



Customer	: ESRIN	Document Ref	: SST_CCI-PVP-UoL-001
WP No	:	Issue Date	: 04 February 2014
		Issue	: 2

Project : CCI Phase 1 (SST)

Title : Product Validation Plan (PVP)

Abstract : This document contains the Product Validation Plan (PVP) for the ESA SST_CCI project

Authors : 
 : Gary Corlett, Chris Merchant,
 Nick Rayner

Approved : 
 : Chris Merchant
 University of Edinburgh
 Science Leader

Accepted : 
 : Craig Donlon
 ESA

Distribution : SST_CCI team members
 Craig Donlon (ESA)

**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	10 Jan 2012	Initial Issue
B		
C		
D	24 Oct 2011	Updated version issued to UoE and MOHC
E	11 Nov 2011	Updated following comments from UoE and MOHC
F	11 Nov 2011	Minor changes prior to submission to ESA
G	14 Nov 2011	First submission to ESA
H	14 Nov 2011	Cosmetic changes by Project Manager
I	11 Jan 2012	Updated to action RIDS raised by ESA TO in ESA-RIDS-SST_cci-PVP-UoL-001-Draft-H-BATCH-1-and-BATCH-2.doc
J	30 Jan 2012	Minor updates following RID review with ESA TO
1	30 Jan 2012	Issue 1 (accepted)
2	04 Feb 2014	Issue 2 – updated with final Phase I product specification and intercomparison datasets

TABLE OF CONTENTS

1. INTRODUCTION.....	6
1.1 Purpose and Scope.....	6
1.2 Structure of the Document	6
1.3 Referenced Documents.....	7
1.4 Acronyms and abbreviations	10
2. DEFINITIONS	13
3. ORGANISATION OF ACTIVITIES.....	14
3.1 Multi-sensor match-up database	15
3.2 Algorithm selection process.....	15
3.3 SST_CCI Products	16
3.4 Uncertainties	17
3.5 Validation and Evaluation	18
3.6 Independence of validation activities.....	18
3.7 Getting Endorsements.....	19
3.8 Release of Products	19
3.9 Master schedule of activities	19
4. REFERENCE DATASET	21
4.1 Introduction.....	21
4.2 Overview of data sources.....	22
4.2.1 SST at approximately 0.2m depth from drifting buoys	23
4.2.1.1 Background.....	23
4.2.1.2 Accuracy	23
4.2.1.3 Stability	24
4.2.2 SST at approximately 1 m depth from moored buoys	26
4.2.2.1 Background.....	26
4.2.2.2 Accuracy	27
4.2.2.3 Stability	28
4.2.3 SST at various depths from Voluntary Observing Ships and Research Vessels	28
4.2.3.1 Background.....	28
4.2.3.2 Accuracy	29
4.2.3.3 Stability	29
4.2.4 SST _{skin} from shipborne radiometers	30
4.2.4.1 Background.....	30
4.2.4.2 Accuracy	31
4.2.5 Near-surface temperature measurements from Argo.....	32
4.2.5.1 Background.....	32
4.2.5.2 Accuracy	33
4.2.5.3 Stability	34
4.2.6 SST at various depths in HadSST3	34
4.3 Criteria for selection	34
4.4 Content of Reference Dataset.....	44
5. MULTI-SENSOR MATCH-UP DATASET	45
5.1 From concept to reality.....	45
5.2 Match-up rules.....	46
5.3 Segregation of drifting buoy match-ups.....	46
5.4 MMD Input data.....	47
5.5 MMD output format.....	47
6. SELECTION OF ALGORITHMS TO BE IMPLEMENTED IN SST_CCI	48
6.1 Purpose and Scope of Algorithm Selection in SST CCI	48
6.1.1 Purpose and Definition of Potential Scope.....	48

6.1.1.1	Observation classification.....	49
6.1.1.2	SST estimation.....	50
6.1.1.3	SST uncertainty estimation.....	50
6.1.1.4	SST product confidence assignment.....	50
6.1.1.5	SST-skin to SST-subskin to SST-depth adjustment.....	51
6.1.1.6	SST time adjustment to 1030 or 2230.....	51
6.1.2	Algorithm types covered in Algorithm Selection.....	51
6.2	Organisation and responsibilities.....	51
6.2.1	Pre-Selection Engagement.....	51
6.2.2	Putting the SST CCI development algorithm outputs in the MMD / RRDP.....	52
6.2.3	Solicit and receive extensions to MMD.....	52
6.2.4	Announcement and dissemination of RRDP.....	53
6.2.5	Round-robin data package consultation (WP 21220).....	53
6.2.6	Algorithm comparison and selection.....	53
6.2.7	Write report and journal paper on algorithm selection.....	54
6.3	Selection criteria and process.....	54
6.3.1	Over-arching principles.....	54
6.3.2	Definition of common metrics.....	55
6.3.2.1	Bias.....	55
6.3.2.2	Non-systematic uncertainty (precision).....	57
6.3.2.3	Stability.....	58
6.3.2.4	Independence from in situ SST.....	58
6.3.2.5	SST sensitivity.....	59
6.3.2.6	Generality.....	60
6.3.2.7	Improvability.....	61
6.3.2.8	Difficulty of implementation.....	61
6.3.3	Selection of SST_CCI algorithms to be implemented.....	62
7.	VALIDATION OF SST_CCI PRODUCTS.....	65
7.1	Endorsement of methods.....	65
7.2	Definitions.....	65
7.3	SST validation.....	65
7.4	Reference data.....	66
7.5	Rules and responsibilities for objective independent product validation.....	66
7.6	Validation criteria.....	67
7.7	Validation confirmation levels.....	67
7.8	Classes of validation.....	67
7.9	Types of validation.....	68
7.10	Analysis procedures.....	69
7.11	Review process and decision sequence.....	69
7.12	Re-validation of upgrades.....	69
8.	SST_CCI PRODUCT INTERCOMPARISON WITHIN THE GMPE.....	70
8.1	Long-term product.....	70
8.2	Demonstration product.....	71
9.	THE SST_CCI CLIMATE ASSESSMENT REPORT.....	72
9.1	Assessment of long-term behaviour of SST_CCI products.....	72
9.1.1	Our analysis.....	72
9.1.2	Engagement of others.....	78
9.2	Assessment of the impact of SST_CCI products on climate model simulations.....	79
9.2.1	Our analysis.....	80
9.2.1.1	Time means.....	81
9.2.1.2	Variability.....	82
9.2.2	Engagement of others.....	83
9.3	Assessment of the consistency of SST_CCI products with other CCI ECVs.....	84
9.3.1	Our analysis.....	85
9.3.2	Engagement of others.....	85

