the ATSR-2 and AATSR periods in the Tropics the number of contributions falls behind those of other reanalyses.

## 6.5 Feature resolution

Figure 6-11 shows horizontal gradients (combined for North-South and East-West directions) for the Gulf Stream region for an example date, 01 July 2007, given in K per 100 km. The gradients are calculated on the native grid for each reanalysis and interpolated to the same 0.25 degree GMPE grid before plotting. The sharpness of the gradients in these plots illustrates the ability of the reanalysis to capture high-resolution features.

OSTIA CCI has the strongest gradients. Visual analysis of animations of the daily gradient field around the time period shown in Figure 6-11 indicates these gradients are likely to be an accurate representation of fronts, and unlikely to be noise. Feature resolution in OSTIA CCI is much improved compared to OSTIA v1.0, through upgrades to the background error covariances and the number of iterations performed by the analysis scheme (Roberts-Jones et al., 2013; RD.333). Feature resolution in the OSTIA CCI demo 1 product is very marginally poorer than for the OSTIA CCI reanalysis (Figure 6-11), due to the inclusion of lower resolution data from microwave instruments in the demo product. However, both OSTIA CCI and the demo 1 product compare well against the other reanalyses. CMC also compares well. MGDSST has strong gradients although some noise is seen. The GMPE median is smoother than OSTIA CCI and CMC, which is to be expected given it is an average of 6 different reanalyses.

## 6.6 Summary of product intercomparison results

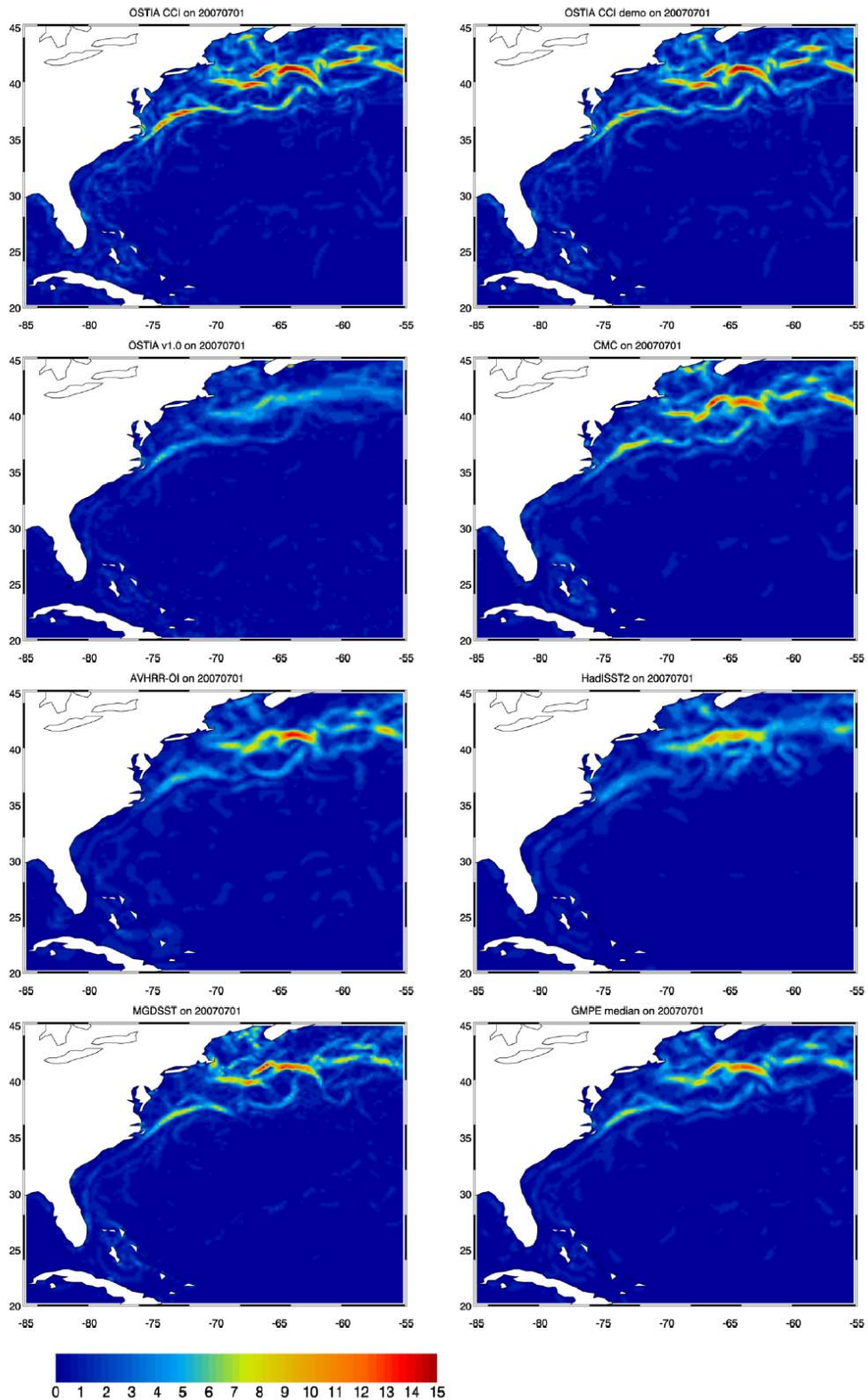
The OSTIA CCI SST reanalysis has been compared to five other long-term SST reanalyses and a median product generated from all six, using a version of the NRT GMPE system.

OSTIA CCI performs well compared to the other products and the GMPE median. The bias and RMS error compared to near-surface, independent Argo data are similar to those found for the two best-performing products, CMC and the GMPE median, for most regions, despite not including any in situ observations. The exceptions to this are a warm bias in the Tropics, reflecting biases in SST CCI L2P and L3U input data, and a seasonal cold bias centred around 50N and largest in the summer. The inclusion of SST CCI L2P microwave data in OSTIA CCI for a three-month period (demo 1) improves the standard deviation compared to Argo data for almost all ocean regions. The bias is more positive for the demo 1 reanalysis than for the long-term reanalysis, for this period. The OSTIA CCI reanalysis and the demo 1 reanalysis also compare very well to the other reanalyses in terms of feature resolution.

A Hovmuller plot of the anomaly to the GMPE median of the OSTIA CCI reanalysis by latitude over the whole time period shows the effect of including the different ATSR instruments on the Tropical bias. The bias is largest for the ATSR-2 period and still apparent for the AATSR period, but is not seen when data from the ATSR-1 instrument is being used. In this earlier period, OSTIA CCI makes the largest global percentage contribution of grid points to the GMPE median of all six reanalyses. In the later period,



especially in the Tropics, the contribution to the GMPE median from OSTIA CCI is reduced.



**Figure 6-11:** Horizontal SST gradients for the Gulf Stream region for each reanalysis including demo 1 and the GMPE median, for 01 July 2007 in K per 100 km.



## 7. SUMMARY AND CONCLUSIONS

The SST\_CCI long-term products have been validated against both independent and pseudo-independent reference data. The SST\_CCI long-term L4 analysis product has also been inter-compared to other long-term datasets using an implementation of the GMPE. Methods to validate both the SSTs and their associated uncertainty were implemented.

The following conclusions are drawn for:

- SST\_CCI long-term L2P AVHRR
  - Regional biases of several 10ths are calculated
  - Day time data generally cooler than night
  - Strong degree of consistency between later sensors; earlier sensors markedly more variable
  - Consistent “cold tail” on histograms, which may indicate CLAVR-x cloud detection failures
  - Night time uncertainties are very good – better than expected; slightly over estimated and less discriminating in day time
- SST\_CCI long-term L3U ATSR
  - Regional biases of few 10ths (generally somewhat larger than ARC, except for ATSR-1)
  - Day time data generally warmer than night
  - Discrepancy between day and night coverage larger than expected (indicates issue with day time cloud mask)
  - Uncertainties generally over estimated and less discriminating than expected
- SST\_CCI long-term L4 analysis
  - Skewed towards day time (c.f. day/night coverage in L2P/L3U)
  - Improvement over OSTIA reanalysis V1; stats comparable to GMPE median
  - Feature resolution highest in GMPE comparison
  - Uncertainties realistic and mostly highly verified

In addition the product validation and inter-comparison activities highlighted:

- Results distorted if time and depth differences not accounted for
- Notable changes in quality and sampling distribution of reference data (particularly during ATSR-2 period)
- Coverage of GTMBA in ICOADS not sufficient for stability assessment

- Depth time should be explicit in products

**It is recommended that all V1 SST\_CCI products should be released for evaluation by the user community once the PVIR and CAR are accepted.**

## APPENDIX A DETAILED AVHRR PRODUCT VALIDATION RESULTS

The following section contains the detailed validation results for the SST\_CCI long-term ECV AVHRR products. For each sensor we provide:

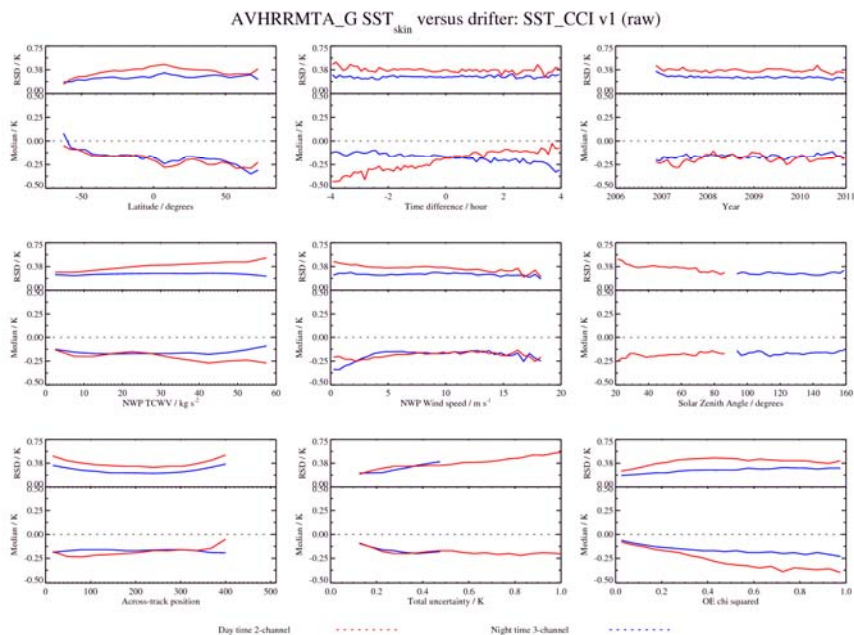
- Dependence plots of median and robust standard deviation of the discrepancy between the satellite and drifting buoys for
  - Satellite SST<sub>skin</sub> versus drifter SST<sub>depth</sub>.
  - Satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub>.
  - Satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub> with additional adjustments for the difference between the satellite and drifter measurement times (satellite at 10:30 am/pm local solar time) from a combined diurnal variability/skin effect model.

Dependences are provided for latitude, time difference between satellite and drifter measurements, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and the retrieval chi squared function.

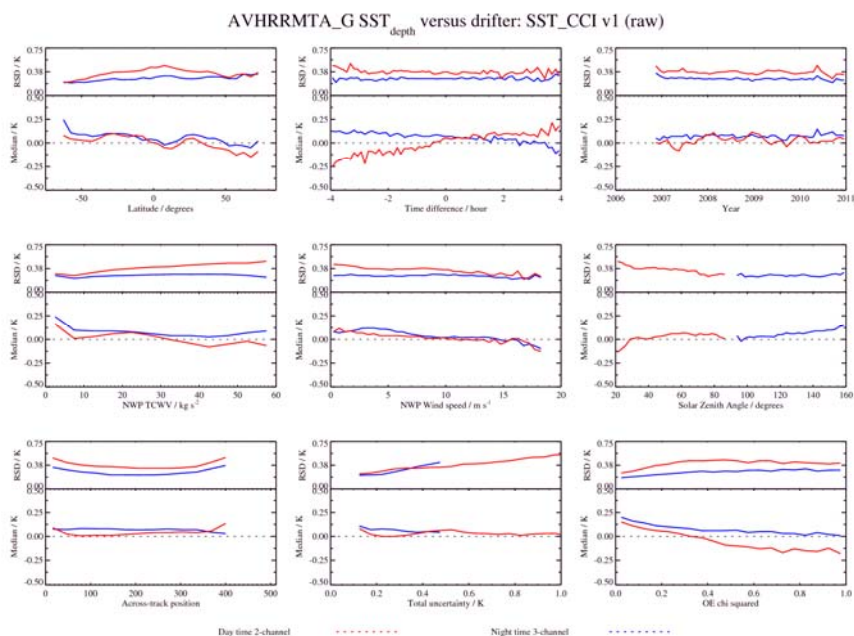
Note: A minimum of 30 match-ups is required for each point on the dependence plots (from central limit theorem). As such, the minimum standard error for a standard deviation of 0.5 K would be roughly 0.1 K.

- Spatial maps and Hovmoller plots of the median discrepancy between the satellite and drifting buoys for the same three comparisons as for the dependence plots.
- Histograms of the distributions of median discrepancies between the satellite and drifting buoys for satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub> with additional adjustments for the difference between the satellite and drifter measurement times.
- Uncertainty validation plots for the total uncertainty applicable to the satellite SST<sub>depth</sub> as a function of the median discrepancies between the satellite and drifting buoys for satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub> with additional adjustments for the difference between the satellite and drifter measurement times. For further details of the uncertainty validation methodologies please see Section 5.
- A table of the median and robust standard deviation of the discrepancy between the satellite products and the various reference datasets for a selection of comparisons.
- A summary of the key findings for each sensor.

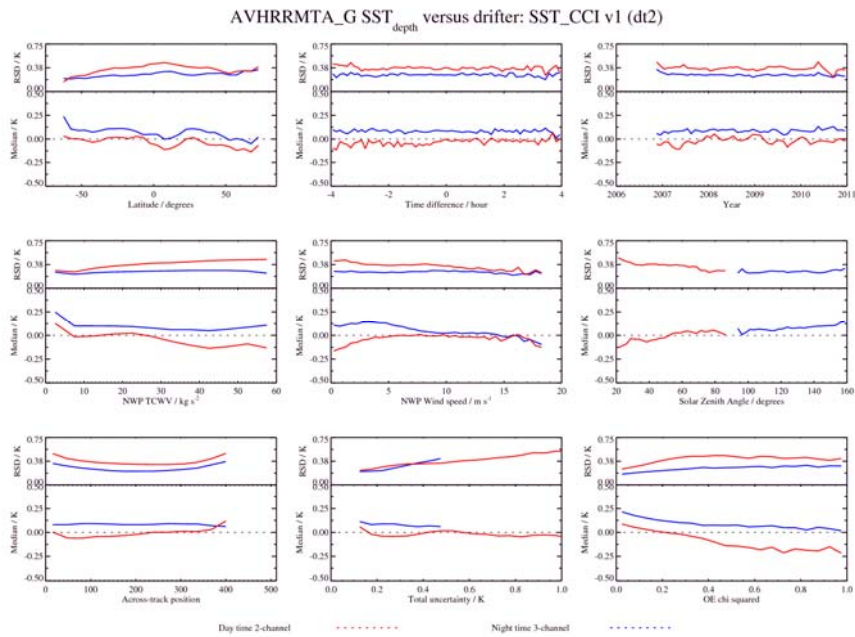
### A.1 AVHRR MTA



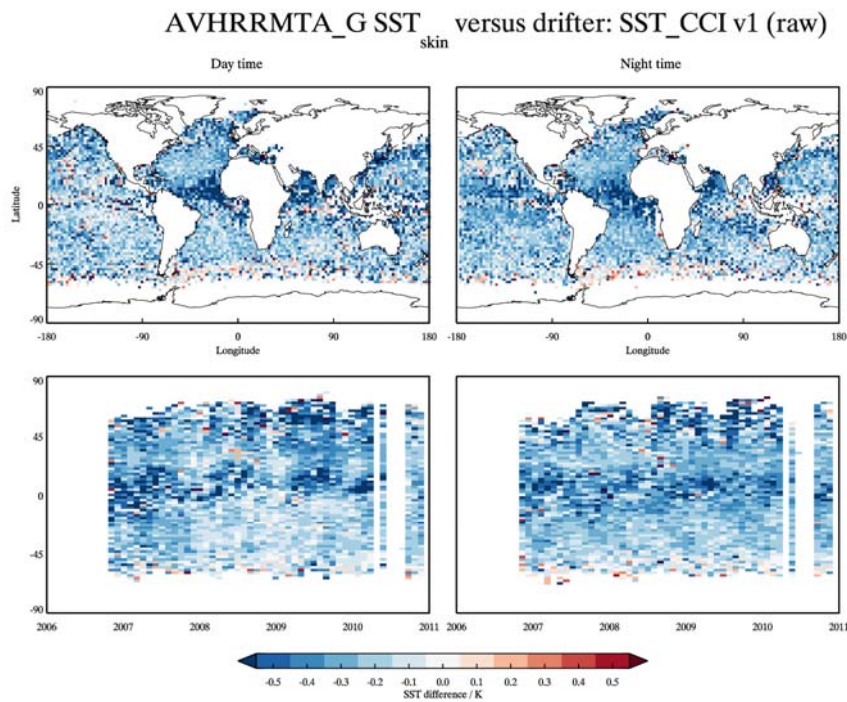
**Figure 7-1:** Dependence of the median and robust standard deviation between AVHRR-MTA SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-2:** Dependence of the median and robust standard deviation between AVHRR-MTA SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

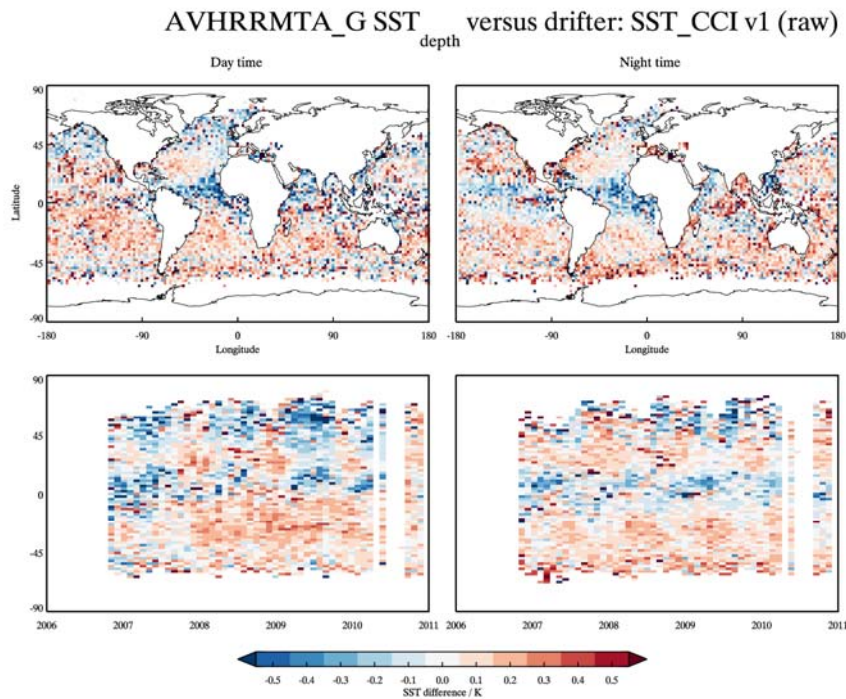


**Figure 7-3:** Dependence of the median and robust standard deviation between AVHRR-MTA SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

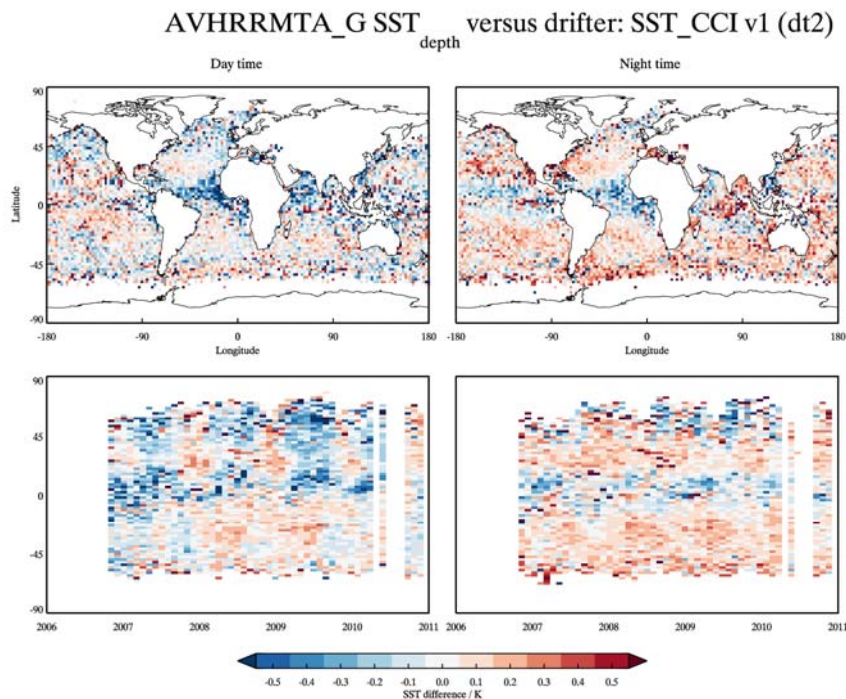


**Figure 7-4:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-MTA SST<sub>skin</sub> and drifter SST<sub>depth</sub>.



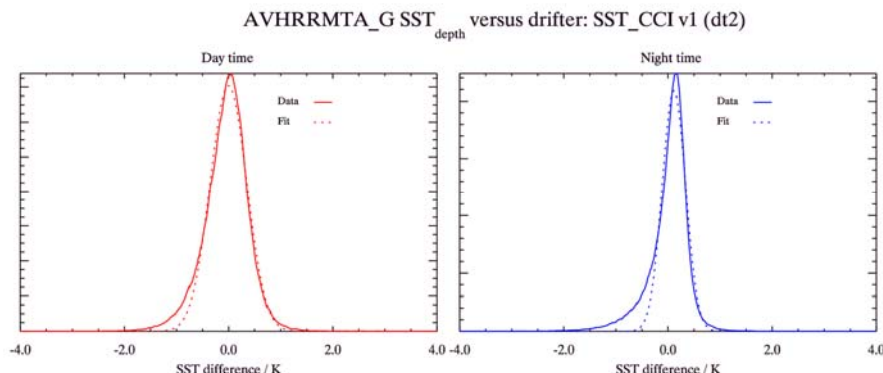


**Figure 7-5:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-MTA SST<sub>depth</sub> and drifter SST<sub>depth</sub>.

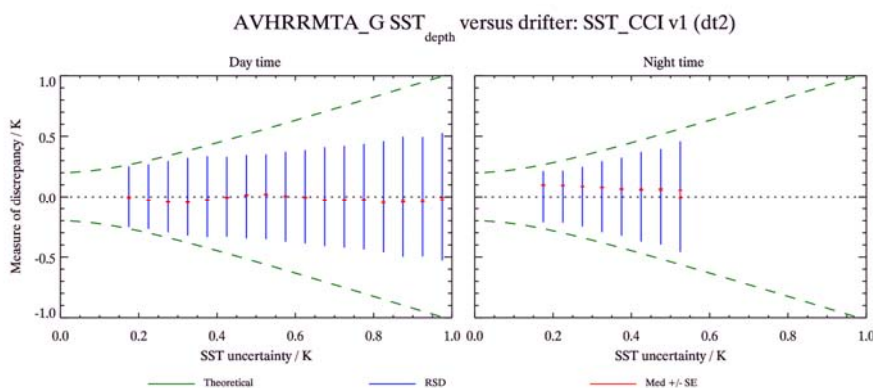


**Figure 7-6:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-MTA SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).





**Figure 7-7:** Histograms of the median discrepancy between AVHRR-MTA SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-8:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-MTA SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.

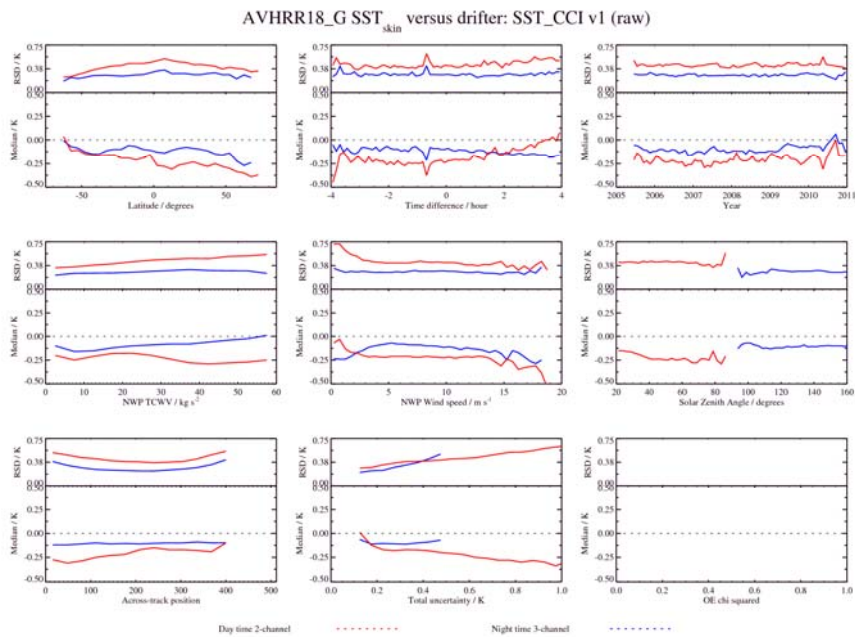
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	130150	-0.02	0.37
	<i>Night</i>	133619	+0.08	0.26
<b>iDrifters</b>	<i>Day</i>	12941	-0.02	0.36
	<i>Night</i>	13305	+0.08	0.26
<b>GT MBA</b>	<i>Day</i>	2073	-0.05	0.41
	<i>Night</i>	2626	+0.04	0.27
<b>Argo</b>	<i>Day</i>	1049	-0.04	0.37
	<i>Night</i>	680	+0.05	0.26
<b>Radiometers</b>	<i>Day</i>	16	-0.03	0.48
	<i>Night</i>	22	+0.01	0.30

**Table 7-1:** Global validation statistics from comparing SST CCI AVHRR-MTA to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GT MBA and Argo; for radiometers only the time difference has been adjusted.

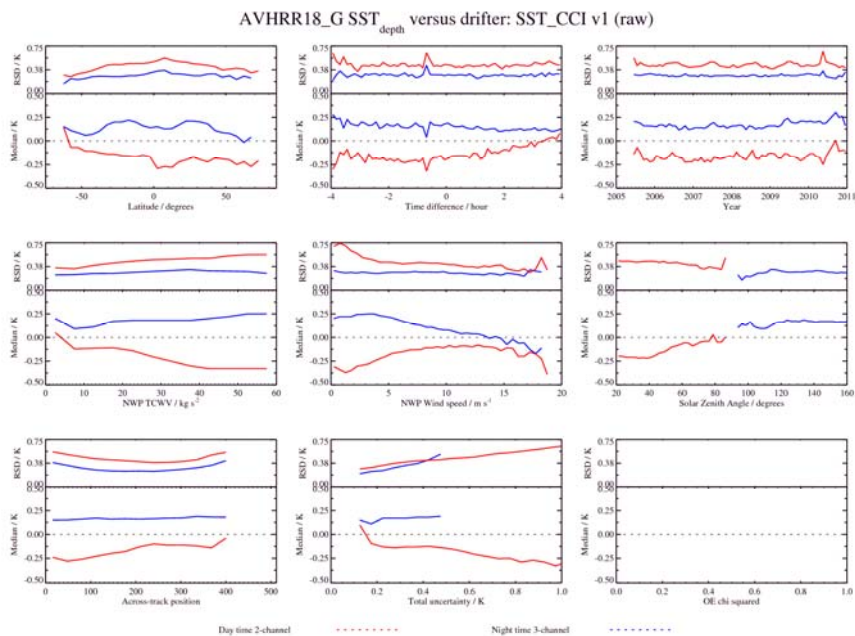
Summary of key findings from AVHRR-MTA validation:

- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel
- Residual bias at low wind speed (stronger in 3- channel)
- Strong regional variations (cooler in Arctic and Southern Oceans)
- Evidence of desert dust effects
- Residual cloud contamination (stronger at night)
- Uncertainty estimates reasonable; better discrimination at night.

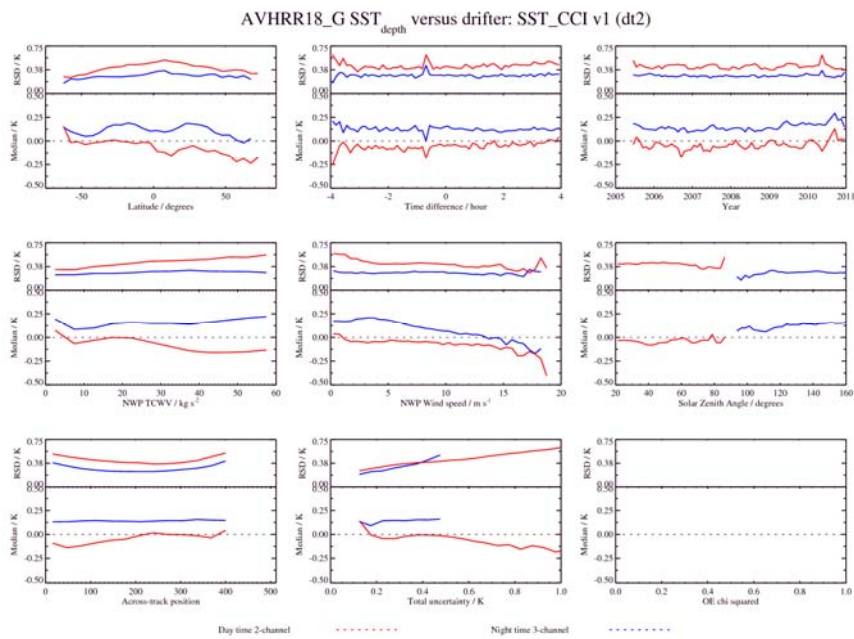
## A.2 AVHRR 18



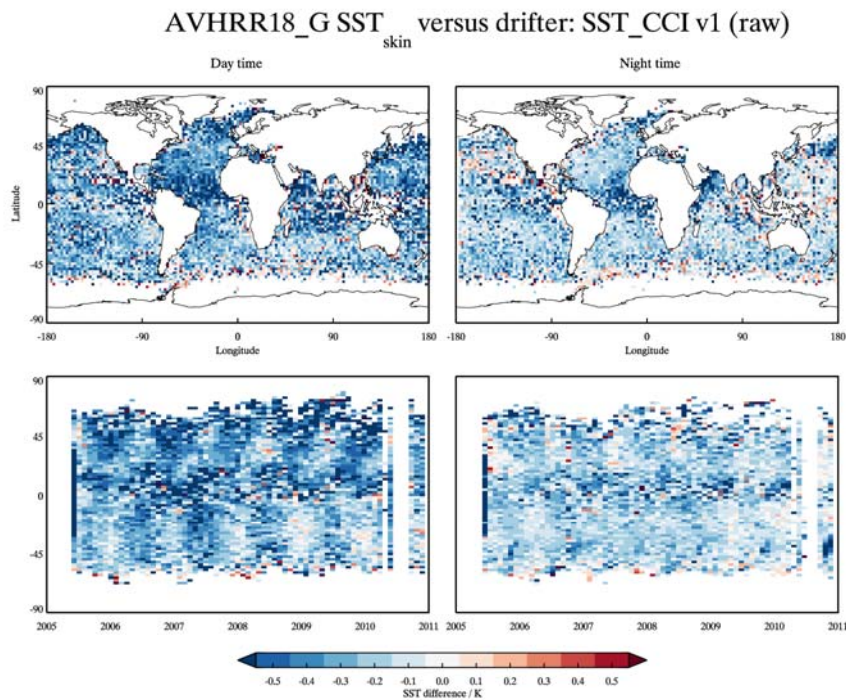
**Figure 7-9:** Dependence of the median and robust standard deviation between AVHRR-18 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-10:** Dependence of the median and robust standard deviation between AVHRR-18 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

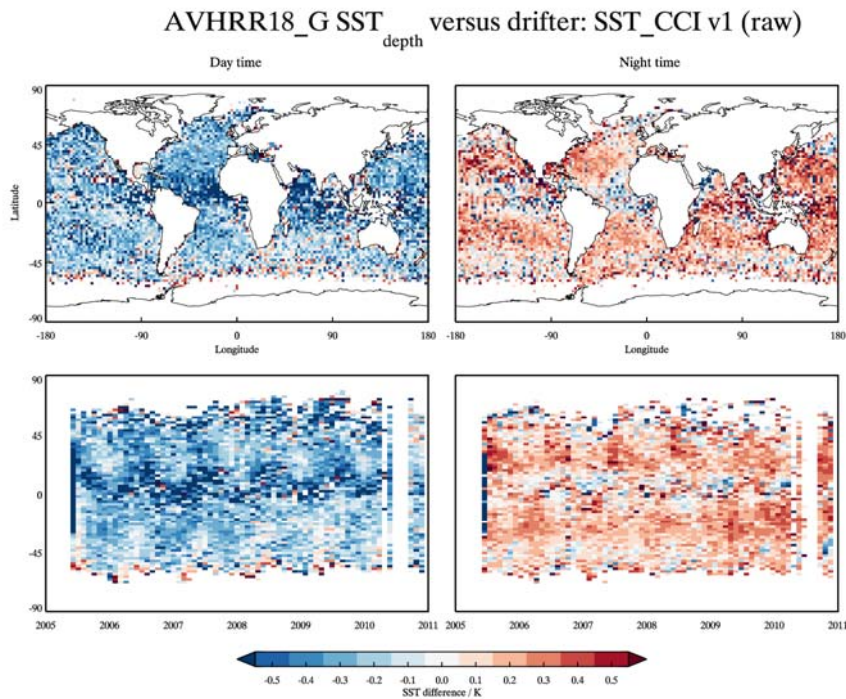


**Figure 7-11:** Dependence of the median and robust standard deviation between AVHRR-18 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

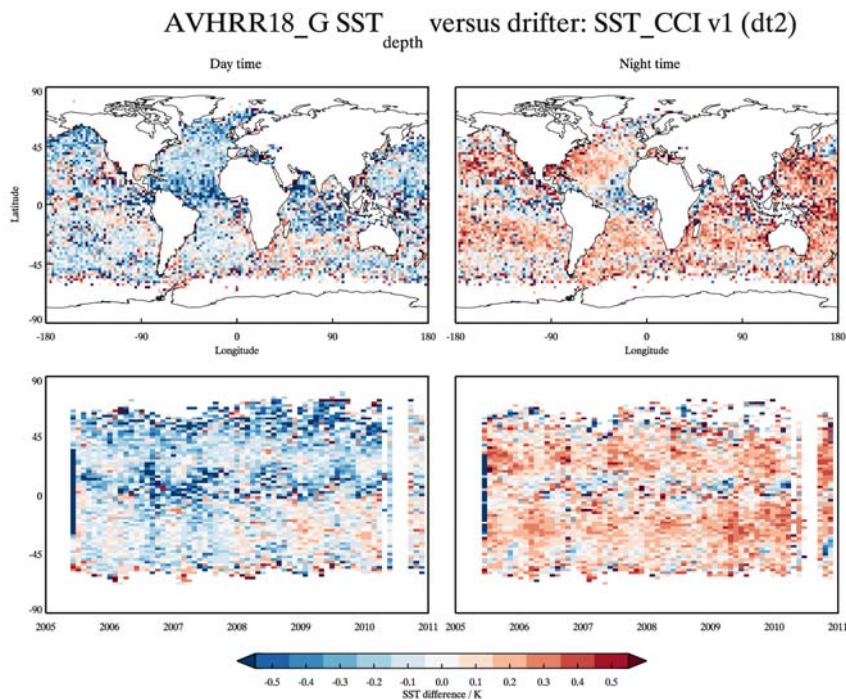


**Figure 7-12:** Spatial distribution and Hovmöller plot of the Dependence of the median discrepancy between AVHRR-18 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.

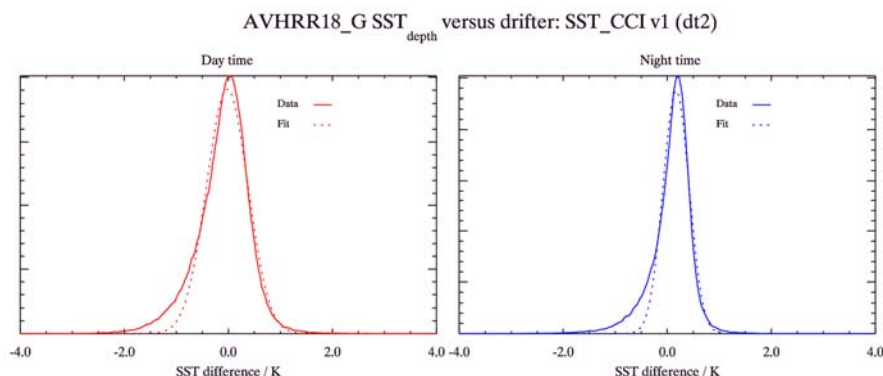




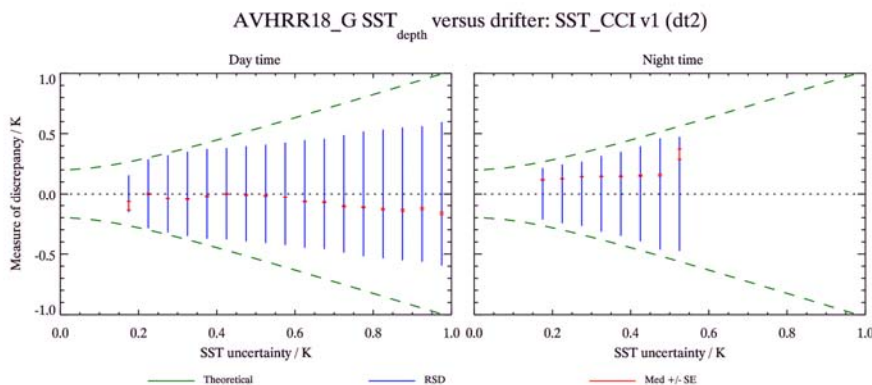
**Figure 7-13:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-18 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.



**Figure 7-14:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-18 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-15:** Histograms of the median discrepancy between AVHRR-18 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-16:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-18 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.



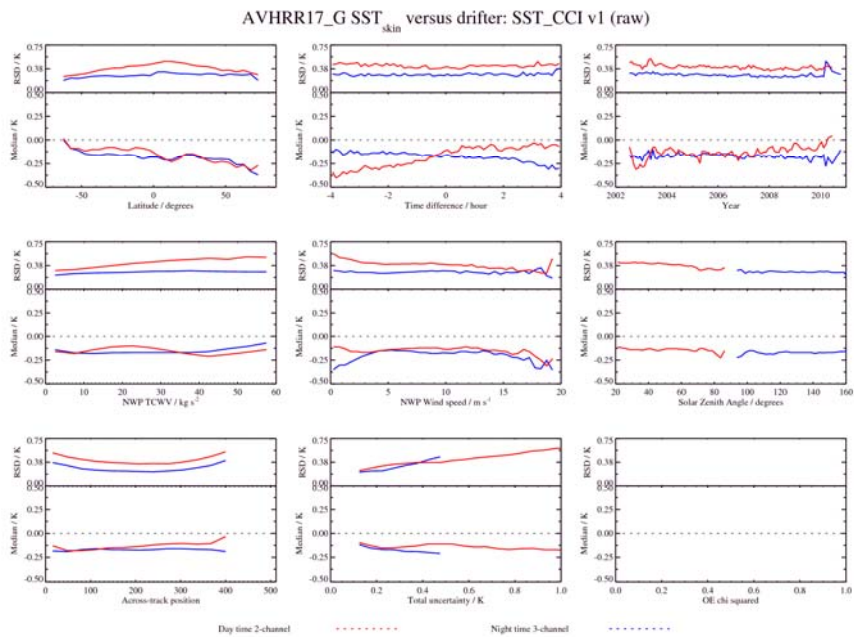
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	203490	-0.05	0.43
	<i>Night</i>	167294	+0.14	0.29
<b>iDrifters</b>	<i>Day</i>	20416	-0.04	0.42
	<i>Night</i>	16700	+0.14	0.29
<b>GTMBA</b>	<i>Day</i>	579	-0.06	0.54
	<i>Night</i>	524	-0.17	0.24
<b>Argo</b>	<i>Day</i>	908	-0.05	0.42
	<i>Night</i>	518	+0.07	0.32
<b>Radiometers</b>	<i>Day</i>	38	-0.06	0.54
	<i>Night</i>	12	-0.16	0.24

**Table 7-2:** Global validation statistics from comparing SST CCI AVHRR-18 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GTMBA and Argo; for radiometers only the time difference has been adjusted.

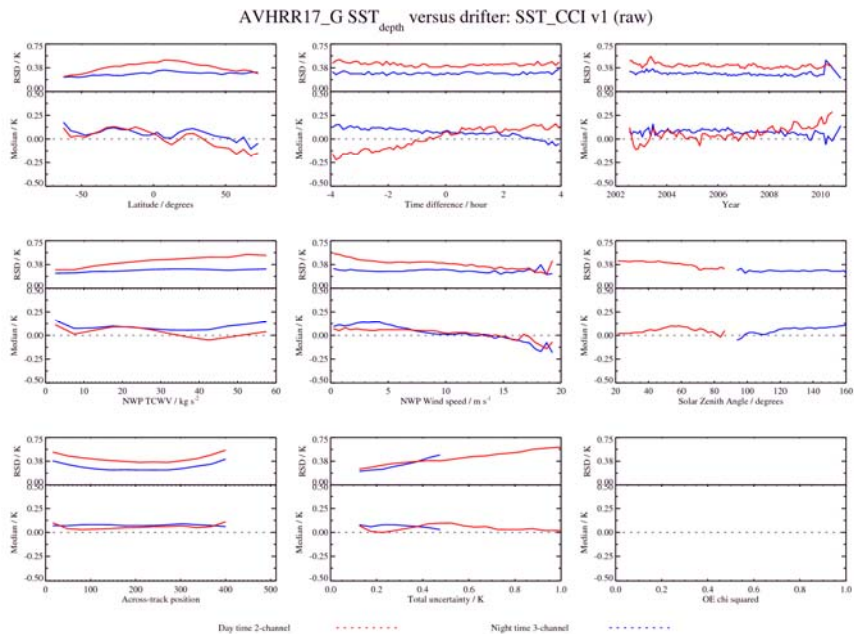
Summary of key findings from AVHRR-18 validation:

- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel (larger than seen for MTA)
- Residual bias wind speed bias notably in 3-channel
- Some regional variations (cooler in Arctic and Southern Oceans)
- Residual TCWV dependences in 2-channel
- Evidence of desert dust effect on data
- Residual cloud contamination (stronger at night)
- Uncertainty estimates reasonable; better discrimination at night.