APPENDIX A	ASSESSMENT OF USER REQUIREMENTS	86
APPENDIX B	ADHERENCE TO CCI PROJECT GUIDELINES	89
APPENDIX C	ROUND-ROBIN PROTOCOL	92
C.1	Participation	92
C.1.1	Who can participate?	92
C.1.2	What do I gain from participating?	92
C.1.3	What am I expected to contribute?	92
C.1.4	What commitment do I give?	93
C.1.5	What happens next?	94
C.1.6	How will progress and results be reported?	94
C.1.7	Will the results and data be made public?	94
C.1.8	What if my sensor is not in the round robin data package?	95
C.2	Schedule	95
C.2.1	What are the time scales?	95
C.3	Experiment Design and Selection criteria	96
C.3.1	Experiment Design	96
C.3.2	Selection Criteria	97
C.4	Data	98
C.4.1	What is in the round robin data package?	98
C.4.2	How do I get the round robin data package?	98
C.4.3	What data do I have to deliver?	98
C.4.4	How do I submit my data?	99
C.4.5	Format specification of participant contributions	99
C.5	Important Contacts	100

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 is holding an open algorithm selection exercise. This consists of algorithm intercomparison (described in ESA documents as the "Round Robin", (RR)) followed by selection of algorithms following criteria defined in this document. The chosen algorithm(s) will then be implemented in an end-to-end system to generate the first SST_CCI data records. It is expected that future algorithm selection exercises will be carried out for each subsequent reprocessing to ensure the best performing algorithm is always implemented.

Following selection and implementation the SST_CCI L2, L3 and L4 products will be independently validated using high quality SST measurements made in situ from a number of sources. In addition the SST_CCI L4 products will be compared to other L4 products as part of the Group for High Resolution SST (GHRSSST) Multi Product Ensemble (GMPE) and other inter-comparisons carried out as part of the Climate Assessment Report (CAR). The CAR will also include other kinds of assessment, as detailed in this document.

1.1 Purpose and Scope

This document summarises the SST_CCI product validation plan (PVP). It describes the approach to algorithm selection, product validation, intercomparison and climate assessment for the SST_CCI products.

1.2 Structure of the Document

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

2 DEFINITIONS

This section defines key terms used within this document.

3 ORGANISATION OF ACTIVITIES

This section provides a summary of the algorithm selection, product validation, intercomparison and climate assessment activities described in this document.

4 REFERENCE DATASET

This section describes the content of the reference dataset for validation and climate assessment.

5 multi-sensor match-up dataset

This section summarises the multi-sensor match-up dataset.

6 SELECTION OF ALGORITHMS TO BE IMPLEMENTED IN SST_CCI

This section summarises the approach to selecting the algorithms for producing SST_CCI data products.

7 VALIDATION OF SST_CCI PRODUCTS

This section describes the procedures for product validation.

8 SST_CCI PRODUCT INTERCOMPARISON WITHIN THE GMPE

This section describes the procedures for product intercomparison.

9 THE SST_CCI CLIMATE ASSESSMENT REPORT

This section describes the procedures for climate assessment.

APPENDIX A ASSESSMENT OF USER REQUIREMENTS

This section summarises user requirements related to algorithm selection, product validation, intercomparison and climate assessment, and how each one has been addressed within this document.

APPENDIX B ADHERENCE TO CCI PROJECT GUIDELINES

This section summarises the adherence to the relevant CCI project guidelines that were defined at the first CCI collocation.

APPENDIX C ROUND-ROBIN PROTOCOL

This section contains the round robin protocol for the algorithm selection exercise.

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:

- 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). *Journal of Atmospheric and Oceanic Technology*, 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, *Journal of Atmospheric and Oceanic Technology*, 21, 268-283.
- RD.052 Minnett, P. J., et al. (2001). The marine-atmospheric emitted radiance interferometer: A high-accuracy, seagoing infrared spectroradiometer, *Journal of Atmospheric and Oceanic Technology*, 18(6), 994-1013.
- RD.058 Lumpkin, R., and Pazos, M.: Measuring surface currents with Surface Velocity Program drifters: the instrument, its data, and some recent results. In: *Lagrangian Analysis and Prediction of Coastal and Ocean Dynamics (LAPCOS)*, ed. A. Griffa, A. D. Kirwan, A. J. Mariano, T. Ozgokmen, and T. Rossby, 500pp
- RD.074 Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., Kaplan, A., 2003, Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century *J. Geophys. Res.* Vol. 108, No. D14, 4407 10.1029/2002JD002670

- RD.076 Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S., Schlax, M.G., 2007, Daily High-Resolution-Blended Analyses for Sea Surface Temperature. *J. Climate*, 20, 5473–5496, doi: 10.1175/2007JCLI1824.1
- RD.079 Smith, T., R. Reynolds, T. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006), *Journal of Climate*, 21 (10), 2283–2296, doi:10.1175/2007JCLI2100.1.
- RD.081 Kaplan, A., Cane, M.A., Kushnir, Y., Clement, A.C., Blumenthal, M.B., Rajagopalan, B., Analyses of global sea surface temperature 1856-1991, *Journal of Geophysical Research-Oceans*, 103, C9, 18567-18539, 1998
- RD.085 Ishii, M., Shouji, A., Sugimoto, S., Matsumoto, T., 2005, Objective Analyses of Sea-Surface Temperature and Marine Meteorological Variables for the 20th Century using ICOADS and the KOBE Collection. *Int. J. Climatol.*, 25, 865-879.
- RD.099 Berry, D. I., Kent, E.C., 2009. A New Air–Sea Interaction Gridded Dataset from ICOADS With Uncertainty Estimates. *Bull. Amer. Meteor. Soc.*, 90, 645–656
- RD.112 Brasnett, B. (2008). The impact of satellite retrievals in a global sea-surface-temperature analysis. *Q. J. R. Meteorol. Soc.*, 134: 1745-1760. DOI: 10:1002/qj.319.
- 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 ESA Climate Change Initiative phase 1 – scientific user consultation and detailed specification – statement of work, Issue 1.4, Revision 1, 09/11/2009, Reference EOP-SEP/SOW/0031-09/SP
- RD.169 ESA CCI Project Guidelines V1, EOP-DTEX-EOPS-SW-10-0002, Issue 1, Revision 0
- RD.171 SST_CCI User Requirements Document, SST_CCI-URD-UKMO-001
- RD.172 SST_CCI Data Access Requirements Document, SST_CCI-DARD-UoL-001
- RD.173 SST_CCI Product Validation Plan, SST_CCI-PVP-UoL-001
- RD.175 SST_CCI Product Specification Document, SST_CCI-PSD-UKMO-001
- RD.191 Bureau International des Poids et Mesures, Guide to the Expression of Uncertainty in Measurement (GUM), JCGM 100:2008, 2008. Available online at <http://www.bipm.org/en/publications/guides/gum.html>
- RD.207 Thiebaut, J., E. Rogers, W. Wang, B. Katz, A new High-resolution blended Real-Time Global Sea Surface Temperature Analysis, *Bulletin of the AMS*, 84, 645-656,. 2003 journals.ametsoc.org/doi/pdf/10.1175/BAMS-84-5-645
- RD.208 Gemmill, W., B. Katz, and X. Li, Daily Real-Time, Global Sea Surface Temperature High-Resolution Analysis: RTG_SST_HR, Technical Note Nr. 260 2007.
- 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. *J. Geophys. Res.*, 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. *J. Geophys. Res.*, 116, D14104, doi:10.1029/2010JD015220
- RD.212 Reynolds, R. W., Rayner, N.A., Smith, T.M., Stokes D.C., Wang, W., 2002, An improved in situ and satellite SST analysis for climate. *J. Climate*, 15, 1609-1625