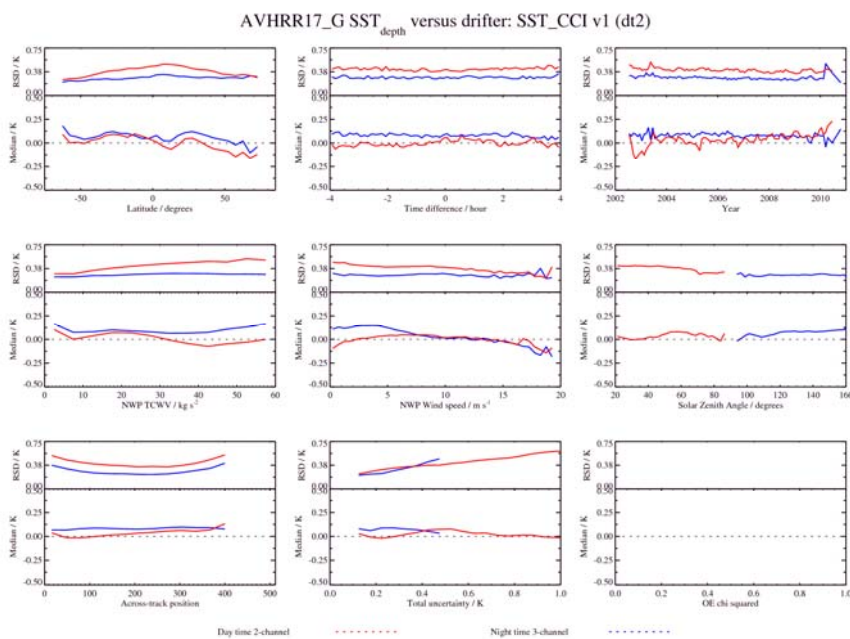
### A.3 AVHRR 17



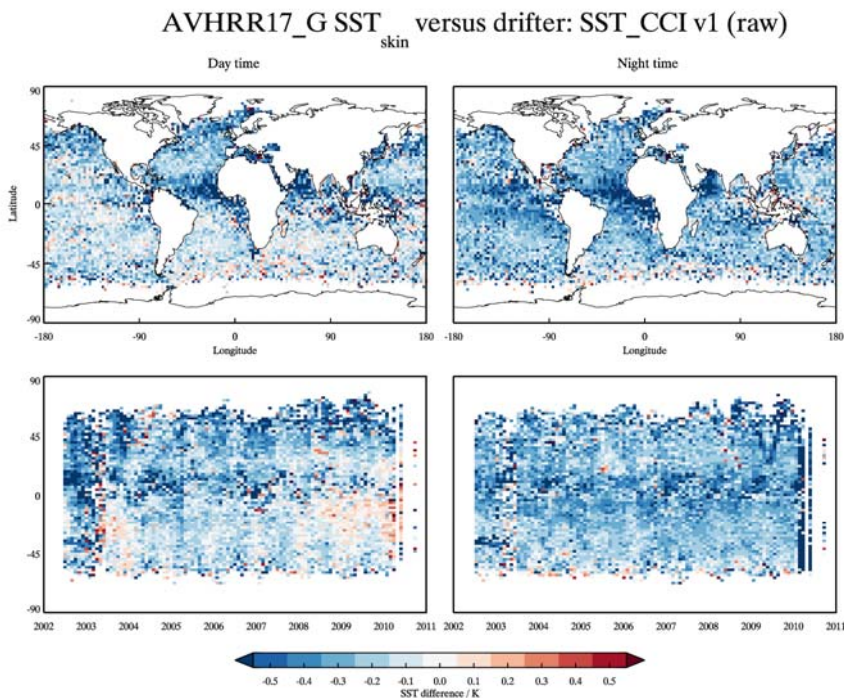
**Figure 7-17:** Dependence of the median and robust standard deviation between AVHRR-17 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-18:** Dependence of the median and robust standard deviation between AVHRR-17 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

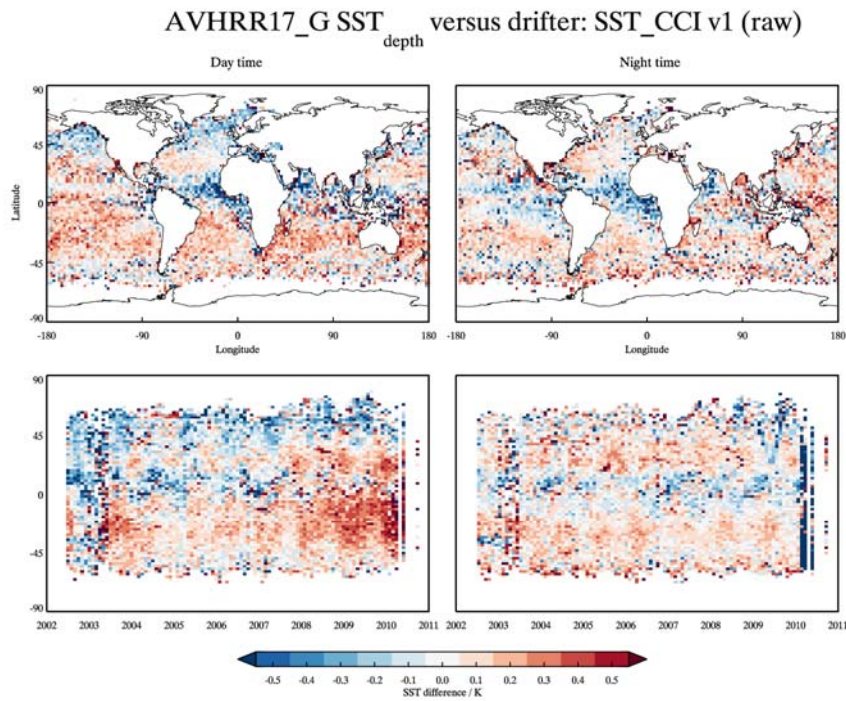


**Figure 7-19:** Dependence of the median and robust standard deviation between AVHRR-17 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

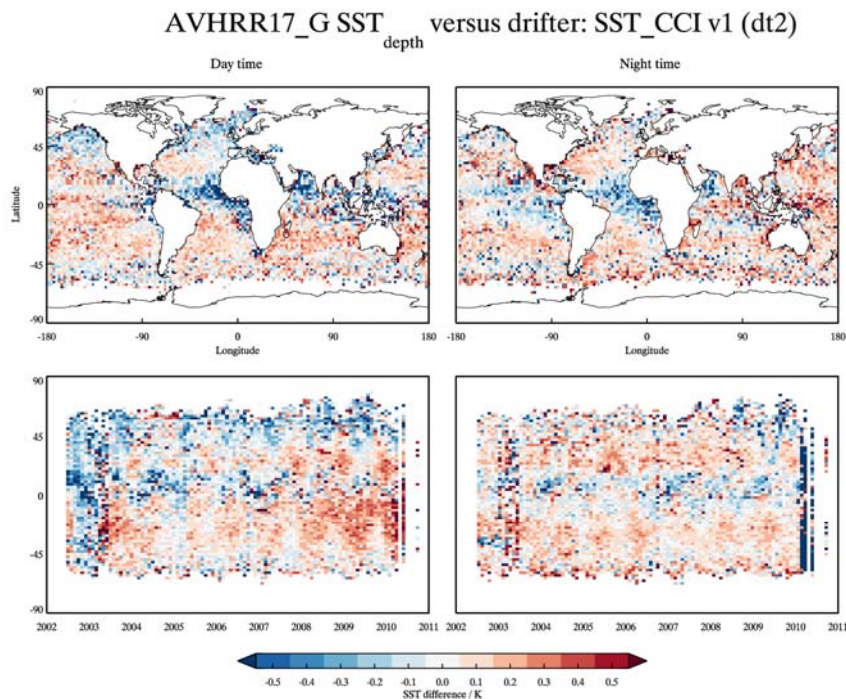


**Figure 7-20:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-17 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.

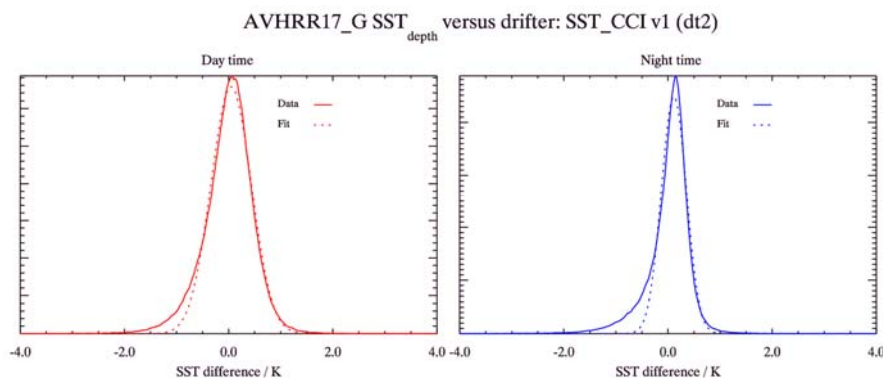




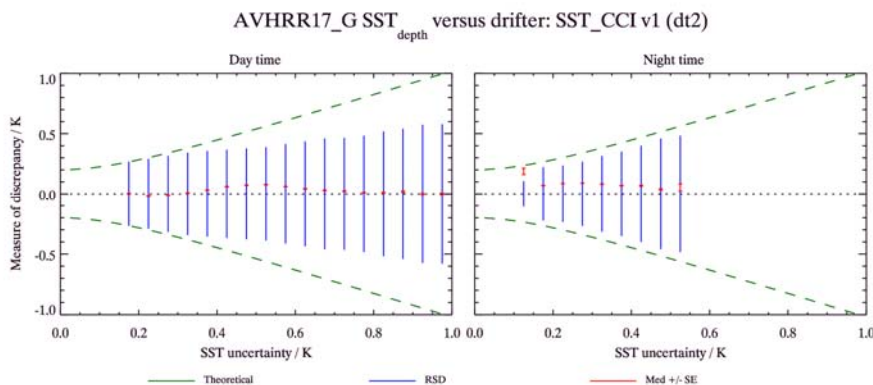
**Figure 7-21:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-17 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.



**Figure 7-22:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-17 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-23:** Histograms of the median discrepancy between AVHRR-17 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-24:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-17 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.



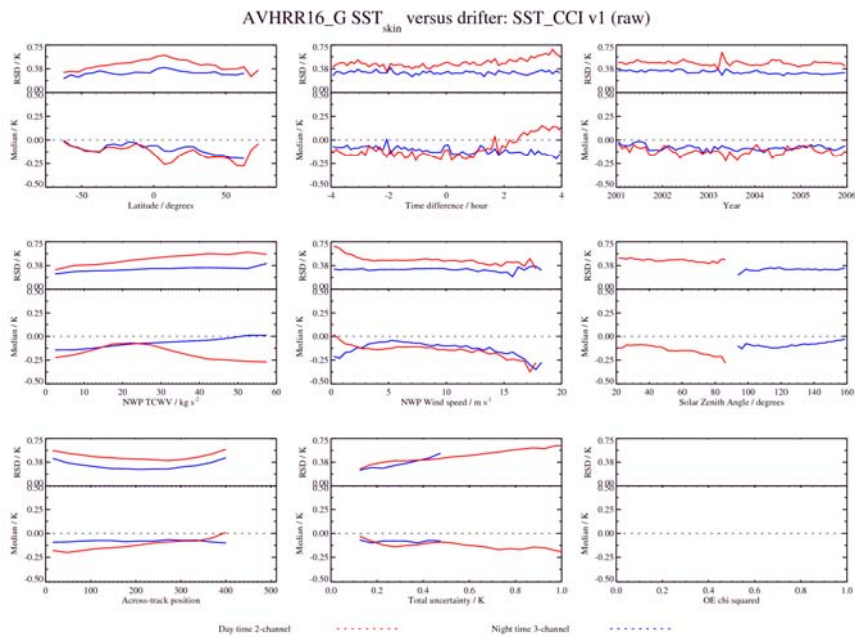
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	335985	+0.03	0.40
	<i>Night</i>	305926	+0.08	0.28
<b>iDrifters</b>	<i>Day</i>	27779	+0.04	0.40
	<i>Night</i>	25529	+0.08	0.28
<b>GT MBA</b>	<i>Day</i>	3766	+0.01	0.45
	<i>Night</i>	2879	+0.04	0.28
<b>Argo</b>	<i>Day</i>	1966	+0.01	0.41
	<i>Night</i>	1328	+0.06	0.26
<b>Radiometers</b>	<i>Day</i>	81	-0.11	0.43
	<i>Night</i>	105	-0.05	0.37

**Table 7-3:** Global validation statistics from comparing SST CCI AVHRR-17 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GT MBA and Argo; for radiometers only the time difference has been adjusted.

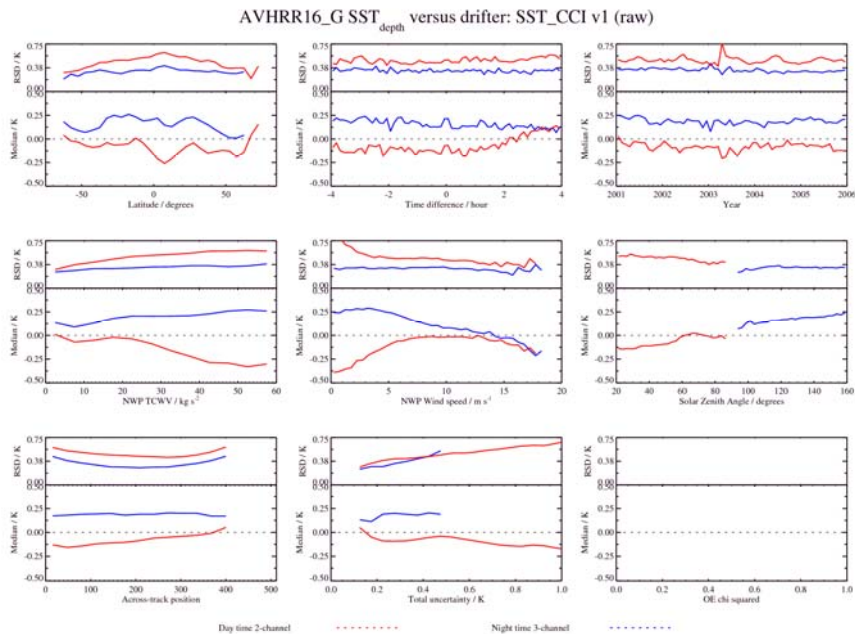
Summary of key findings from AVHRR-17 validation:

- Results similar to AVHRR-18 (both morning orbits)
- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel
- Residual bias at low wind speed (stronger in 3- channel)
- Strong regional variations (cooler in Arctic and Southern Oceans)
- Evidence of desert dust effects
- Residual cloud contamination (stronger at night)
- Uncertainty estimates reasonable; better discrimination at night.
- Multi-year trend in bias of day-time results

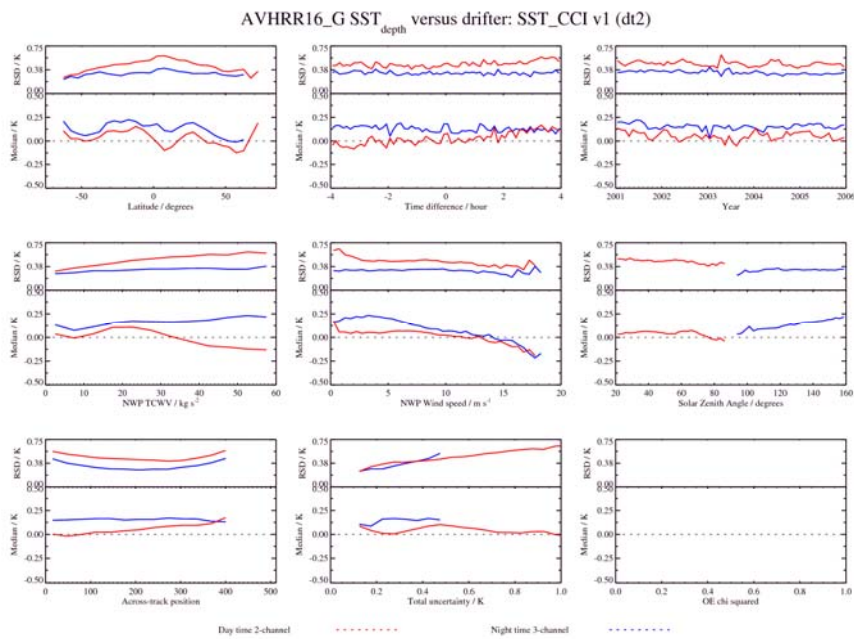
## A.4 AVHRR 16



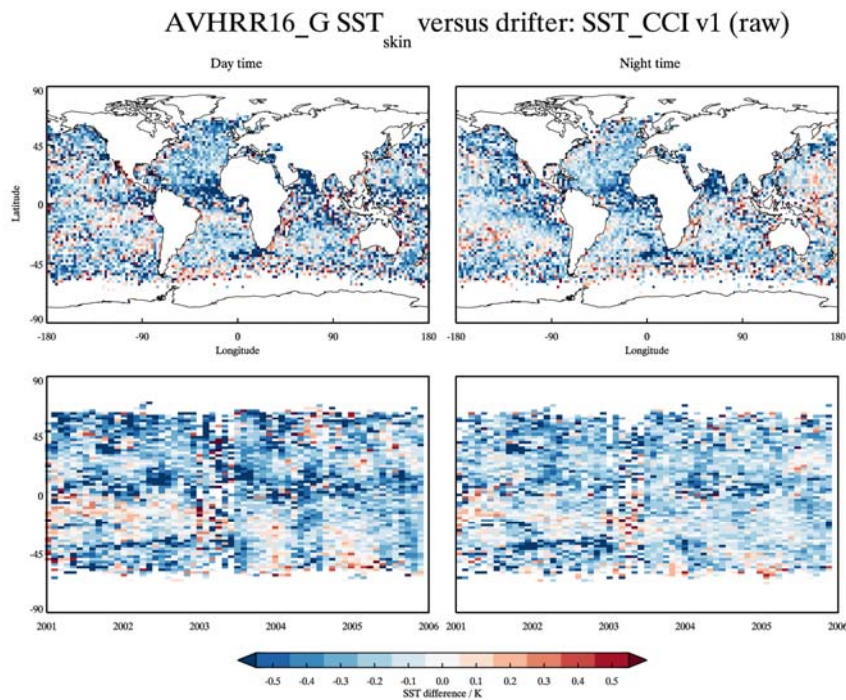
**Figure 7-25:** Dependence of the median and robust standard deviation between AVHRR-16 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-26:** Dependence of the median and robust standard deviation between AVHRR-16 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

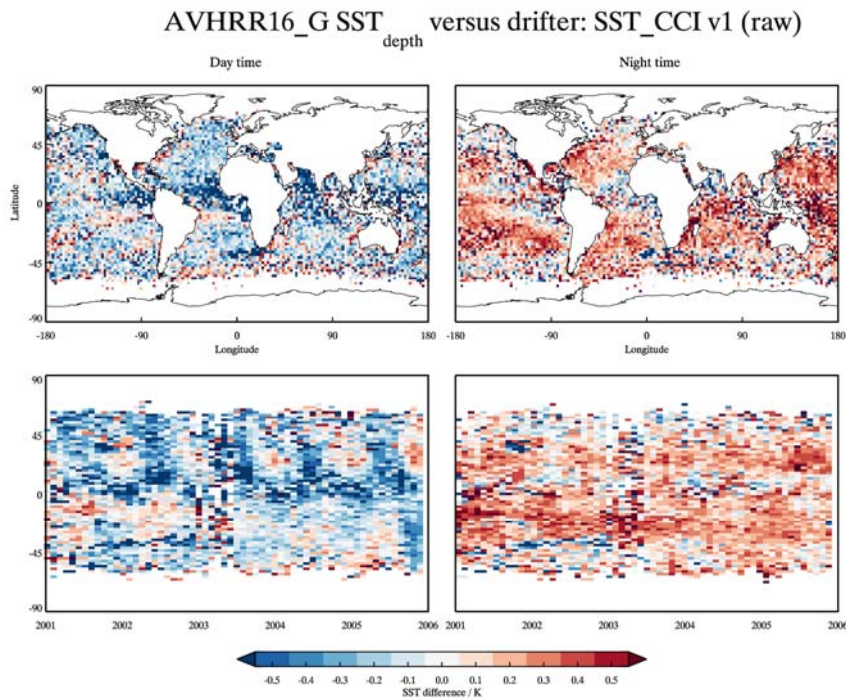


**Figure 7-27:** Dependence of the median and robust standard deviation between AVHRR-16 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

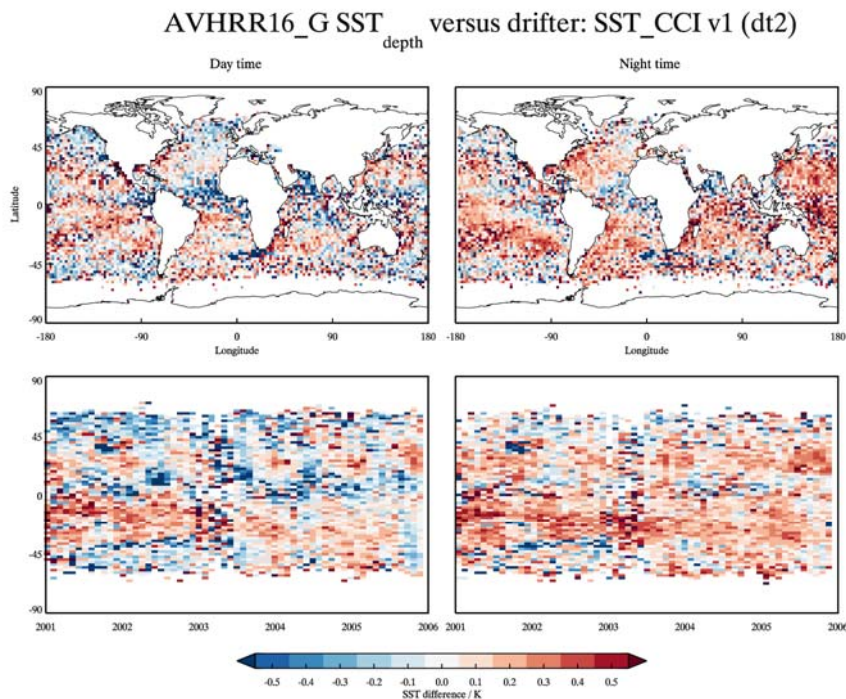


**Figure 7-28:** Spatial distribution and Hovmöller plot of the Dependence of the median discrepancy between AVHRR-16 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.