- RD.213 Donlon, C.J., M. Martin, J. D. Stark, J. Roberts-Jones, and E. Fiedler (2012), The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA), Remote Sensing of the Environment, 116, 140-158.
- RD.214 Gentemann, C. L., F. J. Wentz and M. DeMaria, Near real time global optimum interpolated microwave SSTs: applications to hurricane intensity forecasting, paper presented at 27th conference on hurricanes and tropical meteorology, Monterey, CA, 2006.
- RD.216 Casey, K.S., T.B. Brandon, P. Cornillon, and R. Evans (2010). "The Past, Present and Future of the AVHRR Pathfinder SST Program", in Oceanography from Space: Revisited, eds. V. Barale, J.F.R. Gower, and L. Alberotanza, Springer. DOI: 10.1007/978-90-481-8681-5_16
- RD.217 SST_CCI Round Robin Data Package Specification, SST_CCI-RRDPS-UoL-001
- RD.218 SST_CCI Round Robin Protocol, SST_CCI-RRP-UoL-001
- RD.225 SST CCI Algorithm Theoretical Basis Document – not yet published
- RD.226 SST CCI Product Validation and Algorithm Selection Report – not yet published
- RD.232 SST_CCI Multi-sensor Match-up Specification, SST_CCI-REP-UoL-001
- RD.233 Karspeck, A. et al 2011: Bayesian modelling and ensemble reconstruction of mid-scale variability in North Atlantic SSTs for 1850-2008 QJRM doi:10.1002/qj.900
- RD.234 Minnett, P. J., 1991: Consequences of sea surface temperature variability on the validation and applications of satellite measurements. J. Geophys. Res., 96, 18,475-18,489.
- RD.235 Beggs Helen (2008) GAMSSA – A New Global Australian Multi-Sensor SST Analysis, Submitted to Proceedings of the 9th GHRSSST-PP Science Team Meeting, Perros-Guirec, France, 9-13 June 2008.
http://cawcr.gov.au/projects/SST/GHRSSST9/9th_GHRSSST-PP_Meeting_GAMSSA_paper.doc
- RD.236 Zhong, Aihong and Helen Beggs (2008) Analysis and Prediction Operations Bulletin No. 77 - Operational Implementation of Global Australian Multi-Sensor Sea Surface Temperature Analysis, 2 October 2008.
http://cawcr.gov.au/projects/SST/GAMSSA_BoM_Operational_Bulletin_77.pdf
- RD.237 Kurihara, Y., T. Sakurai, and T. Kuragano (2006), Global daily sea surface temperature analysis using data from satellite microwave radiometer, satellite infrared radiometer and in-situ observations. Weather Bulletin, 73. s1-s18 (in Japanese).
- RD.238 Brasnett, B. (1997). A global analysis of sea surface temperature for numerical weather prediction. J. Atmos. Oceanic Technol. 14: 925-937.
- RD.239 Roberts-Jones, J., E. Fiedler and M. Martin, 2012: Daily, global, high-resolution SST and sea-ice reanalysis for 1985-2007 using the OSTIA system, submitted to J. Climate
- RD.240 Miller, P., 2009: Composite front maps for improved visibility of dynamic sea-surface features on cloudy SeaWIFS and AVHRR data, J. Marine Systems, 78, 327-336.
- RD.241 Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T., 1999: A dipole mode in the tropical Indian Ocean, Nature, 401, 360–363.
- 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, *Remote Sensing of the Environment*, 116, 79-92.
- RD.244 Reverdin G., Boutin J., Martin N., et al., 2010: Temperature Measurements from Surface Drifters, *J. Atmos. Ocean. Tech.*, 27, 1403-1409, doi: 10.1175/2010JTECHO741.1
- 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, *Journal of Geophysical Research*, 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, *J. Atmos. Ocean. Tech.*, 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. *J. Atmos. Oceanic Tech.*, 17(12), pp. 1667–1675
- RD.258 SST_CCI System Requirements Document, SST_CCI-SRD-BC-001 – not yet available
- RD.259 SST_CCI System Specification Document, SST_CCI-SSD-BC-001 – not yet available
- RD.260 He, R., K. Chen, T. Moore and M. Li (2010): Mesoscale variations of sea surface temperature and ocean color patterns at the Mid-Atlantic Bight shelfbreak, *GRL*, 37, doi:10.1029/2010GL042658
- RD.261 Sokolov S. and S.R. Rintoul, 2007, On the relationship between fronts of the Antarctic Circumpolar Current and surface chlorophyll concentrations in the Southern Ocean, *JGR*, 112, doi:10.1029/2006JC004072

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.

Term	Definition
AAI	Aerosol Absorbing Index
AATSR	Advanced ATSR
AMSR-E	Advanced Microwave Scanning Radiometer - EOS
ATSR	Along Track Scanning Radiometer
AVHRR	Advanced Very high Resolution Radiometer
BC	Brockmann Consult
CCI	Climate Change Initiative
CEOS	Committee on Earth Observation Satellites
CMS	Centre de Météorologie Spatiale
CMUG	Climate Modelling User Group

DMI	Danmarks Meteorologiske Institut
DBCP	Data Buoy Cooperation Panel
ECMWF	European Centre for Medium-Range Weather Forecasts
ECV	Essential Climate Variable
EO	Earth Observation
ESA	European Space Agency
FRAC	Full Resolution Area Coverage
GAC	Global Area Coverage
GCOS	Global Climate Observing System
GHR SST	Group for High Resolution SST
GOME	Global Ozone Monitoring Experiment
GMPE	GHR SST Multi-Product Ensemble
GRIB	Gridded Binary file format
GTS	Global Telecommunications System
HDF	Hierarchical Data Format
ICOADS	International Comprehensive Ocean-Atmosphere Data Set
L2	Level 2
L3	Level 3
L3C	L3 collated
L3U	L3 uncollated
L4	Level 4
MD	Match-up Dataset
METOP	Meteorological Operational Satellite
MM	Multi-sensor Match-up
MMD	Multi-sensor Match-up Dataset
MN	Met.No
MOHC	Met Office Hadley Centre
NetCDF	Network Common Data Form
NCEP	NOAA National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NWP	Numerical Weather Prediction
OMI	Ozone Monitoring Instrument
PSD	Product Specification Document
PVASR	Product Validation and Algorithm Selection Report
PVIR	Product Validation and Intercomparison Report
PVP	Product Validation Plan

QA4EO	Quality Assurance for Earth Observation
RFI	Radio Frequency Interference
RR	Round Robin
RRDP	Round Robin Data Package
SCL	Space ConneXions Limited
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SL	Science Leader
SoW	Statement of Work
SST	Sea Surface Temperature
SST_CCI	ESA Climate Change Initiative on SST
TMI	TRMM Microwave Imager
TOMS	Total Ozone Mapping Spectrometer
TRMM	Tropical Rainfall Measuring Mission
UoE	University of Edinburgh
UoL	University of Leicester
UR	User Requirements
URD	User Requirements Document
V	Validation
WGCV	Working Group on Calibration and Validation

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, 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 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 random 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 μm (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 μm .

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 PVP is written on the basis of these definitions.

3. ORGANISATION OF ACTIVITIES

The activities described in this document cover:

- Identification of the best performing retrieval algorithm or combination of algorithms via an open algorithm selection exercise
- Validation of SST_CCI L2, L3 and L4 products, which will be performed independently using high quality SST measurements made in situ from a number of sources
- Other assessments of the new products against other data, referred to as “climate assessment”

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 algorithm selection, product validation, intercomparison and climate assessment is shown schematically in Figure 3-1.

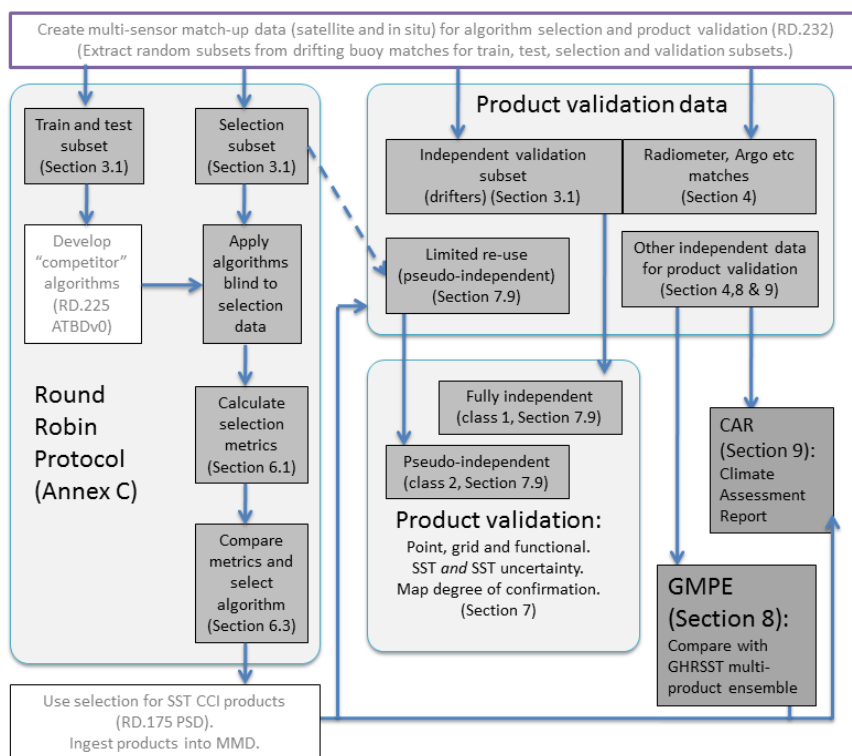


Figure 3-1: Flowchart indicating logical flow of algorithm selection, product validation inter-comparison and climate assessment for the SST_CCI project. Activities and data sets specified in this document are in dark-grey boxes. The top box represents the multi-sensor match-up system which is a key 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 (see Section 3.1 for a brief introduction and Section 5 for further details), which will be the source of subsets of data used for both algorithm selection (see Section 3.2 and Section 6) and product validation (see Section 3.5 and Section 7).

The SST_CCI products will also be compared to other SST analyses (see Section 3.5 and Section 8) and will undergo a climate assessment (see Section 3.5 and Section 9).

A master schedule indicating all key dates such as the release of project reports and products is given in Section 3.8.

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 will pre-match all required in situ data to the set of satellite datasets required for the two different categories of output products (see Section 3.3). The in situ data comprises data from drifting and moored buoys, Argo float, VOS 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].

Each drifting buoy match-up is assigned to one of four categories using the methodology described in Section 5.3:

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

All other multi-sensor match-ups are assigned to the last category, validation. A final subset of the independent validation match-ups, referred to as the reference dataset, will be selected later in the project.

Further details on the generation of the MMD files, including the approach to segregating the drifting match-ups, are given in Section 5.

3.2 Algorithm selection process

The retrieval algorithms will be selected by an open round robin (RR) algorithm selection exercise. Details of the RR phase of the algorithm selection exercise, which includes what is involved, how to participate, how to download the RR dataset, important dates are and who to contact, are given in Section APPENDIX C. The round robin protocol was defined by the validation team and will use datasets generated by the system engineering team using a specification defined by the validation team. Overall management of the RR is the responsibility of the validation team lead, Gary Corlett.

The final step of the algorithm selection process is an evaluation of each submitted algorithm according to a set of predefined criteria. Further details on the algorithm selection criteria are given in Section 6. The final choice of algorithms may require an

element of subjective decision making and may not be solely on the basis of the criteria defined in Section 6. Consequently, the final choice of algorithms will be the responsibility of the algorithm selection team, led by the Science Leader, Chris Merchant.

All of the decisions, final selection criteria and results from the algorithm selection exercise will be available via the Product Validation and Assessment Report, PVASR, RD.226, which will be available by end of July 2012. A detailed description of the final algorithm will be provided in v1 of the Algorithm Theoretical Basis Document, ATBD, RD.225, and the all data will be made available via an SFTP site.

3.3 SST_CCI Products

Following an extensive user requirements review (URR), which is summarised in RD.171, the ESA SST_CCI project will provide 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.
2. Demonstration ECV products, based on wider use of the modern observing system to increase completeness and/or frequency of coverage

A summary of the output products for each category 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 to be selected, validated and evaluated:

- 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)
<u>Long term ECV</u> 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	L4 (0.05°) L2P (variable, ~4 km at centre of swath)
<u>Demonstration ECV</u> 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)	L4 (0.05°) 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.