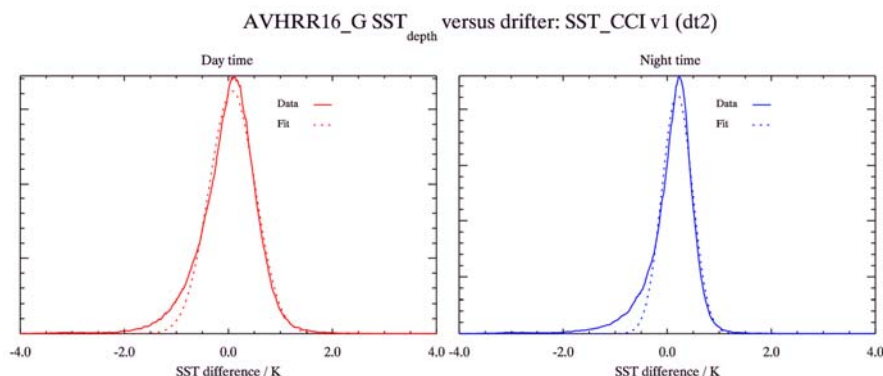




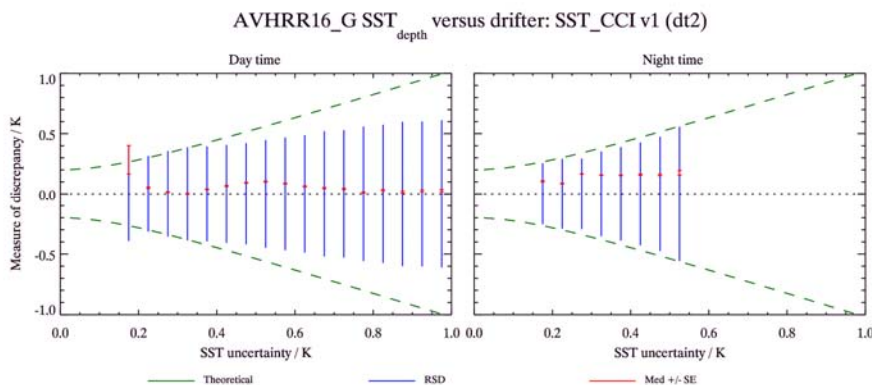
**Figure 7-29:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-16 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.



**Figure 7-30:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-16 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-31:** Histograms of the median discrepancy between AVHRR-16 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-32:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-16 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.



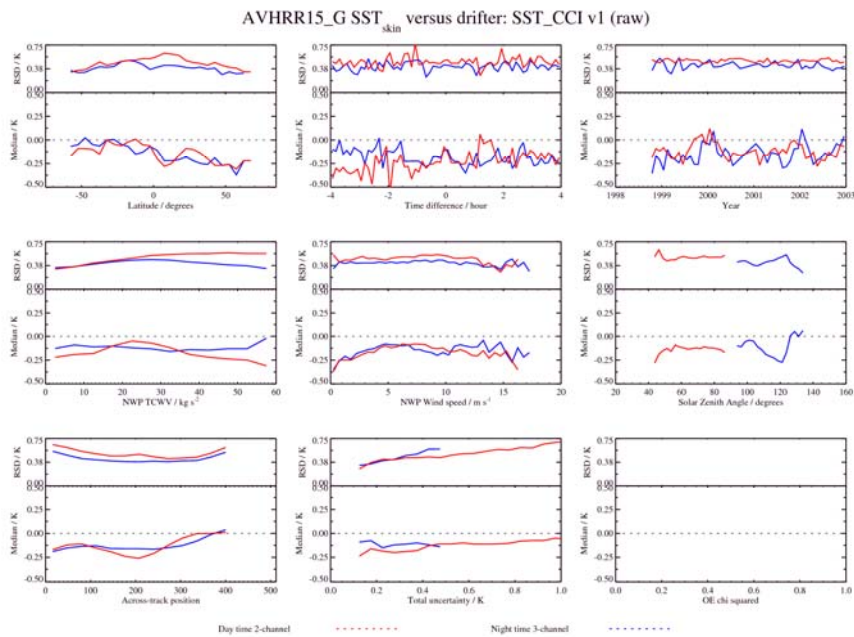
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	189813	+0.05	0.47
	<i>Night</i>	172481	+0.16	0.33
<b>iDrifters</b>	<i>Day</i>	8816	+0.06	0.46
	<i>Night</i>	7848	+0.17	0.31
<b>GTMBA</b>	<i>Day</i>	1146	+0.05	0.49
	<i>Night</i>	125	+0.13	0.38
<b>Argo</b>	<i>Day</i>	336	+0.02	0.49
	<i>Night</i>	294	+0.10	0.34
<b>Radiometers</b>	<i>Day</i>	38	+0.20	0.43
	<i>Night</i>	34	+0.10	0.30

**Table 7-4:** Global validation statistics from comparing SST CCI AVHRR-16 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GTMBA and Argo; for radiometers only the time difference has been adjusted.

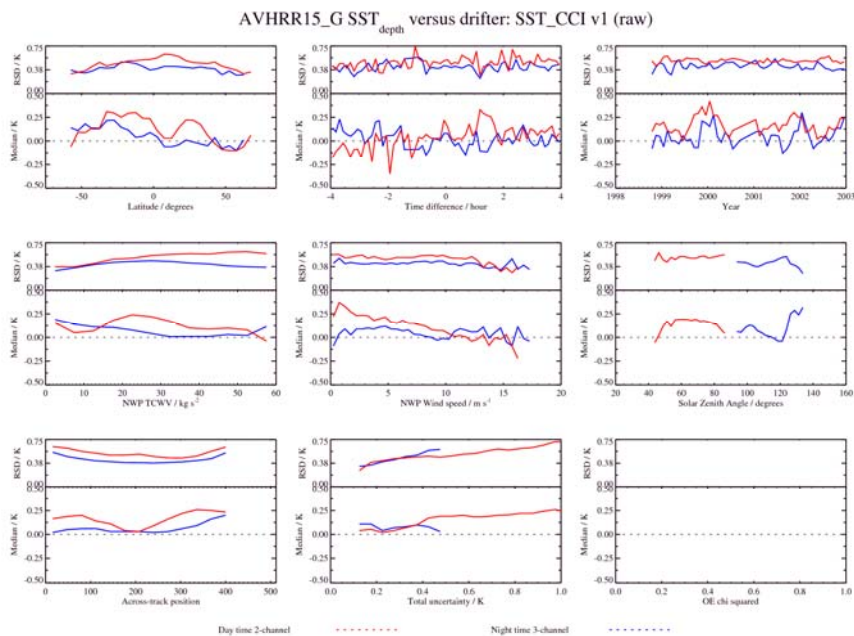
Summary of key findings from AVHRR-16 validation:

- Results similar to AVHRR-18 (both afternoon orbits)
- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel (larger than seen for AVHRR-18)
- Residual bias wind speed bias notably in 3-channel
- Residual TCWV dependence in 2-channel (larger than seen for AVHRR-18)
- Some regional variations (cooler in Arctic and Southern Oceans)
- Evidence of desert dust effect on data
- Residual cloud contamination (stronger at night)
- Uncertainty estimates reasonable; better discrimination at night.

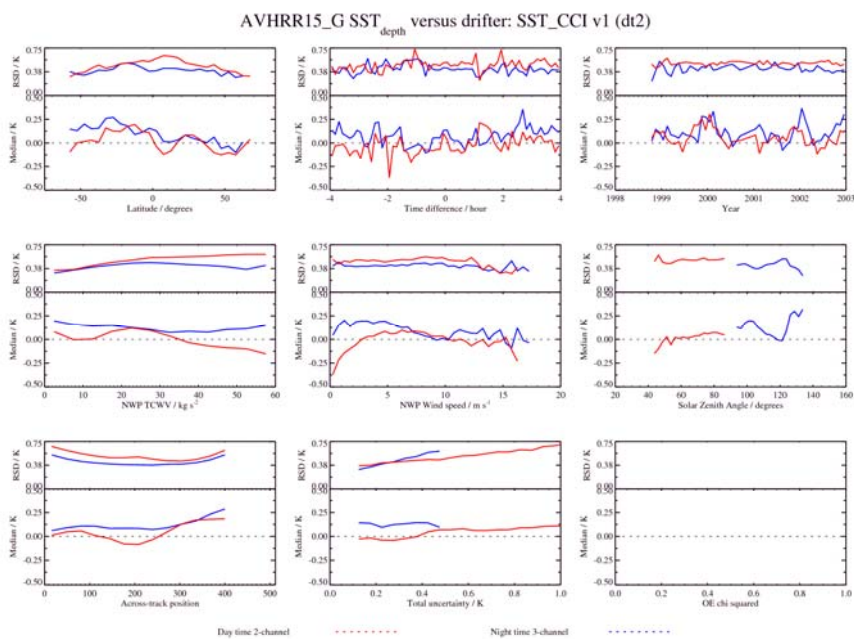
## A.5 AVHRR 15



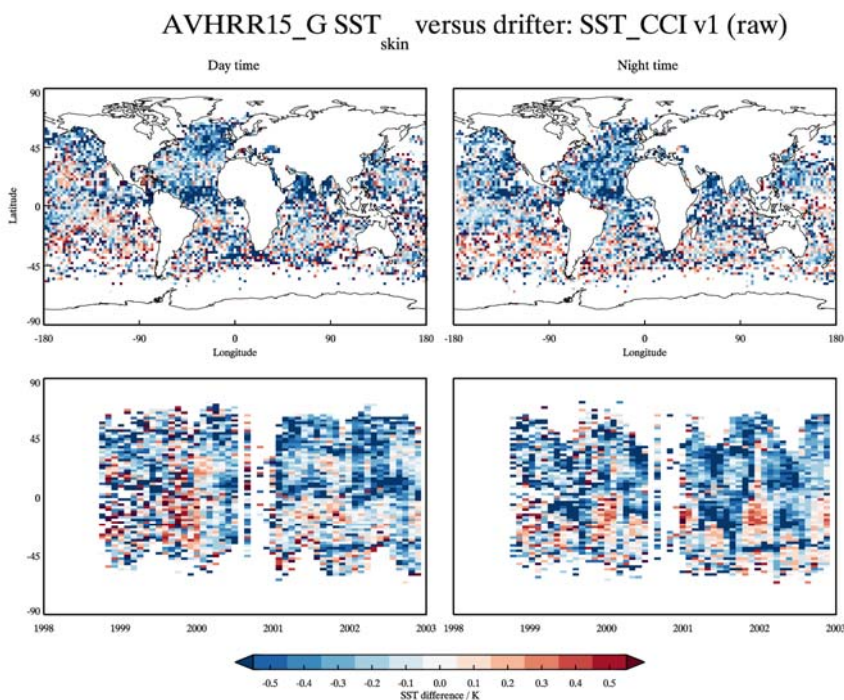
**Figure 7-33:** Dependence of the median and robust standard deviation between AVHRR-15 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-34:** Dependence of the median and robust standard deviation between AVHRR-15 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

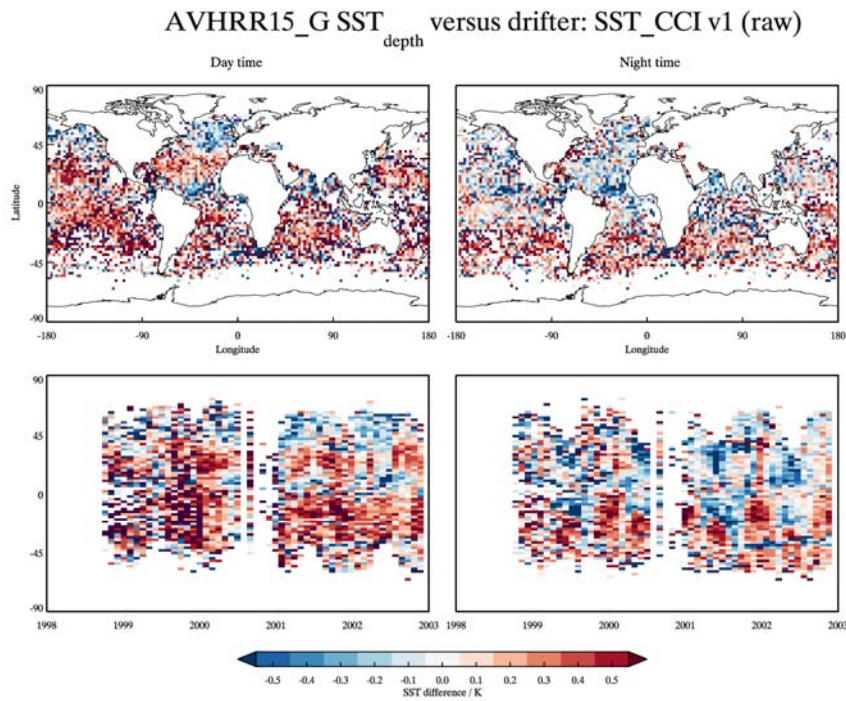


**Figure 7-35:** Dependence of the median and robust standard deviation between AVHRR-15 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

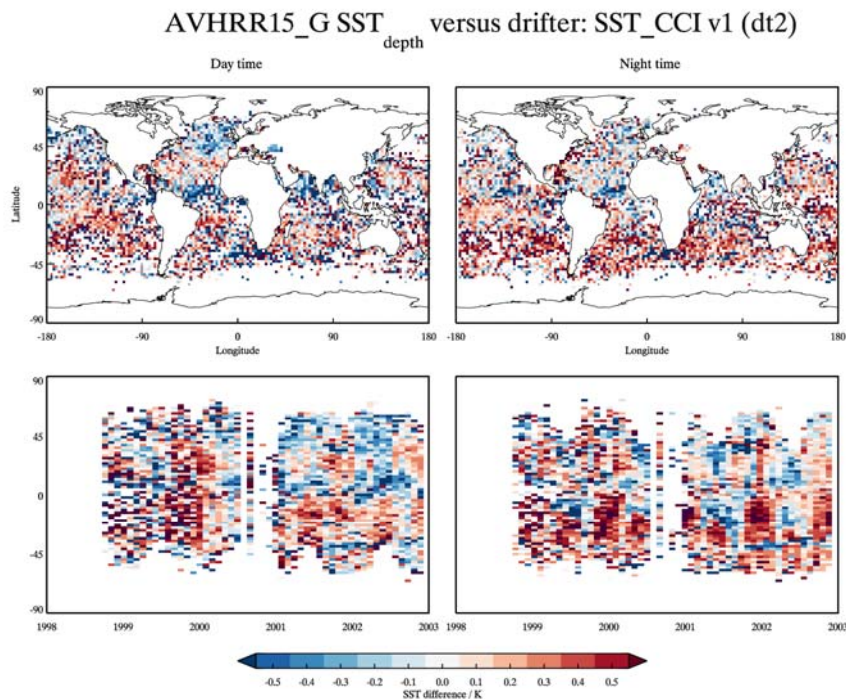


**Figure 7-36:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-15 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.

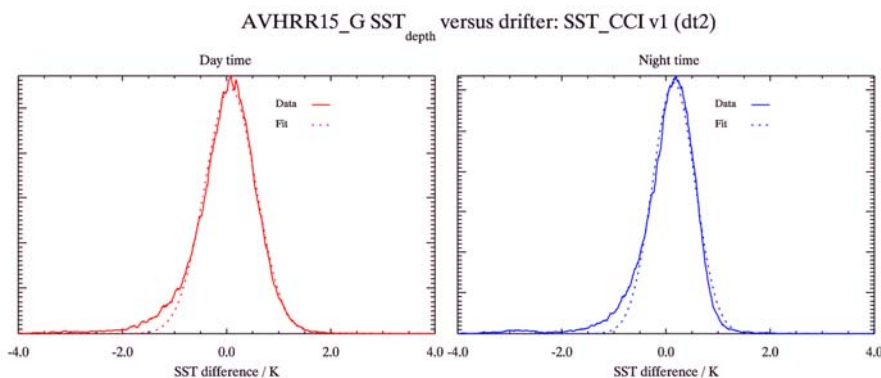




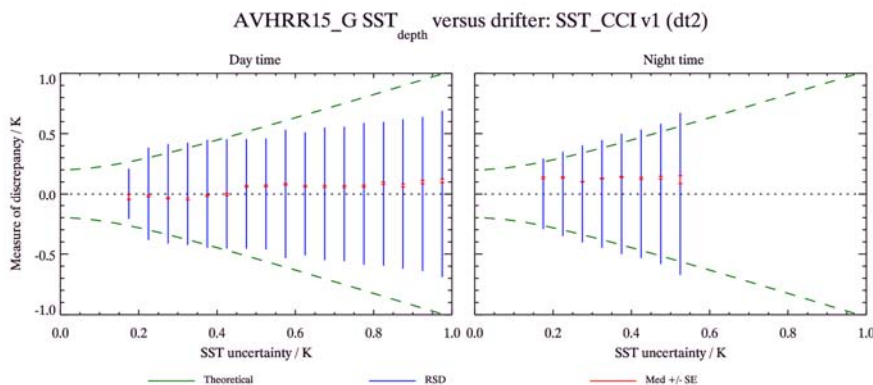
**Figure 7-37:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-15 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.



**Figure 7-38:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-15 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-39:** Histograms of the median discrepancy between AVHRR-15 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-40:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-15 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.



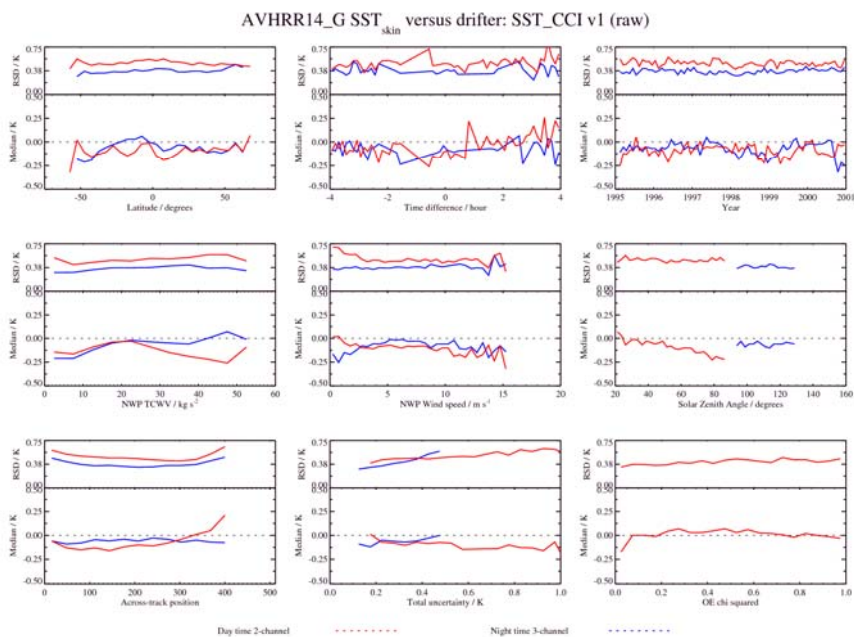
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	70214	+0.04	0.52
	<i>Night</i>	66446	+0.12	0.43
<b>iDrifters</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
<b>GT MBA</b>	<i>Day</i>	914	+0.03	0.54
	<i>Night</i>	28	+0.23	0.33
<b>Argo</b>	<i>Day</i>	40	-0.09	0.48
	<i>Night</i>	43	-0.08	0.48
<b>Radiometers</b>	<i>Day</i>	27	-0.11	0.41
	<i>Night</i>	37	+0.01	0.35

**Table 7-5:** Global validation statistics from comparing SST CCI AVHRR-15 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GT MBA and Argo; for radiometers only the time difference has been adjusted.

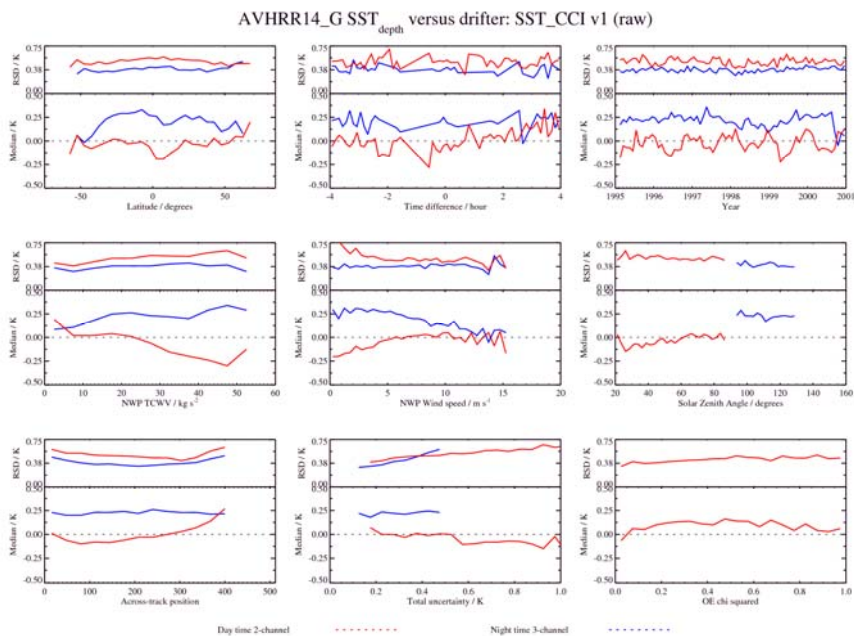
Summary of key findings from AVHRR-15 validation:

- Larger regional biases – not stable in time – change after data gap in 2<sup>nd</sup> ½ of 2000
- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel
- Residual bias wind speed and TCWV dependence in both 2- and 3- channel retrievals
- Some evidence of desert dust effect on data
- Residual cloud contamination not as evident (histograms likely dominated by retrieval uncertainties)
- Uncertainty estimates reasonable – but underestimated at night; better discrimination at night.