3.4 Uncertainties

A key development within the SST_CCI project is to provide 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 spatiotemporal 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).

The uncertainty information attached to SSTs constitutes part of the product, in this approach, and will be validated in its own right.

In all cases, users should exploit the uncertainty information provided within the SST_CCI products within their particular data application and should not use external comparisons to other datasets to estimate uncertainties for SST_CCI products.

3.5 Validation and Evaluation

After production and system verification tasks have been completed the SST_CCI products will undergo validation and evaluation by members of the project team. The process comprises:

Product validation: Independent product validation will be done by the validation team using data not made available for algorithm development or selection. The approach to validating each SST_CCI product is described in Section 7. A key point regarding the validation is that as the products will contain uncertainties then the validation will be used to confirm both the SST and its associated uncertainty and will not be used to derive uncertainty information, which is the more traditional way of deriving uncertainty for satellite derived SST datasets.

Product intercomparison: Intercomparison of SST_CCI products with other satellite based L4 analyses will be carried out by the MOHC by adding the SST_CCI L4 products into the GHRSSST GMPE system. Further details on the inter-comparison of SST_CCI products are given in Section 8.

Climate assessment: The usefulness of the SST_CCI products for climate will be assessed by the climate team and a number of external parties. The initial climate assessment of the SST_CCI products is summarised in Section 9.

The results of the product validation and intercomparison will be presented in the Product Validation and Intercomparison Report (PVIR) and will be submitted for peer review in the scientific literature. The results and findings of the initial climate assessment will be published in the Climate Assessment Report (CAR) and will also be submitted for peer reviewed publication.

3.6 Independence of validation activities

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

Hoeyer (DMI) will contribute a tuned high latitude retrieval algorithm to the algorithm selection process, but will not have the final say if decide if the algorithm is implemented or not, and will also be involved with product validation at high latitudes. Consequently, true independence of all steps will only be compromised if, and only if, the tuned high latitude algorithm is actually implemented. However, in any event the final validation and evaluation steps will still be carried out by other independent personnel.

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)			✓		
Martin (MOHC)				✓	
Rayner (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

3.7 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 will seek endorsement of our methods through external peer review of this document and through submission of journal articles summarising our findings.

Within the CCI programme this document will be reviewed by the CMUG and we will seek external review outside of the CCI programme by the GHRSSST RAN-TAG, ST_VAL, and at the next GHRSSST meeting.

3.8 Release of Products

The SST_CCI products will be openly released (subject to any CCI data policy) as soon as the PVIR and the CAR are accepted by ESA.

3.9 Master schedule of activities

The overall schedule of the project can be summarised through a number of key dates and the release of key project deliverables. These dates and deliverables are:

- Project kick-off
 - 1st August 2010
- User assessment period
 - 1st August 2010 to 31st January 2011
 - URDV2 released on 30th November 2010
- Product specification
 - 1st February 2011 to July 2011
 - PSDv1 released on 1st April 2011
 - PSDv2 released on 23rd June 2011
- Round robin algorithm selection exercise
 - September 2011 to April 2012 *
 - PVP released on 11th January 2012
 - PVASR released on 30th April 2012 *
 - ATBDv0 released on 11th January 2012
- System prototyping
 - November 2011 to June 2012 *
 - SPDv1 released on 30th September 2012 *
 - SVR released on 31st March 2013 *
 - DPMv1 released on 30th June 2012 *
 - IODDv1 released on 30th June 2012 *
 - SRDv1 released on 31st January 2012 *
 - SSDv0 released on 30th June 2012 *
- Product generation
 - June 2012 to March 2013 *
 - Prototype products available on 31st March 2013 *
 - PUG released on 31st August 2012 *
- Product validation and intercomparison
 - May 2013 to June 2013 *
 - PVIR released on 30th September 2013 *
- Climate assessment
 - July 2013 to September 2013 *
 - CAR released on 30th September 2013 *
- Public release of SST_CCI products
 - 31st October 2013 *

* Estimated date(s) based on the expected SST_CCI project schedule at the time of the release of this document.

4. REFERENCE DATASET

4.1 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

The remainder of 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 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 always traceable to an 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 TAO/TRITON/PIRATA arrays are 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.

As well as the individual in situ measurements we will also use the HadSST3 1° gridded product to increase the coverage of available data for validation. HadSST3 comes with quantified uncertainties in the form of multiple, equally-likely realizations of each gridded value and other uncertainty information with covariances. Where data are sparse, this will

be invaluable because the uncertainties can be properly accounted for in the comparisons to the SST_cci data.

Although HadSST3 comprises measurements from different platforms and therefore from different depths below the surface, Kennedy et al. (2011b; RD.211) applied bias adjustments to the gridded anomalies to create an homogeneous record. Adjustments were applied that accounted for the bias in each measurement type from the “true” SST (see Kennedy et al, 2011b, RD.211, for details) and the composition of the measurements contributing to each gridded average. Because there are uncertainties in these adjustments, HadSST3 is presented as an ensemble of 100 equally-likely data sets which allows the user, in principle, to repeat their analysis up to 100 times to assess the effect of the uncertainty on their results. No specific reference depth is used in the adjustments, but HadSST3 is consistent with measurements made at a few cm depth, e.g. those made by drifting buoys. Other components of uncertainty in HadSST3 arising from under-sampling of grid boxes and residual biases in individual measurement platforms are provided as error fields and error covariance matrices.

Special attention will be given to the Arctic and Antarctic regions. Because of the expected scarcity of matches to in situ measurements in these regions, additional ‘dummy’ match-ups have been created within these areas. A dummy match-up is one with no in situ data and is solely for satellite/satellite intercomparisons. This allows at least some inter-comparison of the output products even if they cannot be matched to an in situ measurement. Target research cruise data from synoptic ships observations of SST, ice coverage and cloud coverage will be used as reference data in these areas if available.

Recent developments within GHRST include how to use data from extremely stable satellite instruments such as AATSR as a reference data source for other satellite sensors, as AATSR provides data that has a lower uncertainty than the current GHRST reference dataset (O’Carroll et al., 2008; RD.246). However, as the ATSR series plays a crucial role within the project as part of the algorithm selection, the reference dataset will not include satellite data and will comprise solely in situ and other surface measurements.

4.2 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, uncertainties are divided into 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 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.