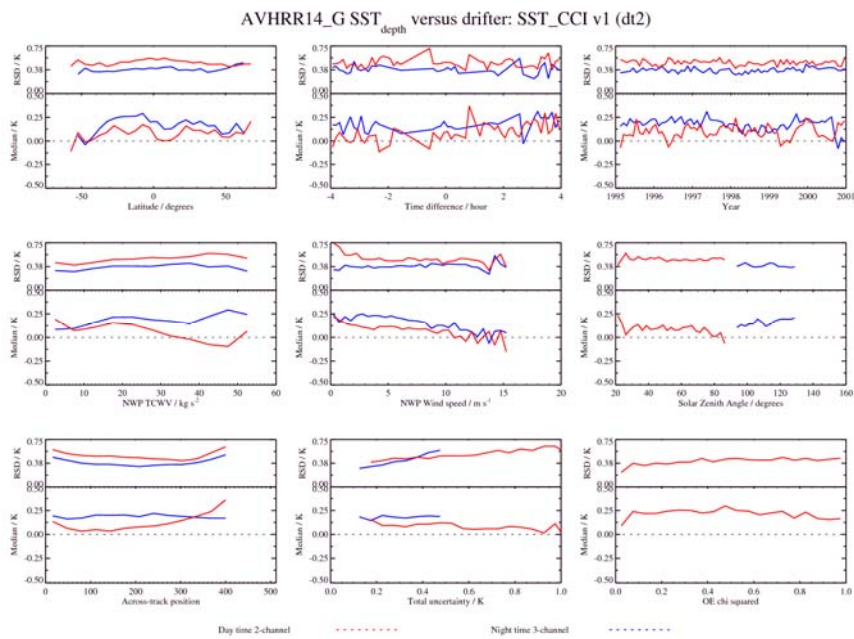
## A.6 AVHRR 14



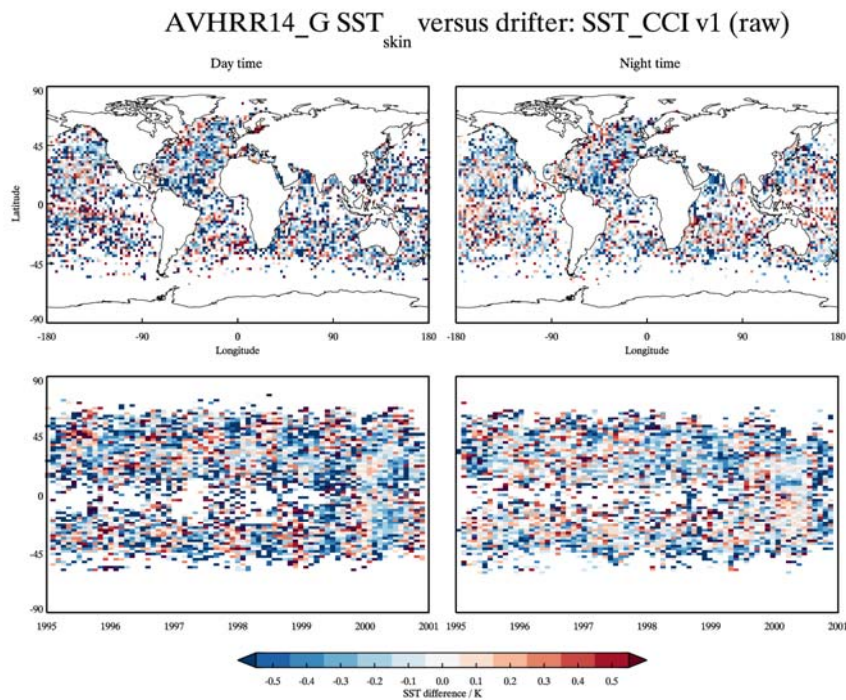
**Figure 7-41:** Dependence of the median and robust standard deviation between AVHRR-14 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-42:** Dependence of the median and robust standard deviation between AVHRR-14 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

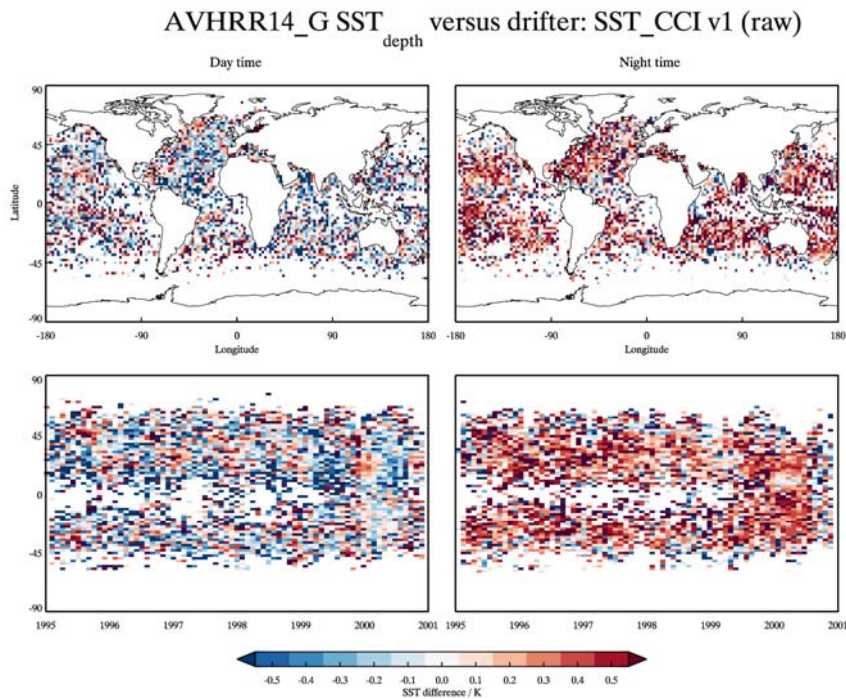


**Figure 7-43:** Dependence of the median and robust standard deviation between AVHRR-14 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

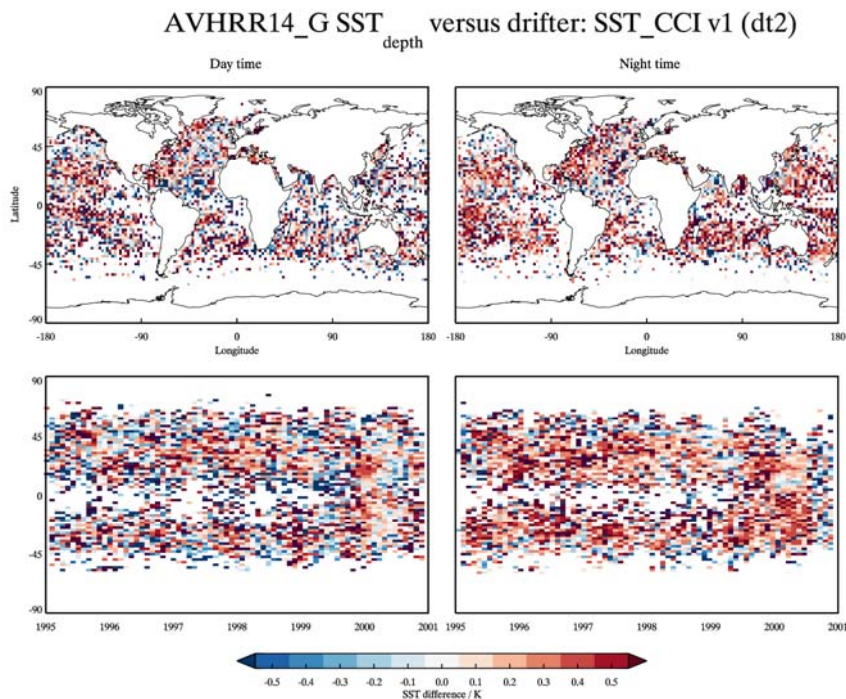


**Figure 7-44:** Spatial distribution and Hovmöller plot of the Dependence of the median discrepancy between AVHRR-14 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.

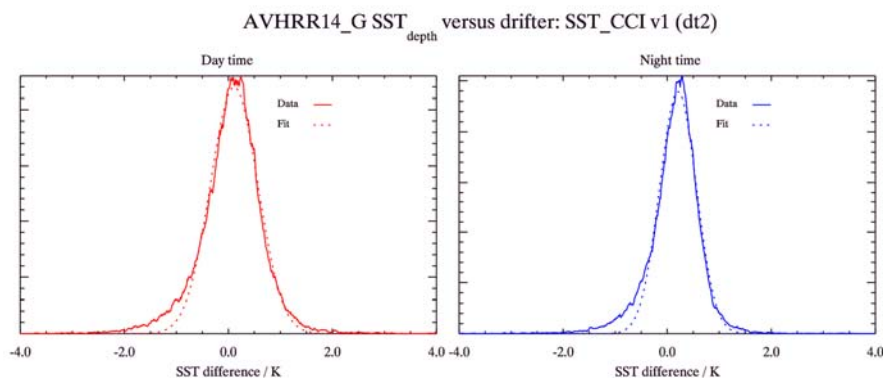




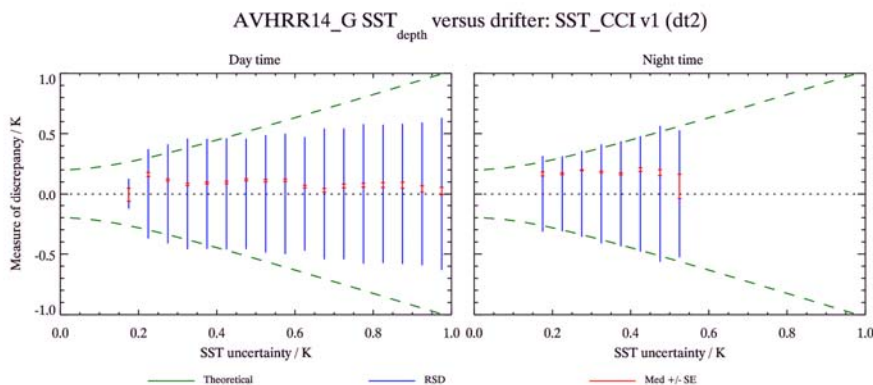
**Figure 7-45:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-14 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.



**Figure 7-46:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-14 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-47:** Histograms of the median discrepancy between AVHRR-14 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-48:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-14 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.



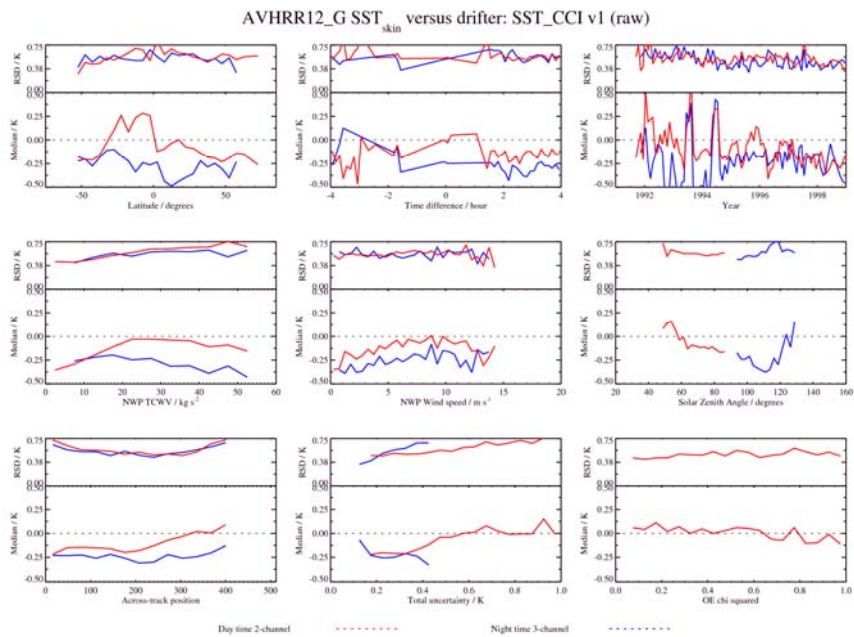
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	26933	+0.09	0.49
	<i>Night</i>	22856	+0.19	0.38
<b>iDrifters</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
<b>GTMBA</b>	<i>Day</i>	1046	+0.15	0.50
	<i>Night</i>	913	+0.19	0.32
<b>Argo</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
<b>Radiometers</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-

**Table 7-6:** Global validation statistics from comparing SST CCI AVHRR-14 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GTMBA and Argo; for radiometers only the time difference has been adjusted.

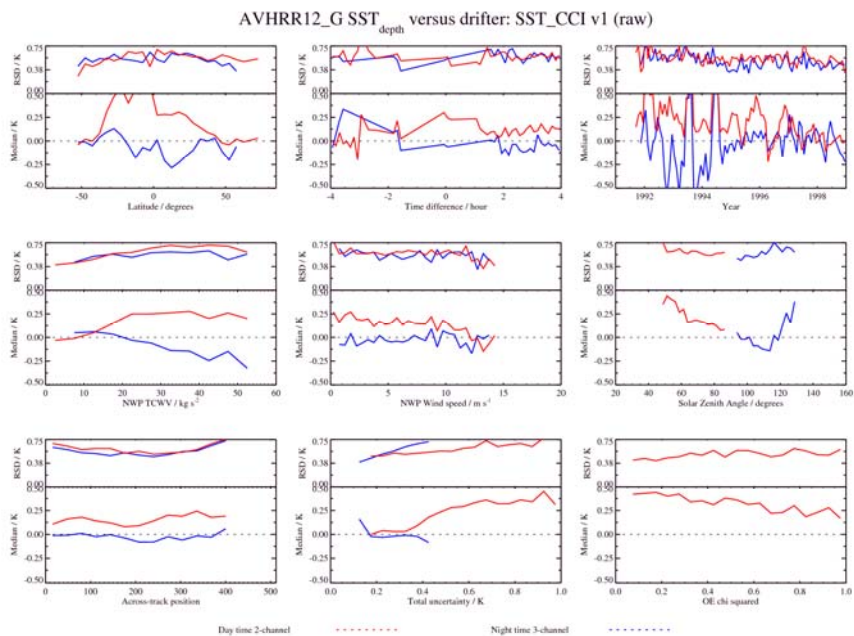
Summary of key findings from AVHRR-14 validation:

- Fewer match-ups than for other sensors; do not have complete global coverage
- Results similar to AVHRR-18 (both afternoon orbits)
- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel (larger than seen for AVHRR-18)
- Residual bias wind speed bias and TCWV dependence in both retrievals; also in satellite view angle (maybe evidence of solar contamination)
- Fewer regional variations but data apparently noisier
- Some residual cloud contamination
- Uncertainty estimates reasonable; better discrimination at night.

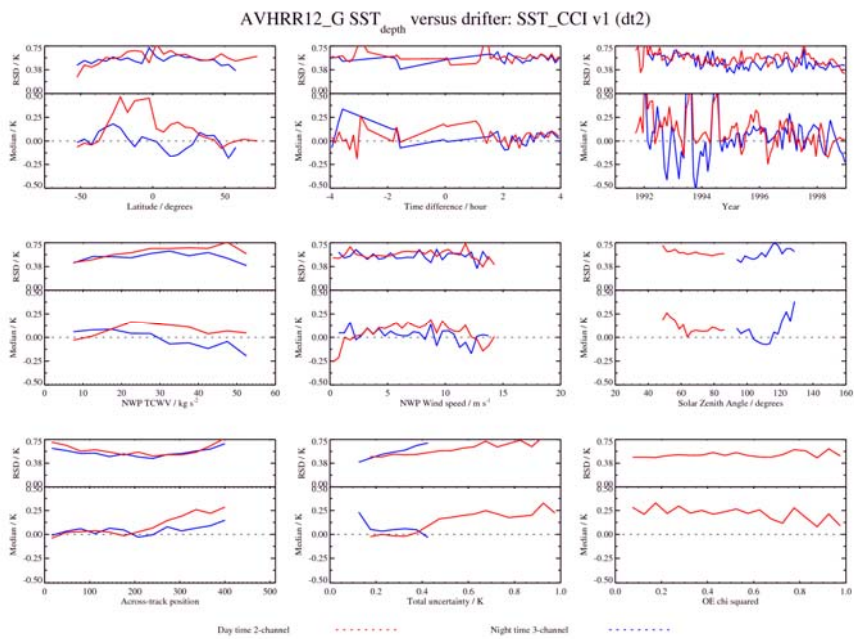
## A.7 AVHRR 12



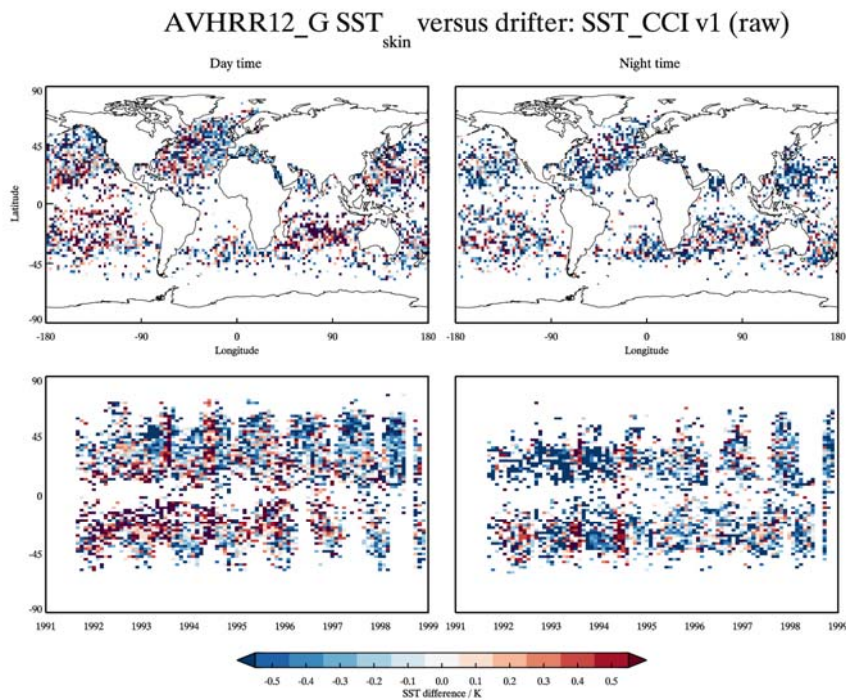
**Figure 7-49:** Dependence of the median and robust standard deviation between AVHRR-12 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-50:** Dependence of the median and robust standard deviation between AVHRR-12 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

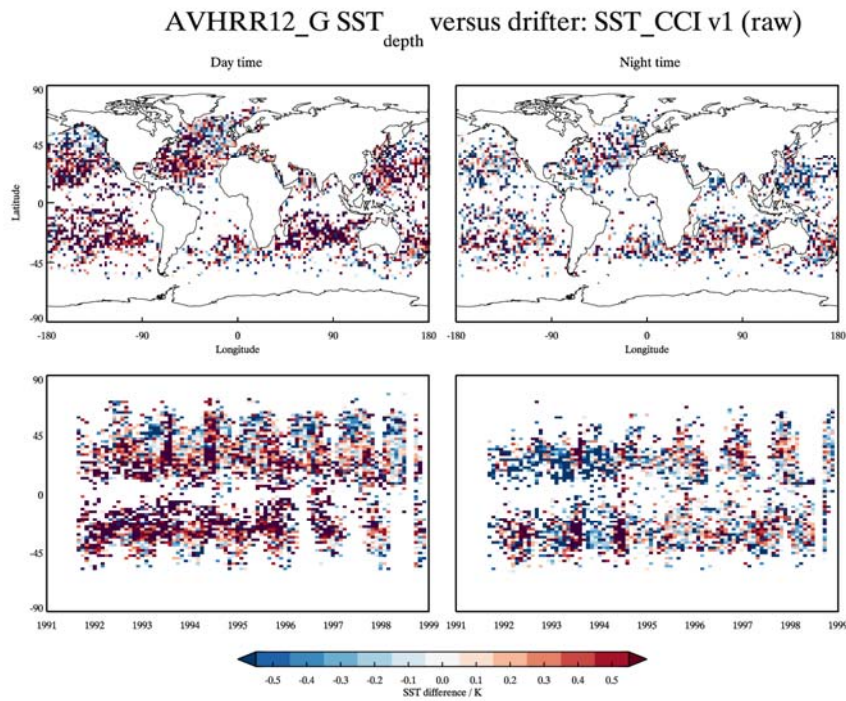


**Figure 7-51:** Dependence of the median and robust standard deviation between AVHRR-12 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

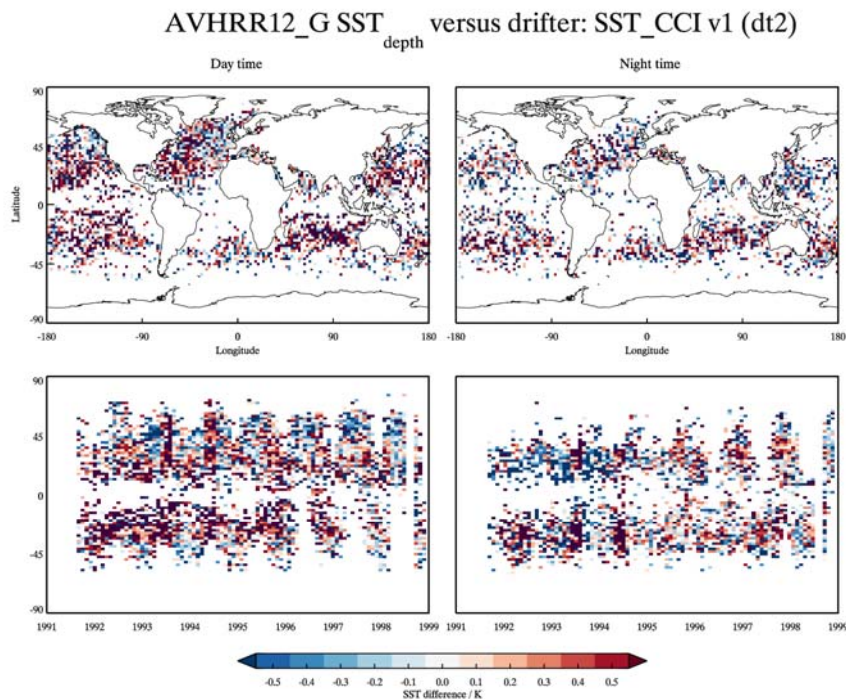


**Figure 7-52:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-12 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.

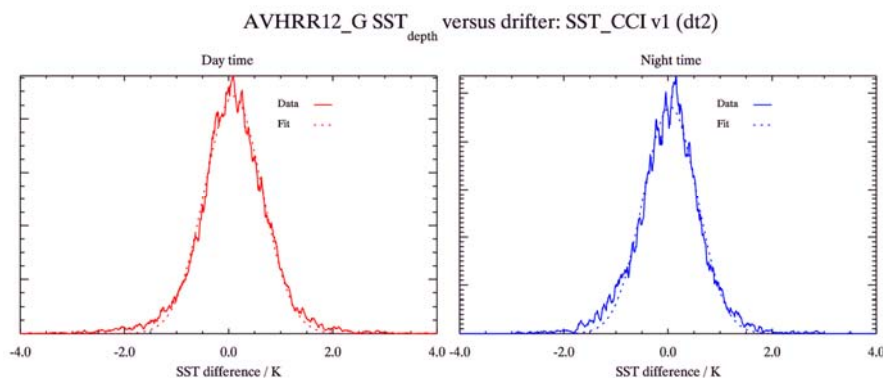




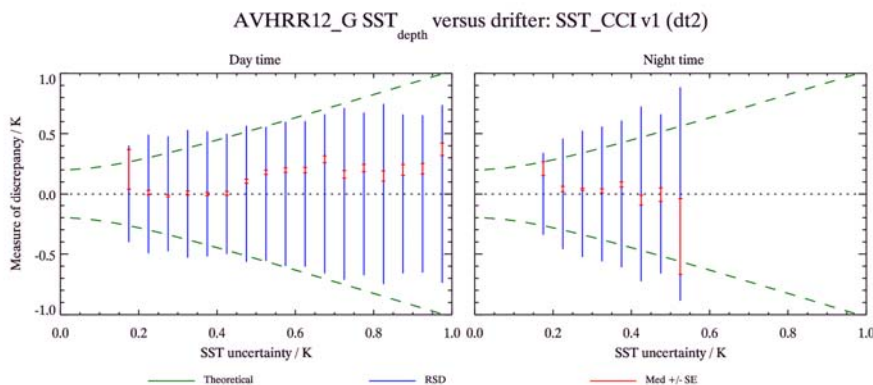
**Figure 7-53:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-12 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.



**Figure 7-54:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AVHRR-12 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-55:** Histograms of the median discrepancy between AVHRR-12 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-56:** Uncertainty validation plots for day time (left) and night time (right) AVHRR-12 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.



Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	12852	+0.08	0.58
	<i>Night</i>	6430	+0.04	0.54
<b>iDrifters</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
<b>GTMBA</b>	<i>Day</i>	1139	+0.24	0.54
	<i>Night</i>	848	-0.10	0.54
<b>Argo</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
<b>Radiometers</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-

**Table 7-7:** Global validation statistics from comparing SST CCI AVHRR-12 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GTMBA and Argo; for radiometers only the time difference has been adjusted.

Summary of key findings from AVHRR-12 validation:

- Fewer match-ups than for other sensors; do not have complete global coverage
- Results similar to AVHRR-18 (both afternoon orbits)
- Residual 2- and 3-channel bias with 3-channel warmer than 2- channel (larger than seen for AVHRR-18)
- Residual bias wind speed bias and TCWV dependence in both retrievals; also strong dependence on satellite view angle
- Fewer regional variations but data much noisier
- Little residual cloud contamination – retrieval uncertainties domination results
- Uncertainty estimates reasonable in day time – underestimated at night; better discrimination at night.
- Intermittent fluctuations in bias at all latitudes in earlier years
- Little evidence of residual stratospheric aerosol biases from Pinatubo eruption

## APPENDIX B DETAILED ATSR PRODUCT VALIDATION RESULTS

The following section contains the detailed validation results for the SST\_CCI long-term ECV ATSR products. For each sensor we provide:

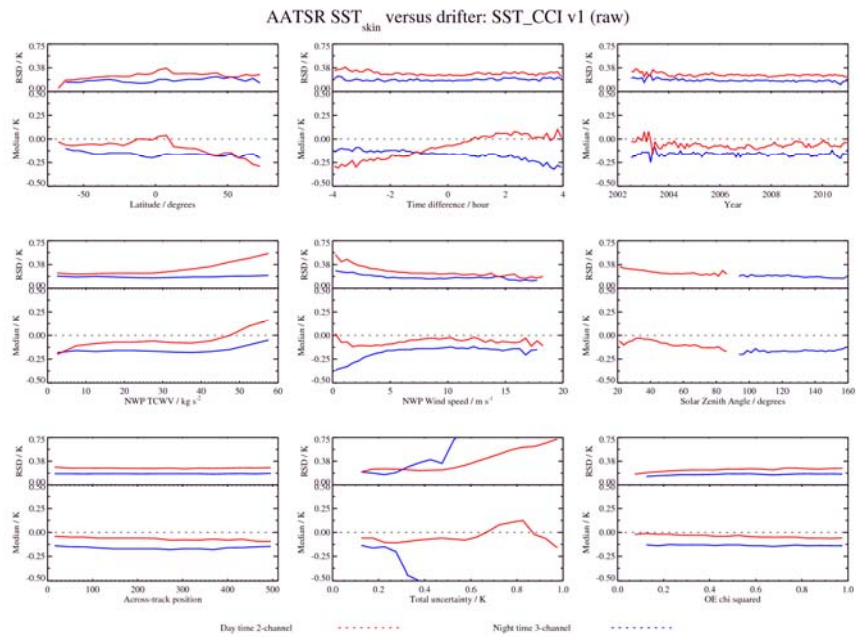
- Dependence plots of median and robust standard deviation of the discrepancy between the satellite and drifting buoys for
  - Satellite SST<sub>skin</sub> versus drifter SST<sub>depth</sub>.
  - Satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub>.
  - Satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub> with additional adjustments for the difference between the satellite and drifter measurement times (satellite at 10:30 am/pm local solar time).

Dependences are provided for latitude, time difference between satellite and drifter measurements, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and the retrieval chi squared function.

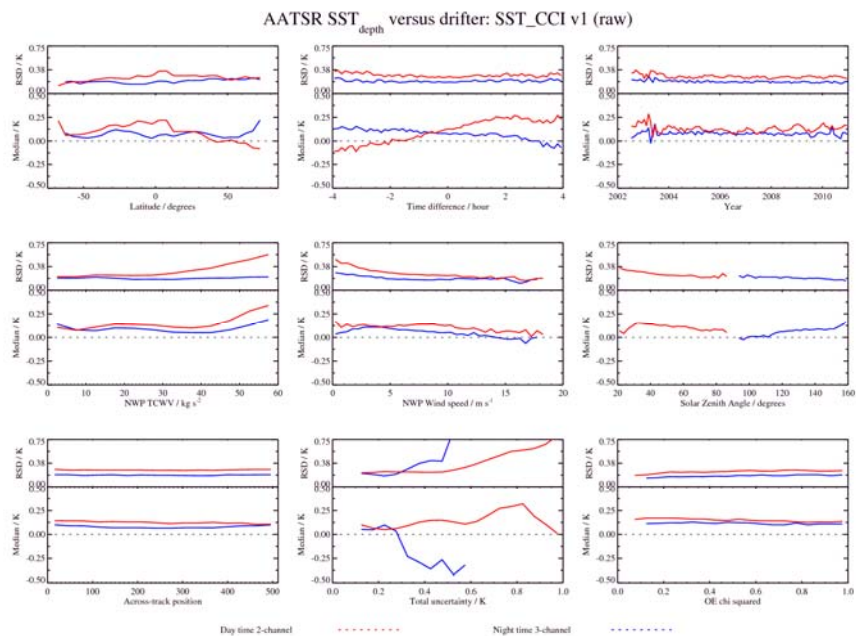
Note: A minimum of 30 match-ups is required for each point on the dependence plots (from central limit theorem). As such, the minimum standard error for a standard deviation of 0.5 K would be roughly 0.1 K.

- Spatial maps and Hovmoller plots of the median discrepancy between the satellite and drifting buoys for the same three comparisons as for the dependence plots.
- Histograms of the distributions of median discrepancies between the satellite and drifting buoys for satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub> with additional adjustments for the difference between the satellite and drifter measurement times.
- Uncertainty validation plots for the total uncertainty applicable to the satellite SST<sub>depth</sub> as a function of the median discrepancies between the satellite and drifting buoys for satellite SST<sub>depth</sub> versus drifter SST<sub>depth</sub> with additional adjustments for the difference between the satellite and drifter measurement times. For further details of the uncertainty validation methodologies please see section 5.
- A table of the median and robust standard deviation of the discrepancy between the satellite products and the various reference datasets for a selection of comparisons.
- A summary of the key findings for each sensor.

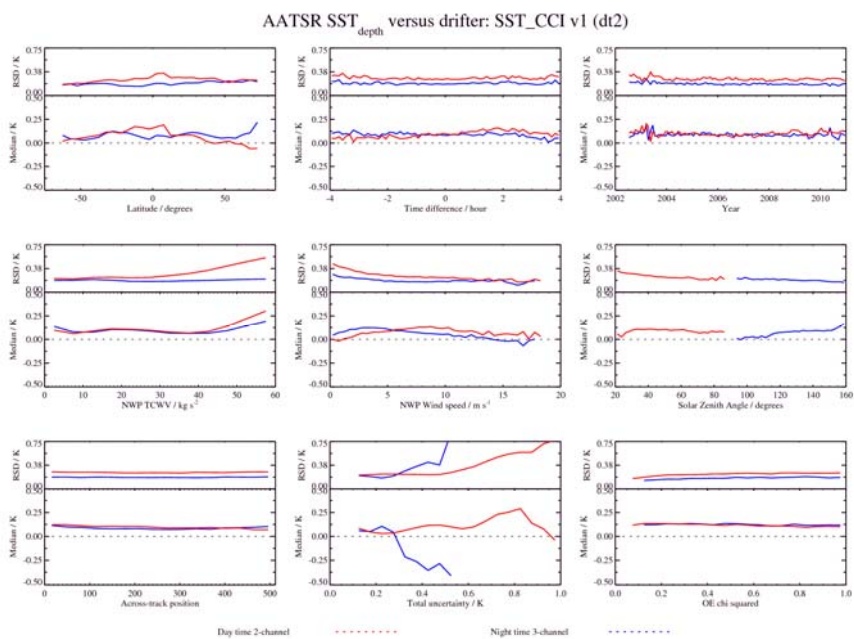
## B.1 AATSR



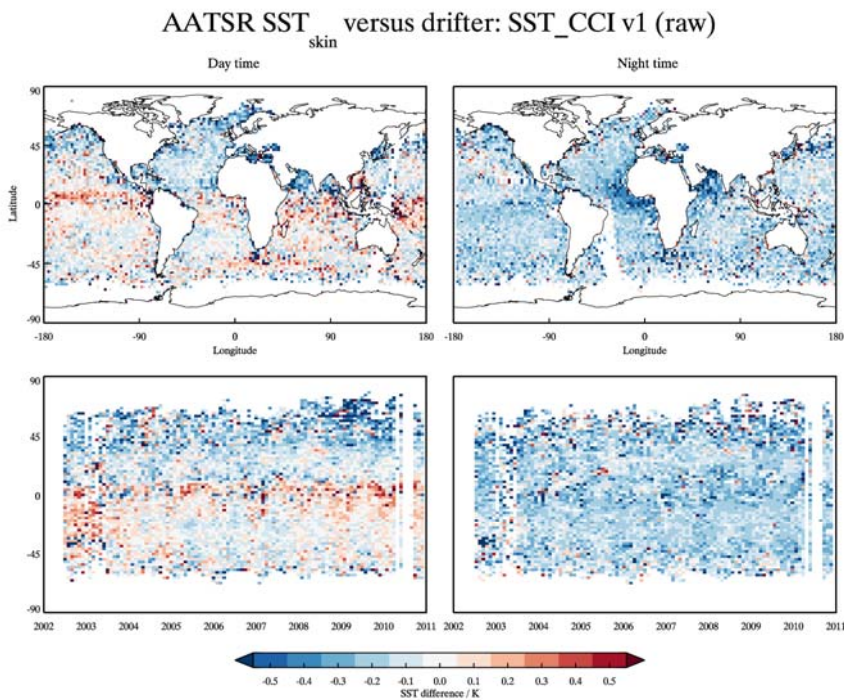
**Figure 7-57:** Dependence of the median and robust standard deviation between AATSR SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-58:** Dependence of the median and robust standard deviation between AATSR SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

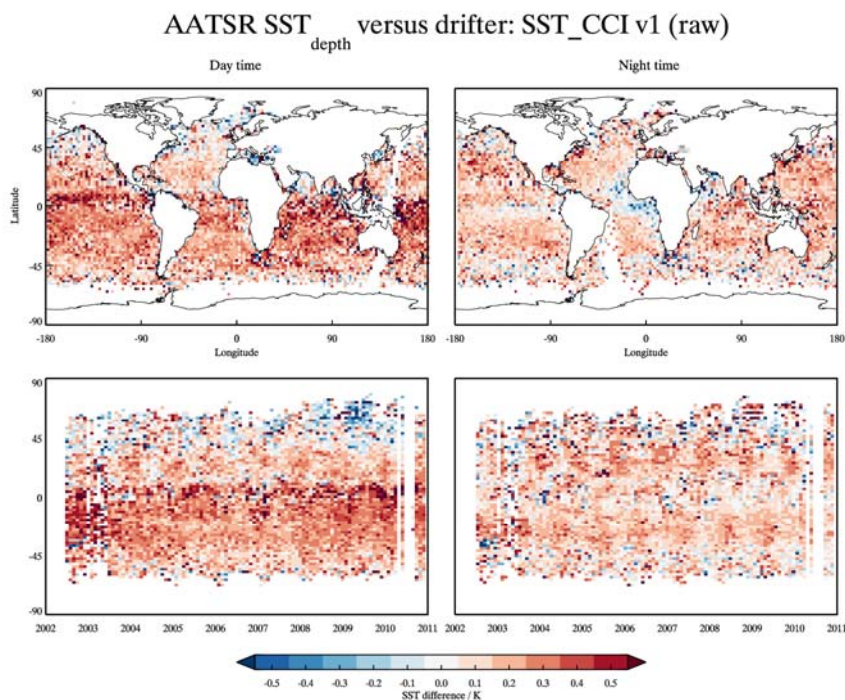


**Figure 7-59:** Dependence of the median and robust standard deviation between AATSR SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

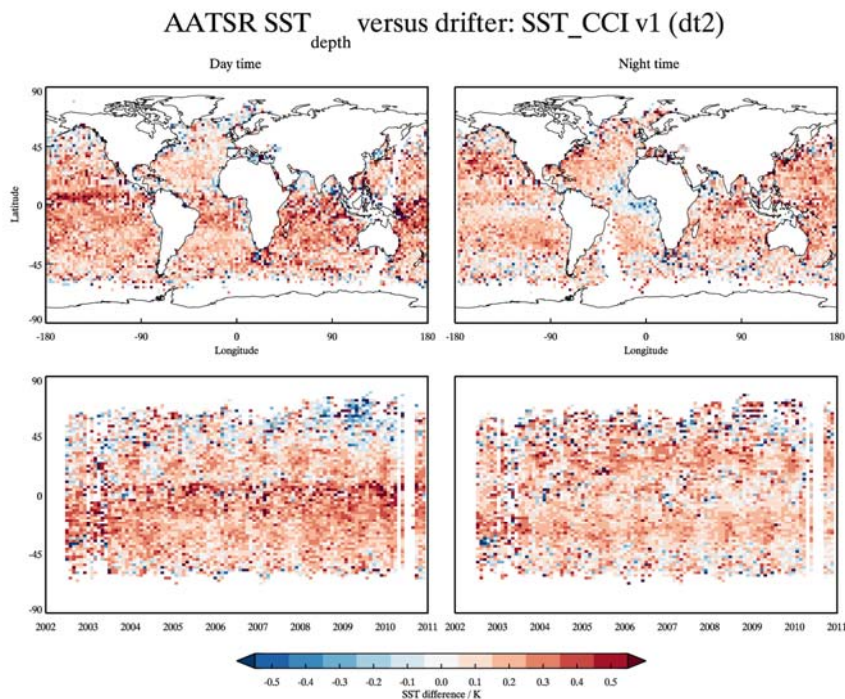


**Figure 7-60:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AATSR SST<sub>skin</sub> and drifter SST<sub>depth</sub>.



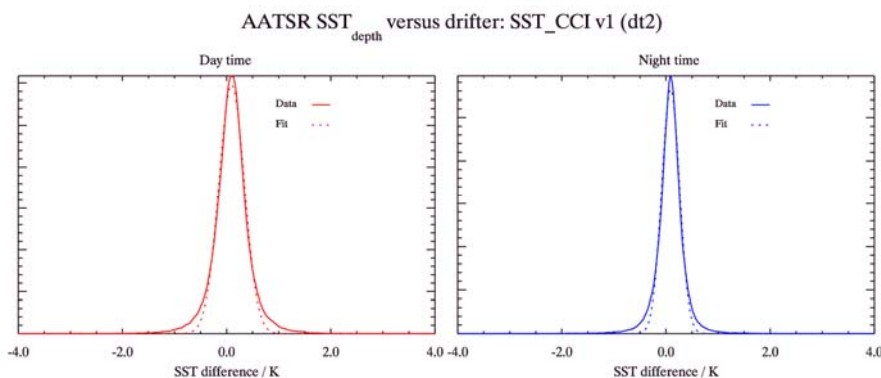


**Figure 7-61:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AATSR SST<sub>depth</sub> and drifter SST<sub>depth</sub>.

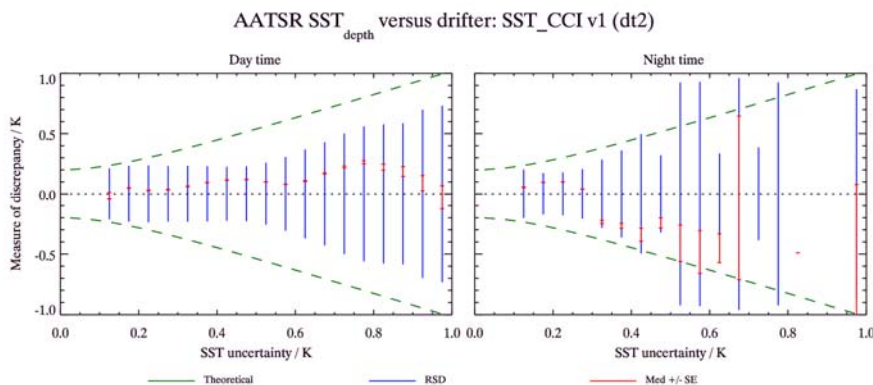


**Figure 7-62:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between AATSR SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).





**Figure 7-63:** Histograms of the median discrepancy between AATSR SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-64:** Uncertainty validation plots for day time (left) and night time (right) AATSR SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time. For a detailed explanation for the uncertainty validation plots please see section 5.

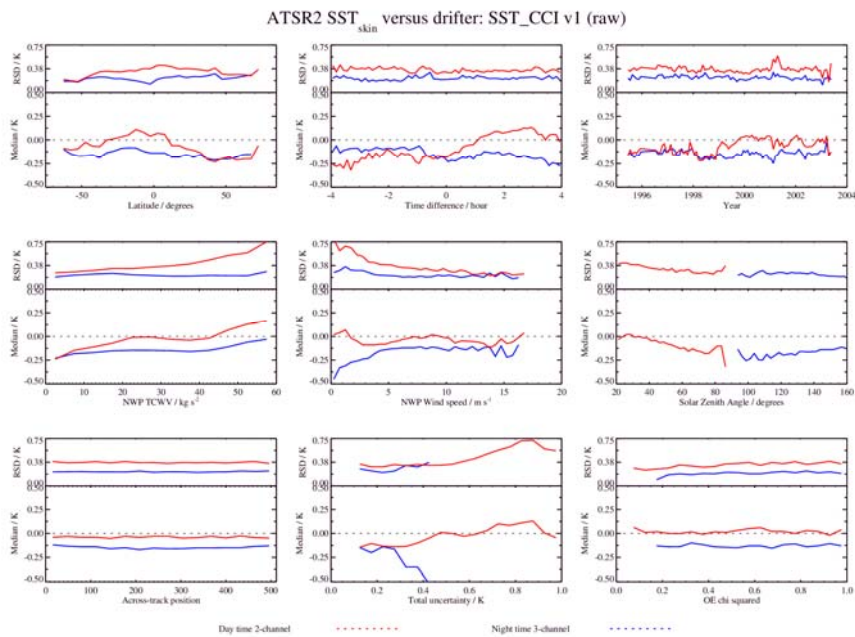
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	197853	+0.10	0.26
	<i>Night</i>	131944	+0.09	0.18
<b>iDrifters</b>	<i>Day</i>	16745	+0.10	0.26
	<i>Night</i>	11316	+0.08	0.18
<b>GT MBA</b>	<i>Day</i>	9977	+0.12	0.27
	<i>Night</i>	3878	+0.04	0.15
<b>Argo</b>	<i>Day</i>	3507	+0.08	0.25
	<i>Night</i>	1731	+0.08	0.16
<b>Radiometers</b>	<i>Day</i>	159	-0.04	0.35
	<i>Night</i>	142	-0.01	0.20

**Table 7-8:** Global validation statistics from comparing SST CCI AATSR to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GT MBA and Argo; for radiometers only the time difference has been adjusted.