4.2.1 SST at approximately 0.2m depth from drifting buoys

4.2.1.1 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-anchor spanning the upper 10 to 15 m of the water column, which allows the buoy to follow currents in the ocean mixed layer (Lumpkin and Pazos, 2006; RD.058). The SST sensor is embedded in the underside of the buoy and measures at a depth of approximately 25cm 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.

4.2.1.2 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).

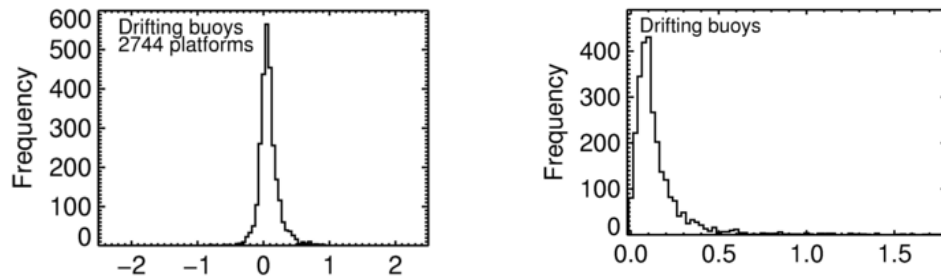


Figure 4-1: Distribution of inter-platform (left) and intra-platform (right) uncertainty for drifting buoys between August 2002 and December 2007. Only platforms with more than 25 ATSR-drifter pairs are shown. (Figure adapted from Kennedy et al, 2012; RD.243.)

A range of intra- and inter-drifter uncertainties was found. The inter-platform uncertainties exhibited a very peaked distribution with standard error of about 0.29K. The intra-platform uncertainties 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.

4.2.1.3 Stability

A recent study has examined differences between two 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.

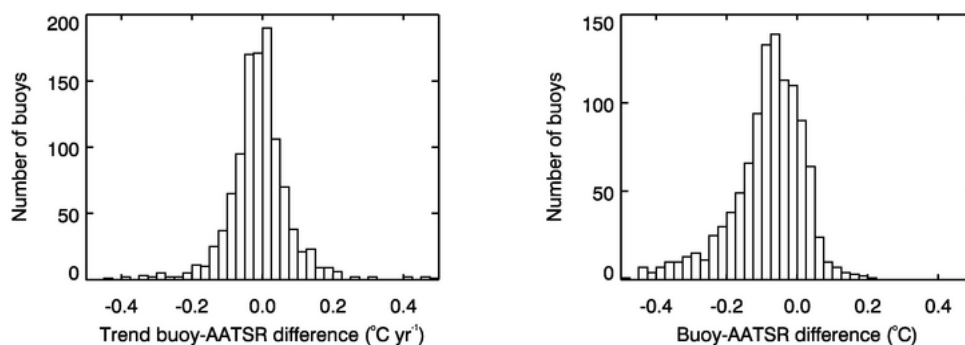


Figure 4-2: Comparisons between SST measured by drifting buoys over their lifetime and retrieved from AATSR measurements. Left: relative trends ($^{\circ}\text{C}/\text{yr}$). Right: average differences ($^{\circ}\text{C}$).

Rob Smith (personal communication) 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.

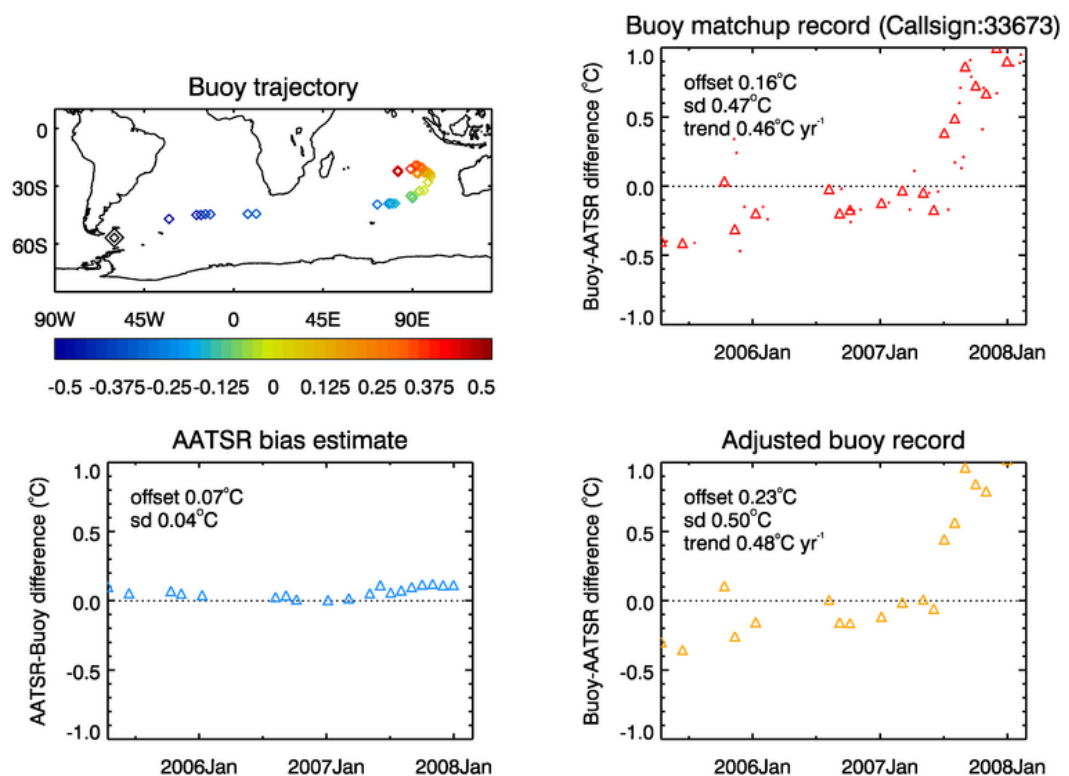


Figure 4-3: Calibration drift of buoy 33673, 2005-2007. Top left: trajectory of the buoy through its lifetime (about 2.5 years). Top right: difference between SST as measured by the buoy and as retrieved from coincident AATSR measurements. Bottom left: estimate of AATSR SST bias from comparison with 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 will be excluded from the SST_CCI reference data set. Methods used in creating blacklists, such as those maintained by Météo France and the Met Office, will also be utilised to exclude erroneous measurements.

4.2.2 SST at approximately 1 m depth from moored buoys

4.2.2.1 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.icommops.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.