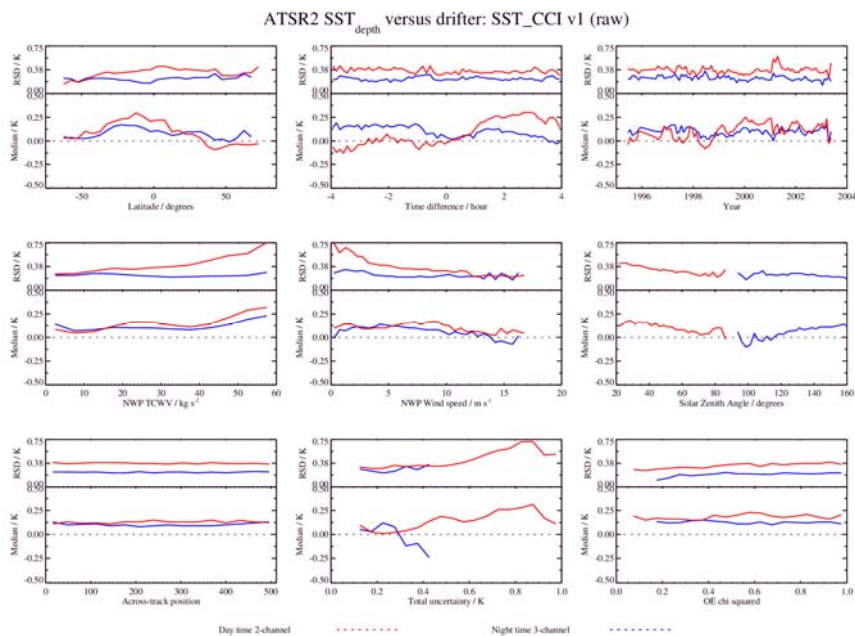
Summary of key findings from AATSR validation:

- Residual bias in wind speed and TCWV dependence in both retrievals
- Cool bias in Arctic
- Evidence of desert dust effects
- Residual cloud contamination in tropics in daytime
- Uncertainty estimates marginal for day-time; slightly better discrimination at night.
- Persistent data gaps, different in location day and night, reflecting processing bug (fixed, so gaps will not appear in next version)

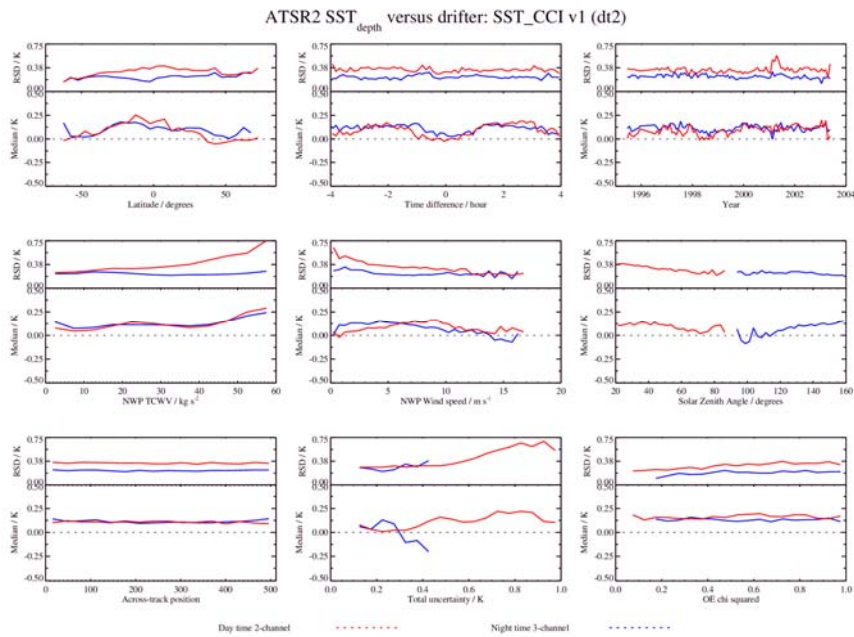
## B.2 ATSR-2



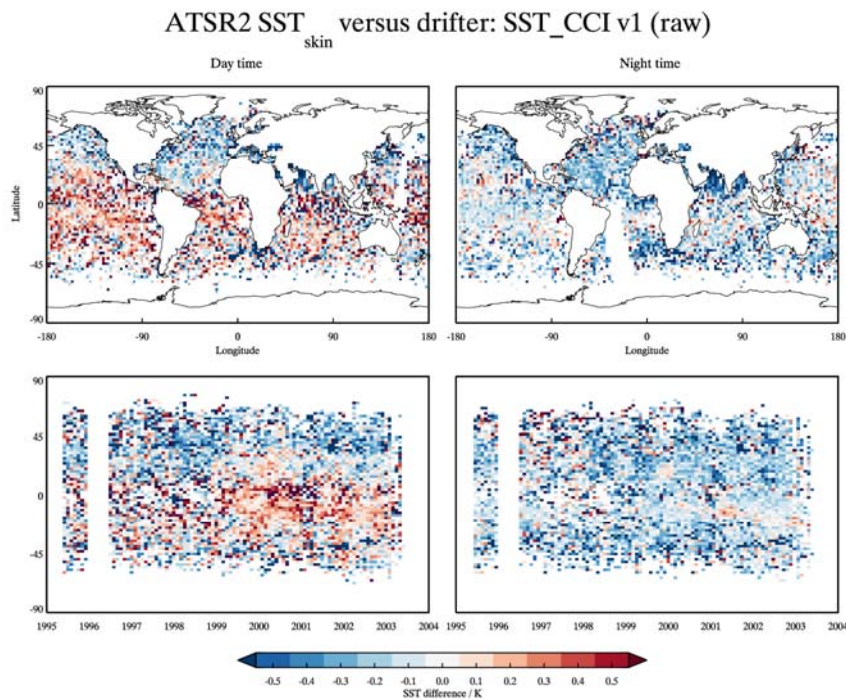
**Figure 7-65:** Dependence of the median and robust standard deviation between ATSR-2 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.



**Figure 7-66:** Dependence of the median and robust standard deviation between ATSR-2 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue.

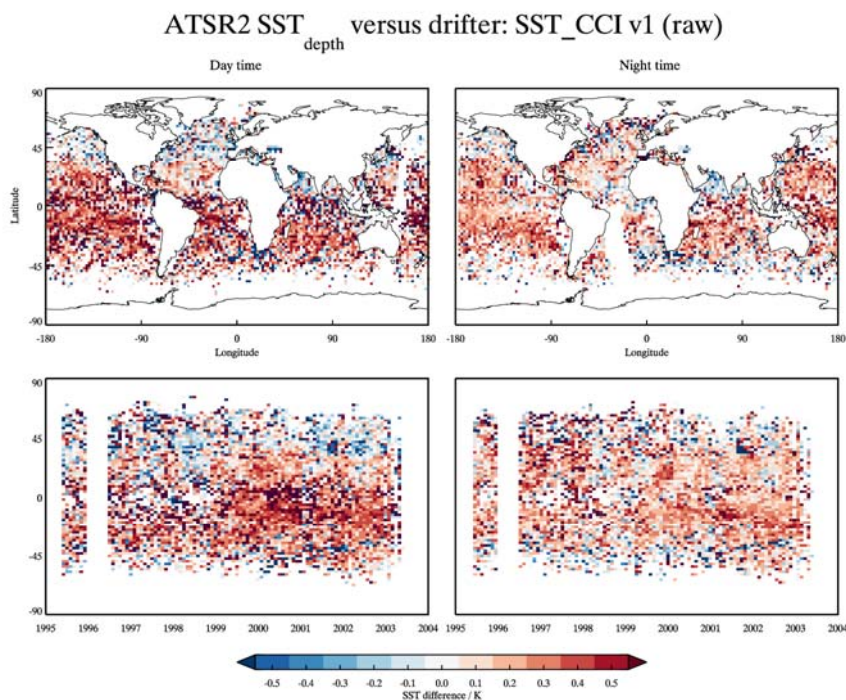


**Figure 7-67:** Dependence of the median and robust standard deviation between ATSR-2 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red and night time results are shown in blue. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

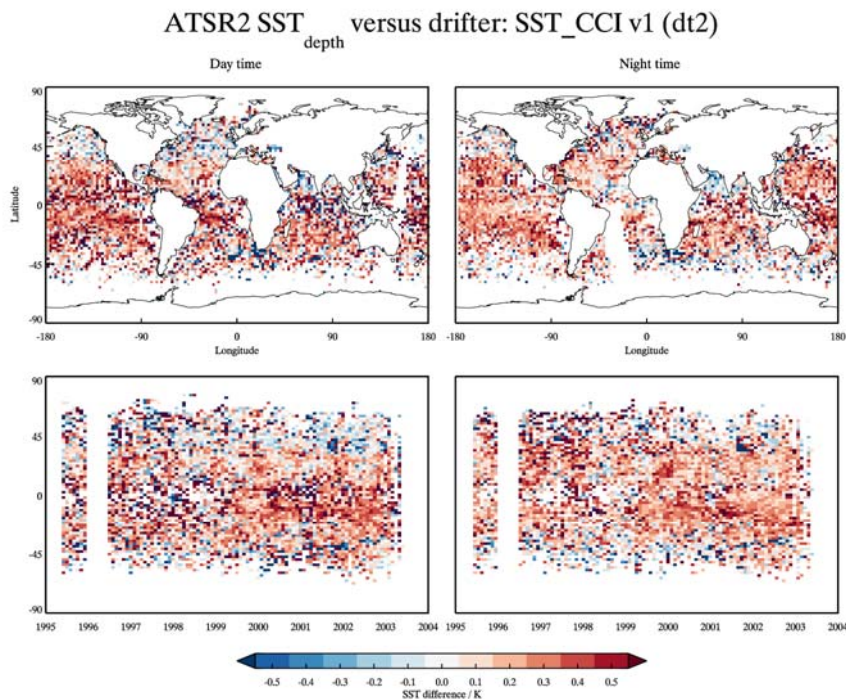


**Figure 7-68:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between ATSR-2 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.



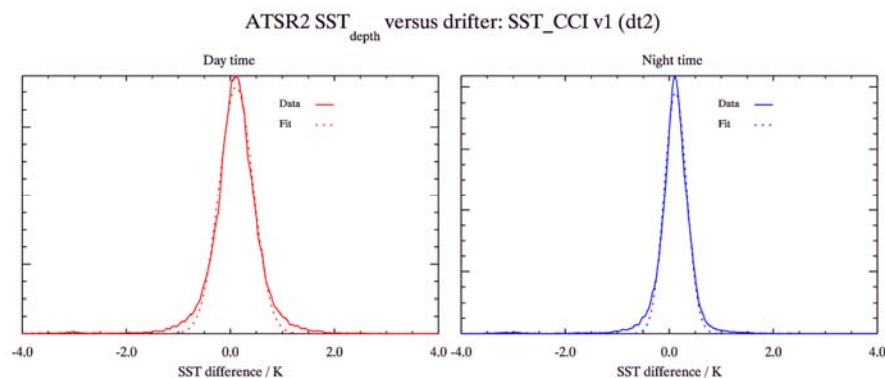


**Figure 7-69:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between ATSR-2 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.

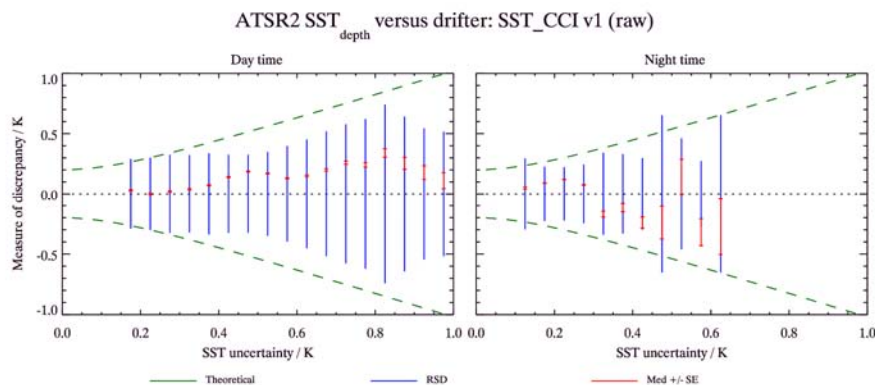


**Figure 7-70:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between ATSR-2 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).





**Figure 7-71:** Histograms of the median discrepancy between ATSR-2 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left) and night time (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-72:** Uncertainty validation plots for day time (left) and night time (right) ATSR-2 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.

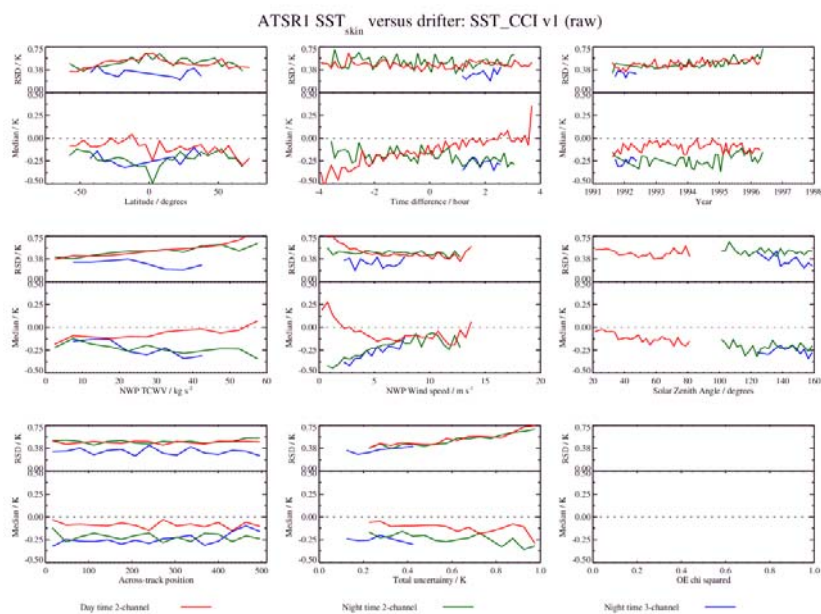
Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	62547	+0.11	0.35
	<i>Night</i>	45211	+0.11	0.23
<b>iDrifters</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
<b>GTMBA</b>	<i>Day</i>	7940	+0.11	0.32
	<i>Night</i>	1590	+0.11	0.16
<b>Argo</b>	<i>Day</i>	138	+0.10	0.36
	<i>Night</i>	106	+0.13	0.20
<b>Radiometers</b>	<i>Day</i>	66	-0.08	0.41
	<i>Night</i>	81	-0.03	0.21

**Table 7-9:** Global validation statistics from comparing SST CCI ATSR-2 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GTMBA and Argo; for radiometers only the time difference has been adjusted.

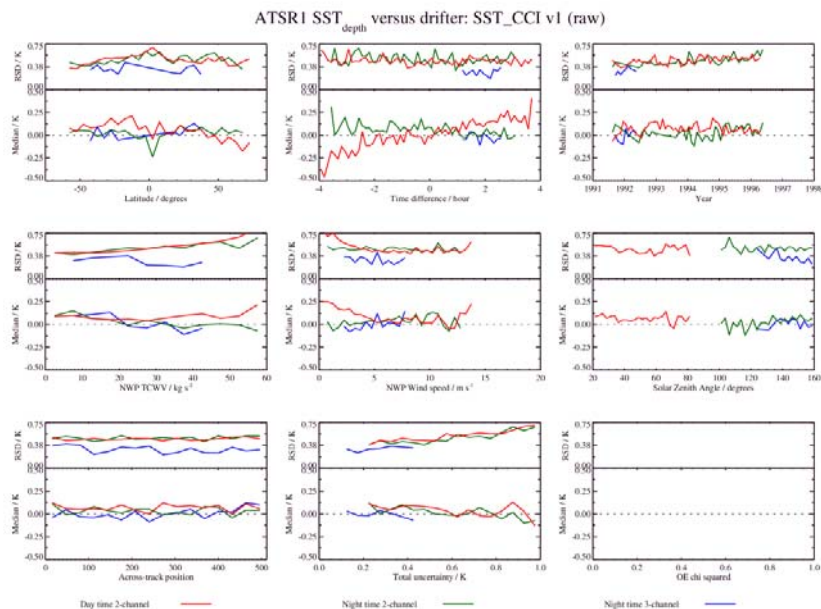
Summary of key findings from ATSR-2 validation:

- Results similar to AATSR
- Spatial patterns influenced by varying drifter coverage over time
- Residual bias in wind speed and TCWV dependence in both retrievals
- Cool bias in Arctic
- Evidence of residual Saharan Dust effects
- Uncertainty estimates acceptable – better consistency between day and night than for AATSR
- Residual time difference effects suggests issues with drifting buoy reporting within this time period
- Persistent data gaps, different in location day and night, reflecting processing bug (fixed, so gaps will not appear in next version)

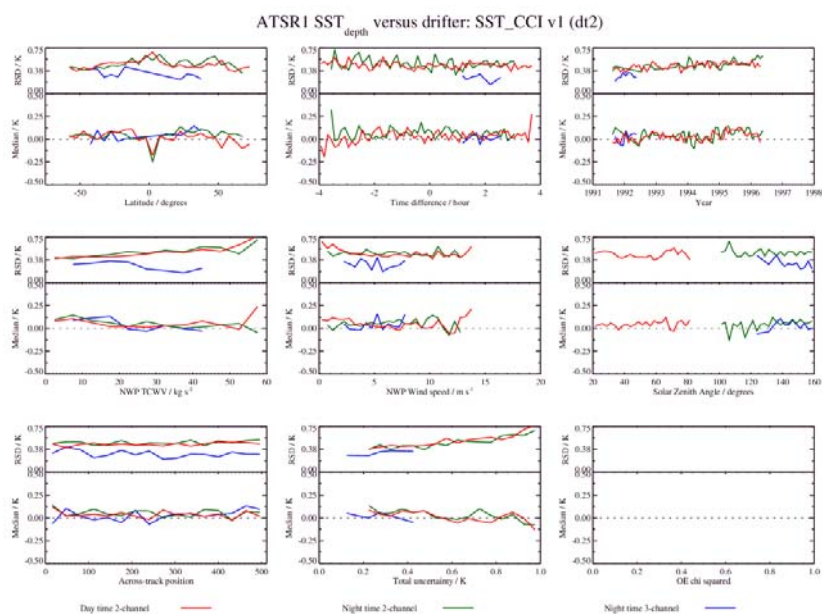
### B.3 ATSR-1



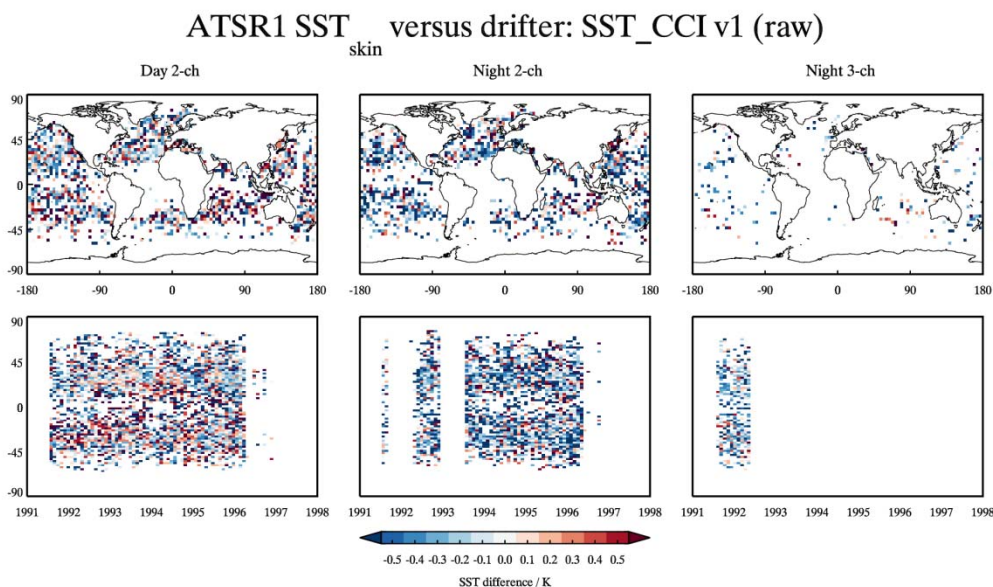
**Figure 7-73:** Dependence of the median and robust standard deviation between ATSR-1 SST<sub>skin</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red, night time 3-channel results are shown in blue and night time 2-channel results are shown in green.



**Figure 7-74:** Dependence of the median and robust standard deviation between ATSR-1 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red, night time 3-channel results are shown in blue and night time 2-channel results are shown in green.

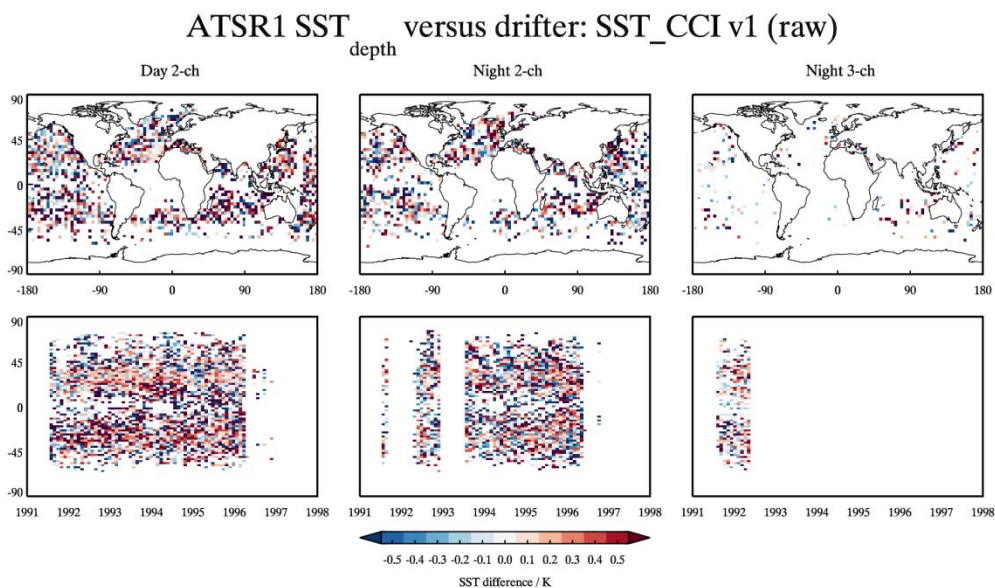


**Figure 7-75:** Dependence of the median and robust standard deviation between ATSR-1 SST<sub>depth</sub> and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, total column water vapour, wind speed, solar zenith angle, across-track position, total uncertainty and retrieval chi squared function. Day time results are shown in red, night time 3-channel results are shown in blue and night time 2-channel results are shown in green. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).

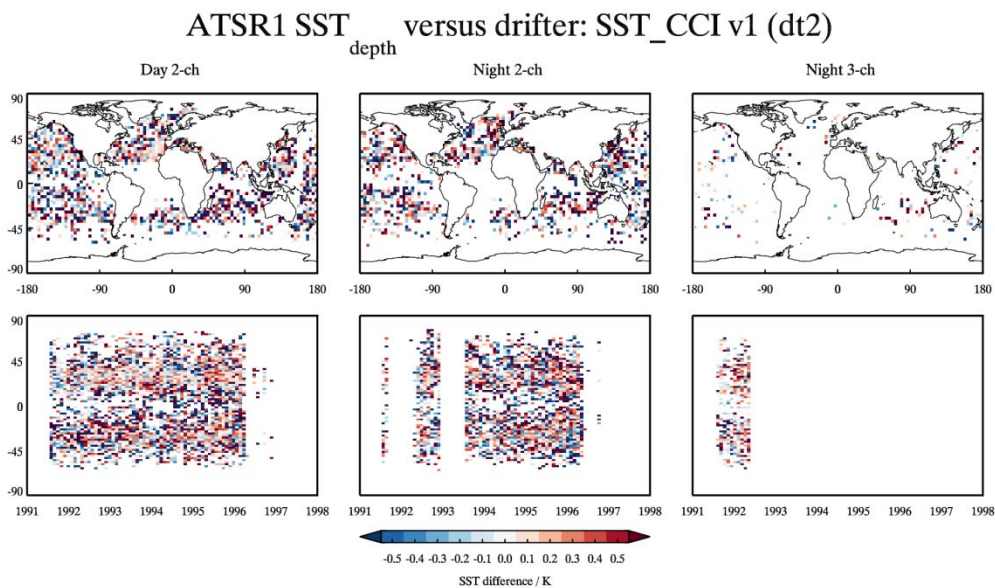


**Figure 7-76:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between ATSR-1 SST<sub>skin</sub> and drifter SST<sub>depth</sub>.



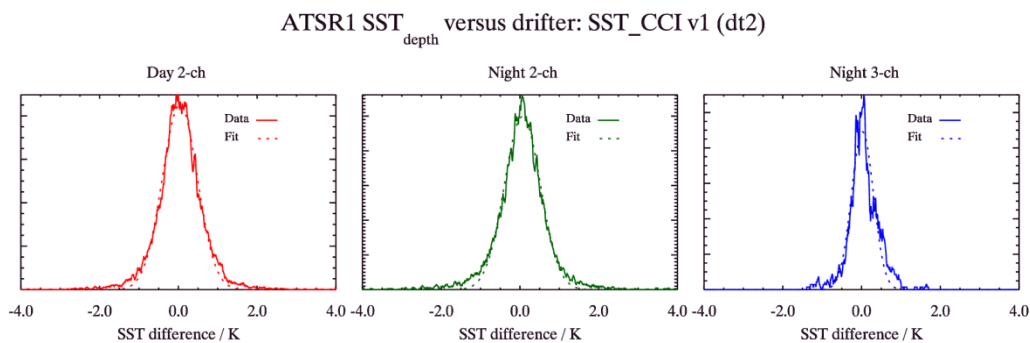


**Figure 7-77:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between ATSR-1 SST<sub>depth</sub> and drifter SST<sub>depth</sub>.

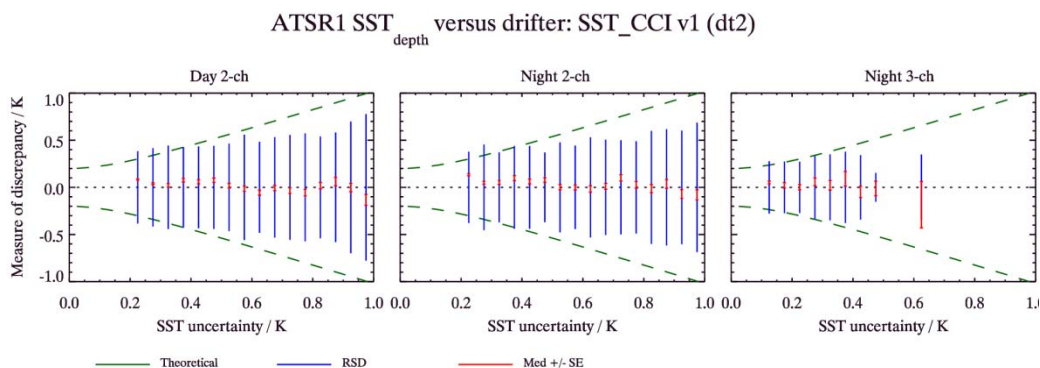


**Figure 7-78:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between ATSR-1 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).





**Figure 7-79:** Histograms of the median discrepancy between ATSR-1 SST<sub>depth</sub> and drifter SST<sub>depth</sub> for day time (left), night time 2-channel (middle) and night time 3-channel (right). An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time).



**Figure 7-80:** Uncertainty validation plots for day time (left), night time 2-channel (middle) and night time 3-channel (right) ATSR-1 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and drifter measurements (satellite at 10:30 am/pm local solar time). For a detailed explanation for the uncertainty validation plots please see section 5.

Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	10093	+0.04	0.47
	<i>Night 2-</i>	6151	+0.06	0.49
	<i>Night 3-</i>	681	+0.04	0.30
<b>iDrifters</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
	<i>Night 3-</i>	-	-	-
<b>GTMBA</b>	<i>Day</i>	3528	+0.04	0.47
	<i>Night</i>	1990	+0.01	0.46
	<i>Night 3-</i>	138	-0.01	0.15
<b>Argo</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
	<i>Night 3-</i>	-	-	-
<b>Radiometers</b>	<i>Day</i>	-	-	-
	<i>Night</i>	-	-	-
	<i>Night 3-</i>	-	-	-

**Table 7-10:** Global validation statistics from comparing SST CCI ATSR-1 to the reference dataset. An additional adjustment has been made using a combined diurnal variability/skin model to account for the difference in time between the satellite and reference measurements (satellite at 10:30 am/pm local solar time) for drifters, GTMBA and Argo; for radiometers only the time difference has been adjusted.

Summary of key findings from ATSR-1 validation:

- Results different to ATSR-2 and AATSR
- Low number of match-ups – not all geographic regions sampled
- Some residual bias in TCWV dependence at low end
- No residual effects of stratospheric aerosol (improvement compared to ARC results)
- Uncertainty estimates give marginal discrimination – although consistent between day and night
- Arguably best performing retrieval of all sensors

## APPENDIX C DETAILED ANALYSIS PRODUCT VALIDATION RESULTS

The following section contains the detailed validation results for the SST\_CCI analysis products. For each analysis we provide:

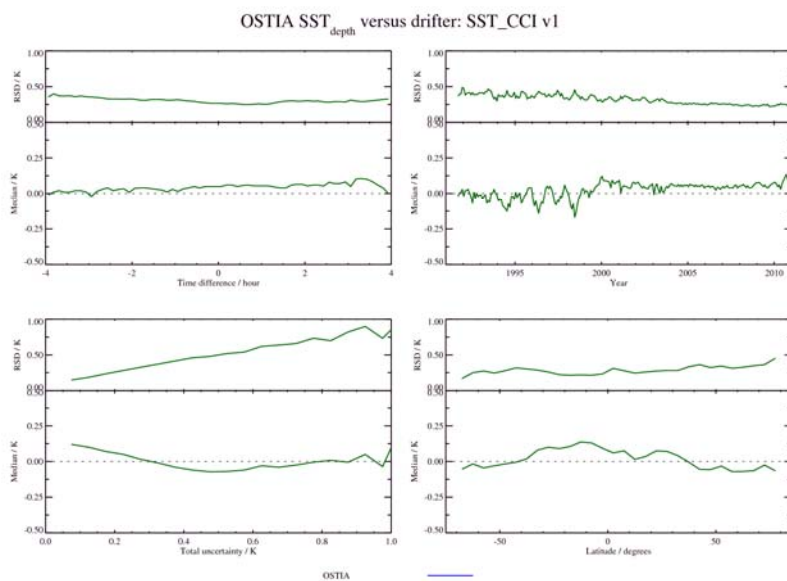
- Dependence plots of median and robust standard deviation of the discrepancy between the analysis and drifting buoys for analysis SST<sub>depth</sub> versus drifter SST<sub>depth</sub>.

Dependences are provided for latitude, time difference between analysis time and drifter measurements, year, and the analysis uncertainty.

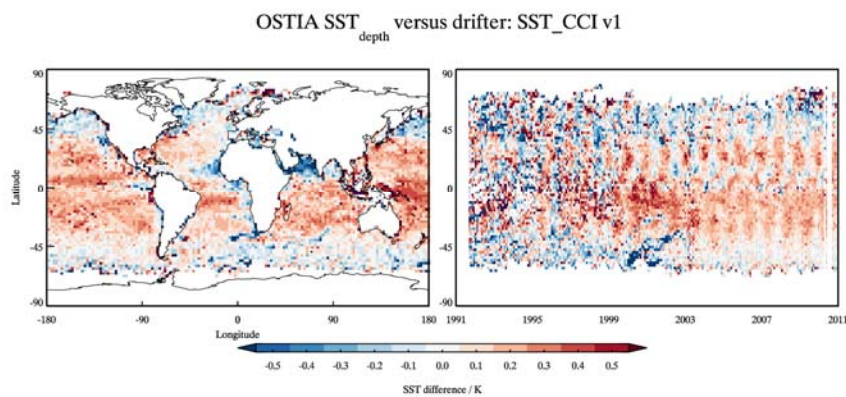
Note: A minimum of 30 match-ups is required for each point on the dependence plots (from central limit theorem). As such, the minimum standard error for a standard deviation of 0.5 K would be roughly 0.1 K.

- Spatial maps and Hovmoller plots of the median discrepancy between the analysis and drifting buoys for analysis SST<sub>depth</sub> versus drifter SST<sub>depth</sub>.
- Histograms of the distributions of median discrepancies between the analysis and drifting buoys for analysis SST<sub>depth</sub> versus drifter SST<sub>depth</sub>.
- Uncertainty validation plots for the total uncertainty applicable to the analysis SST<sub>depth</sub> as a function of the median discrepancies between analysis and drifting buoys for analysis SST<sub>depth</sub> versus drifter SST<sub>depth</sub>. For further details of the uncertainty validation methodologies please see section 5.
- A table of the median and robust standard deviation of the discrepancy between the analysis and the various reference datasets.
- A summary of the key findings for each analysis.

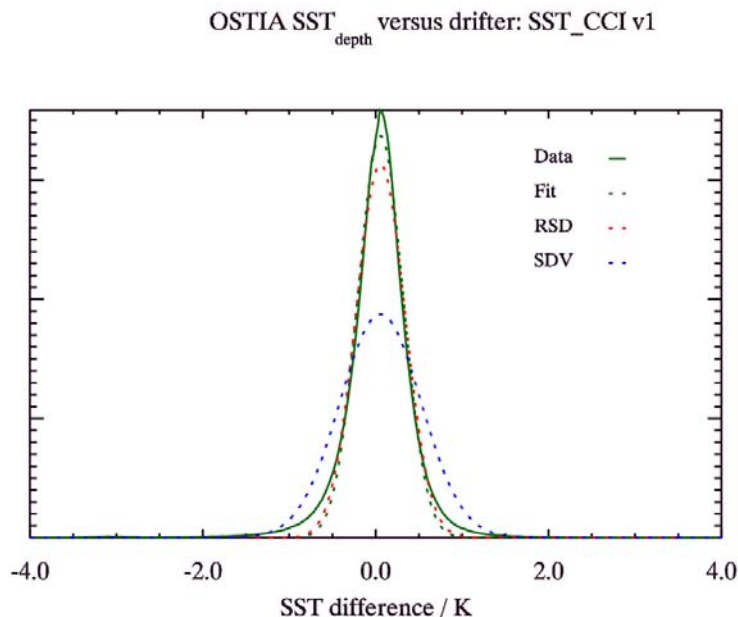
### C.1 SST\_CCI analysis long-term product



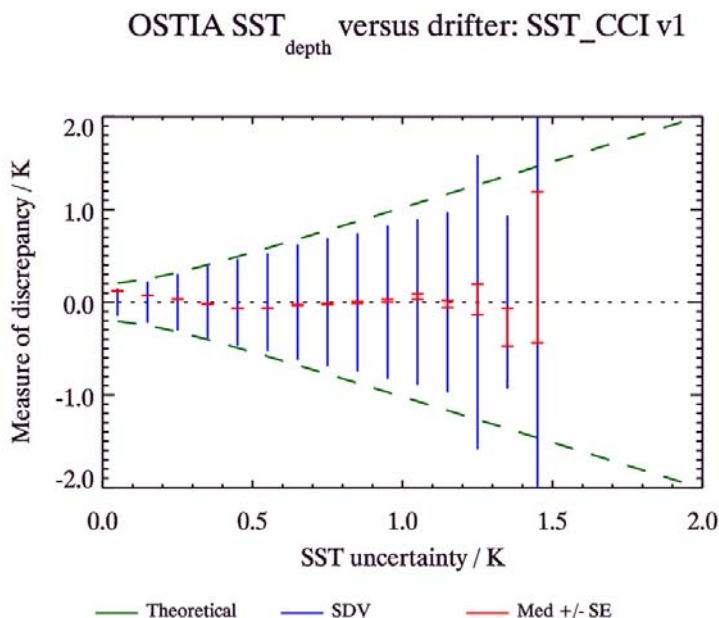
**Figure 7-81:** Dependence of the median and robust standard deviation between SST\_CCI L4 and drifter SST<sub>depth</sub> discrepancies as a function of latitude, time difference, year, and analysis uncertainty.



**Figure 7-82:** Spatial distribution and Hovmoller plot of the Dependence of the median discrepancy between SST\_CCI L4 and drifter SST<sub>depth</sub>.



**Figure 7-83:** Histogram of the median discrepancy between SST\_CCI L4 SST<sub>depth</sub> and drifter SST<sub>depth</sub>. Also shown are the results of a Gaussian fit to the distribution as well as Gaussian functions derived from the robust and non-robust standard deviations of the discrepancies.



**Figure 7-84:** Uncertainty validation plot for SST\_CCI L4 SST<sub>depth</sub> assessed using drifter SST<sub>depth</sub>. For a detailed explanation for the uncertainty validation plots please see section 5.



Reference	Retrieval	Number	Median (K)	RSD (K)
<b>Drifters</b>	<i>Day</i>	2392462	+0.05	0.28
<b>iDrifters</b>	<i>Day</i>	142902	+0.06	0.25
<b>GTMBA</b>	<i>Day</i>	25492	+0.09	0.22
<b>Argo</b>	<i>Day</i>	8867	+0.04	0.26
<b>Radiometers</b>	<i>Day</i>	696	+0.05	0.46

**Table 7-11:** Global validation statistics from comparing SST\_CCI L4 to the reference dataset.

Summary of key findings from SST CCI L4 validation:

- Spatial patterns have strong similarity to ATSRs
- Evidence of day time bias i.e. not always daily mean in spatial patterns
- Daily bias confirmed by residual time dependence indicating ocean warming
- Not possible to adjust in situ measurement OSTIA time so strong residual cycling evident for some periods
- Residuals much noisier in early part of record (influence by drifter coverage)
- Cool bias in Arctic and Southern Oceans
- Desert Dust effects
- Uncertainty estimates are very good

## APPENDIX D ASSESSMENT OF USER REQUIREMENTS

The first activity of the SST\_CCI project was a detailed user requirements review. The results and conclusions from the user requirements review are provided in the SST\_CCI URD, RD.171. An extract of those requirements that have direct implications for product validation and intercomparison is given in Table 7-12 along with an indication of how each requirement has been addressed in this document. User requirements that have an indirect bearing on the approach to product validation and intercomparison are not included here as they are addressed elsewhere (e.g. requirements on what types of products are to be produced are addressed via the Product Specification Document (RD.175), whose content is assumed here).

Requirement identifier	Requirement	Comments (from URD)	How we have addressed this UR in the PVIR
SST_CCI-UR-QUF-48	The most common acceptable levels of bias were 0.1 and 0.3°C (threshold), and 0.1°C (breakthrough and objective). The most common response was that the achievement of this should be demonstrated over a spatial scale of 100 km.		The difference maps give an estimate of bias across spatial scales determined by the statistical power of available validation data.
SST_CCI-UR-QUF-49	The most common response was that 0.1°C is the required precision and that the achievement of this should be demonstrated over a spatial scale of 100 km.		The RSD statistics quantify precision against latitude and other influential factors.
SST_CCI-UR-QUF-50	At the threshold, breakthrough, and objective requirement levels, 0.1°C per decade was the most common response for the acceptable level of drift. The most common response for the spatial scale that the achievement of this should be demonstrated over was 100 km.	However, a significant number of users have stricter requirements, particularly at the breakthrough and objective levels.	The prime discussion of stability is in the SST CCI Climate Assessment Report, since trends in the SST data are analysed therein.
SST_CCI-UR-QUF-51	At the threshold, breakthrough and objective requirement levels, the most common response for the acceptable drift in	However, many users have stricter requirements.	As for QUF-59

Requirement identifier	Requirement	Comments (from URD)	How we have addressed this UR in the PVIR
	relative bias between day and night SSTs was 0.1°C per decade. The most common requirement was that the achievement of this should be demonstrated over a spatial scale of 100 km.		
SST_CCI-UR-QUF-52	At all requirement levels, the most common response was that 0.1°C per decade is the acceptable change in bias over the annual cycle. The most common requirement was that the achievement of this should be demonstrated over a spatial scale of 100 km.		As for QUF-50
<b>Uncertainty information</b>			
SST_CCI-UR-REF-7	Uncertainty characteristics should be verified by comparison against independent observations.	[RD-3]	Uncertainty estimates in products have been validated against independent measurements as part of the product validation (see Section 5)
<b>Requirements for features of the data</b>			
SST_CCI-UR-QUF-78	Verification against independent data.	Classed as essential or preferable by 83% of respondents.	Product validation against reference data set (see Section 4.1.2 and Section 4.2).
SST_CCI-UR-DIS-125	Independent validation/verification by a separate [independent] group is required.		Product validation and climate assessment are undertaken by project members not involved in retrieval algorithm development (see Section 4.1.3)

**Table 7-12:** Summary of SST\_CCI user requirements relevant to product validation and intercomparison.

## APPENDIX E ADHERENCE TO CCI PROJECT GUIDELINES

The first collocation meeting of the ESA CCI was held at ESA ESRIN, Frascati, Italy on 12th-15th September 2010. The collocation brought together representatives of all eleven CCI project teams to discuss areas of common interest. The output of the collocation was a series of recommendations (RD.169). These recommendations are intended to assist the CCI teams to implement their projects and generate ECV data products in a consistent manner, as explicitly required by GCOS.

Two sets of the series of recommendations are relevant to this document, those on round robin (RR) and those on validation (V). Table 7-13 summarises the recommendations for validation and explains how each one has been addressed within the SST\_CCI project.

Number	Recommendation	Adhered to in SST_CCI	Comment where required
V1	All CCI projects should use the definition of validation approved by the CEOS-WGCV.	Yes	The definition is given in Section 2.
V2	All CCI project Product Validation Plans (PVP) shall adhere to the following three requirements regarding independence: <ol style="list-style-type: none"> <li>1. CCI project teams shall use, for validation, in situ or other suitable reference datasets that have not been used during the production of their CCI products.</li> <li>2. CCI project teams shall consider the independence of the geophysical process and ensure that if a particular auxiliary dataset is used in the production of their CCI products then the same dataset is not used in the validation and, if required, alternative auxiliary data are used.</li> <li>3. CCI project teams shall ensure that the validation is carried out (or at least verified) by staff not involved in the final algorithm selection; ideally the validation of the CCI products should be carried out by external parties, i.e. by staff / institutions not involved in the production of the ECVs products.</li> </ol>	Yes	<ol style="list-style-type: none"> <li>1. Product validation will use the reference dataset, which was not used in production.</li> <li>2. Auxiliary datasets used for validation have not been used in production.</li> <li>3. Most validation activities are carried out by personnel not involved in algorithm selection or product generation.</li> </ol>
V3	The CCI consortia shall use established, community accepted, traceable validation protocols where they exist. If such protocols do not exist then CCI projects may adapt existing protocols if appropriate and in any event shall offer their final protocol for future community acceptance.	Yes	PVP circulated to GHRSS ST-VAL group.
V4	Each CCI project shall select appropriate validation data to ensure that an adequate level of validation (confidence) is applied to all output products. The level of validation (confidence) should be indicated in the output product.	Yes	See Section 4.

Number	Recommendation	Adhered to in SST_CCI	Comment where required
V5	The CCI programme should hold a dedicated session (or workshop) on common validation infrastructure during (or prior to) the next collocation meeting.	Yes	The relevant interactions occur on an annual basis via involvement in GHRSSST.
V6	The PVP shall fully describe the validation process for each CCI project. An independent international review board of experts should be invited to review the PVP of each project team. Each CCI project should involve experts from the CMUG throughout their validation activities. A CCI product will be deemed to be validated once all steps of the validation process documented in the PVP have been completed and documented accordingly.	Yes	The PVP was presented at the 2012 meeting of GHRSSST (Tokyo), giving the PVP international scrutiny. GHRSSST ST-VAL group invited to review document.

**Table 7-13:** Summary of recommendations relevant to the round robin and product validation from the first CCI collocation and adherence within the SST\_CCI project



***This Page Is Intentionally Blank***