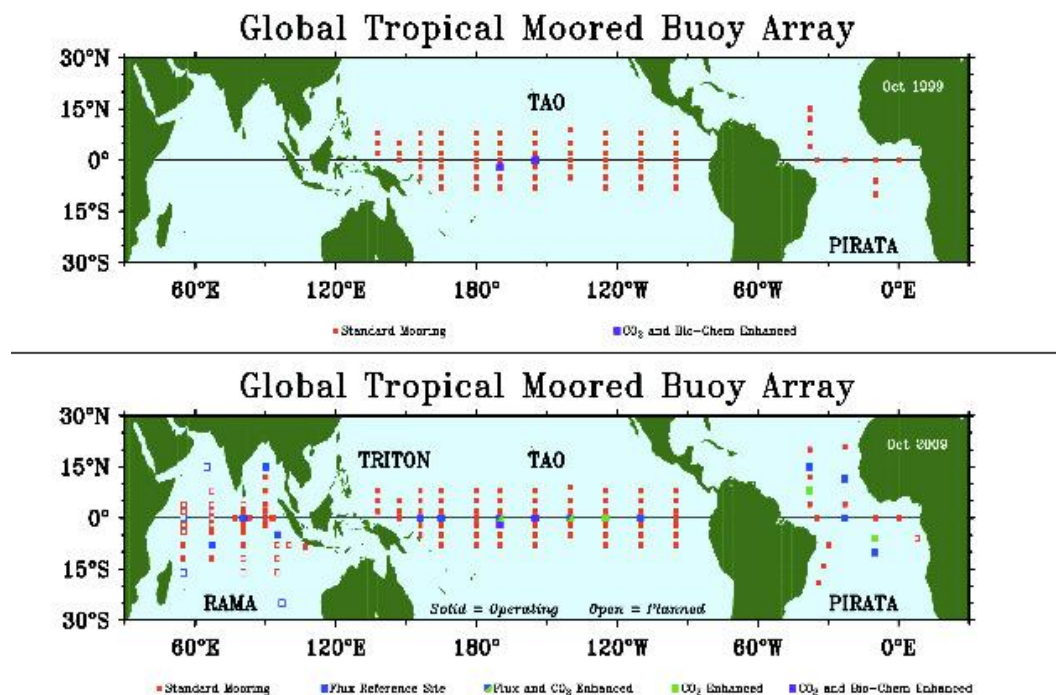


Figure 4-4: The evolving GTMBA, showing both existing and planned moorings (as of 2009). The upper panel shows the arrays as they existed in 1999; the lower panel shows the arrays in 2009 (solid circles) plus planned additions (open circles). (Taken from <http://www.atmos.washington.edu/~ackerman/GTMBA.pdf>)

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 uncertainties (Figure 4-5 and Figure 4-6).

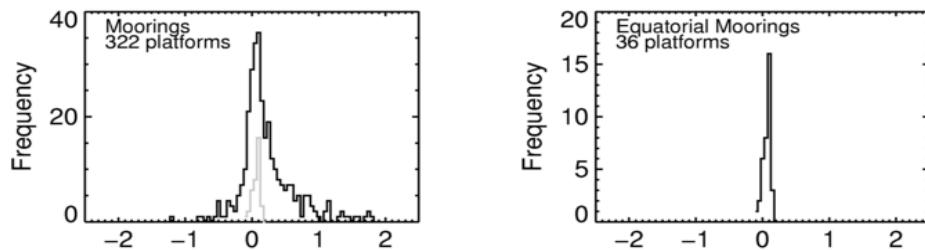


Figure 4-5: Distributions of inter-platform uncertainty for moorings (left, equatorial moorings are shown in grey) and equatorial moorings only (GTMBA, right) between August 2002 and December 2007. Only platforms with more than 25 ATSR-*in situ* pairs are shown. (Figure adapted from Kennedy et al, 2012;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

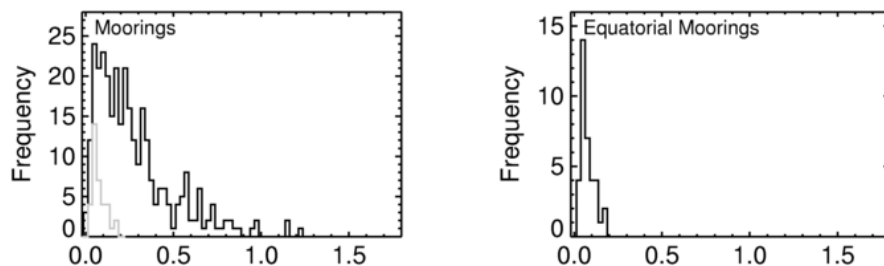


Figure 4-6: Distributions of intra-platform uncertainty for moorings (left, equatorial moorings are shown in grey) and equatorial moorings (GTMBA, right) between August 2002 and December 2007. Only platforms with more than 25 ATSR-*in situ* pairs are shown. (Figure adapted from Kennedy et al, 2012; RD.243.)

Currently, uncertainties are not available for each mooring.

4.2.2.3 Stability

As part of the ARC project, National Oceanography Centre, Southampton (NOCS) have developed a list of moored buoys found to contain discontinuities, which will be used to exclude the least stable moored buoys.

4.2.3 SST at various depths from Voluntary Observing Ships and Research Vessels

4.2.3.1 Background

The Voluntary Observing Ship (VOS) Scheme is an international programme comprising member countries of the WMO that recruit ships to take, record and transmit weather observations, including sea surface temperature, whilst at sea. The repository for VOS data is the International Comprehensive Ocean-Atmosphere Data Set (ICOADS). Measurements of SST from VOS are available from before 1850 onwards. ICOADS comprises a number of "decks", which include all the measurements obtained from a particular collection or source.

The size of the VOS fleet peaked around 1985, when there were more than 7500 ships on the World Meteorological Organisation's VOS fleet list. Numbers have declined since, with fewer than 4000 ships remaining on the list today (http://www.vos.noaa.gov/vos_scheme.shtml).

Historically, SST measurements have been made using a combination of methods: using a bucket to haul a sample of water on board in order to take a temperature reading; noting the temperature of water as it comes into the engine room to cool the ships' engines and via a dedicated hull contact sensor.

Ship's observations are made at the standard synoptic hours of 0000, 0600, 1200 and 1800 UTC and have been for the last several decades.

Each ship could be identified using its call sign; where present, call sign information is recorded in ICOADS metadata. However, for many historical reports, this information is absent. Updates to ICOADS v2.0 were taken from the NCEP GTS stream. Call sign information was also recorded in GTS reports until November 2007. After this date, callsign information was removed from the NCEP GTS data stream owing to concerns about ship security. The Met Office GTS archive contains call sign information for many ships and we are working to restore this information to the reference data set used in the SST_CCI.

VOSclim is class of VOS. VOSclim aims to provide a high-quality subset of marine meteorological data, with extensive associated metadata, to support global climate studies (<http://wf.ncdc.noaa.gov/oa/climate/vosclim/about.html>). This class was designed to provide ground truth for calibrating satellite observations and to provide a high quality reference data set for possible re-calibration of observations from the entire VOS fleet. VOSclim first went operational in 2001. The list of participating ships is available from <http://wf.ncdc.noaa.gov/oa/climate/vosclim/shipinfo.html>, along with their dates of recruitment and withdrawal (where applicable) from the project.

VOSclim data can be retrieved from ICOADS v2.5 through 2007 (deck 700) and downloaded from <http://lwf.ncdc.noaa.gov/oa/climate/vo clim/vo climdata.html> and <http://www7.ncdc.noaa.gov/CDO/CDOMarineSelect.jsp> thereafter.

ICOADS deck 740 contains quality controlled SST measurements from research vessels, sourced from the Research Vessel Surface Meteorology Center (RVSMDC), located at the Center for Ocean-Atmospheric Prediction Studies at Florida State University. The RVSMDC archive contains data from over 60 research vessels, including those operating at high latitudes, such as the R/V Polarstern (<http://www.coaps.fsu.edu/RVSMDC/>).

In addition, ICOADS deck 735 contains Russian R/V data.

4.2.3.2 Accuracy

Kennedy et al (2012; RD.243) utilised coincident match ups between VOS SST measurements and SST retrieved from ATSR measurements (adjusted to sub-skin depth) for 2002-2007 to assess inter- and intra-ship uncertainties (Figure 4-7).

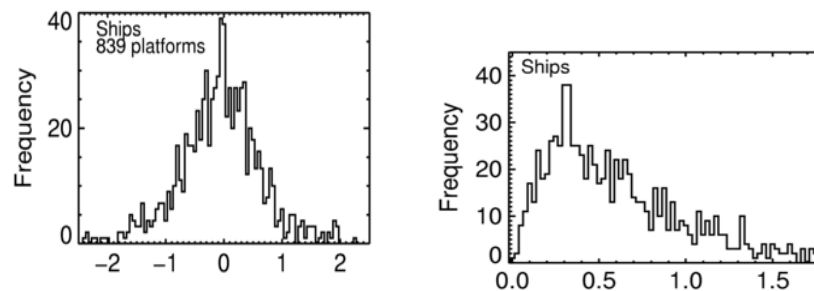


Figure 4-7: Distributions of inter-platform (left) and intra-platform uncertainty for ships between August 2002 and December 2007 (right). Only platforms with more than 25 ATSR-*in situ* pairs are shown. (Figure adapted from Kennedy et al, 2012; RD.243.)

The range of intra- and inter-platform errors is much wider than for drifting or moored buoys, with some ships exhibiting large errors. Correlated intra-ship errors have a large effect on the uncertainty of area averages as ships travel long distances.

The relative accuracy of VOSclim has not yet been assessed. It is not expected that the measurements from its first decade would be of significantly better quality than the wider VOS, as the same measurement methods have generally been used. However, a large portion of the VOSclim fleet now uses Automatic Weather Systems with dedicated hull sensors, which are expected to produce more accurate measurements, when correctly sited (Sarah North, personal communication).

Currently, uncertainties are not available for each VOS.

4.2.3.3 Stability

Instrumentation used by VOS to gather sea surface temperature measurements has changed over time and differs between ships recruited by different nations. This causes both sudden and slowly varying relative biases to be introduced into SST measurements

in ICOADS. Currently, no adjustments have been calculated which can be applied to individual measurements, but relative biases between measurement types have been assessed and adjustments developed for gridded data sets (see also Section 4.2.6).

By retaining metadata on measurement type in the reference data set, it will be possible to utilise information on expected relative biases between VOS SST measurements and satellite retrievals to inform the comparisons.

As for drifting buoys, work underway under the FP7 project ERA-CLIM will assess the VOS SST measurements throughout each ship's record and flag those portions of ships' records which are unsuitable for our reference data set.

4.2.4 SST_{skin} from shipborne radiometers

4.2.4.1 Background

There are a number of infrared radiometers, designed to measure SST_{skin} 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_{skin} 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 ±0.1 K root mean squared error for deployment periods of up to 3 months. It uses two precision calibration blackbody cavities (Theocharus et al, 2010; RD.242). 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_{skin} 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 µm;
- 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 µm;
- the Calibrated Infrared In situ Measurement System (CIRIMS), with a design accuracy of ±0.1 K, passing radiation with wavelengths 9.6–11.5 µm and
- the DAR011 radiometer, a single-channel, self-calibrating, infrared radiometer passing radiation with wavelengths between 10.5 and 11.5 µm.

Ship radiometers do not provide global coverage. However, they are complementary to reference measurements at depth because they provide direct measurements of SST_{skin}.

4.2.4.2 Accuracy

Intercomparisons between the SST_{skin} 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 too largely within $\pm 0.1K$ of each other. Figure 4-8 is taken from Barton et al (2004; RD.050).

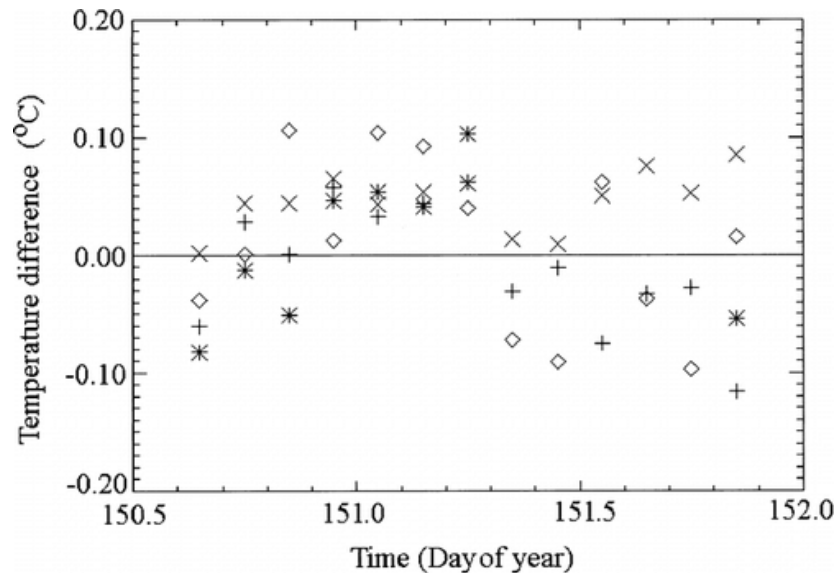


Figure 4-8: The differences between the M-AERI skin SST and those derived using the other radiometers averaged over the intercomparison period: ISAR-5, *; SISTeR, x; JPL, and DAR011, + Reproduced from Barton et al (2004; RD.050), their Figure 5 and Figure 4-9 from Theocharus et al (2010; RD.242).

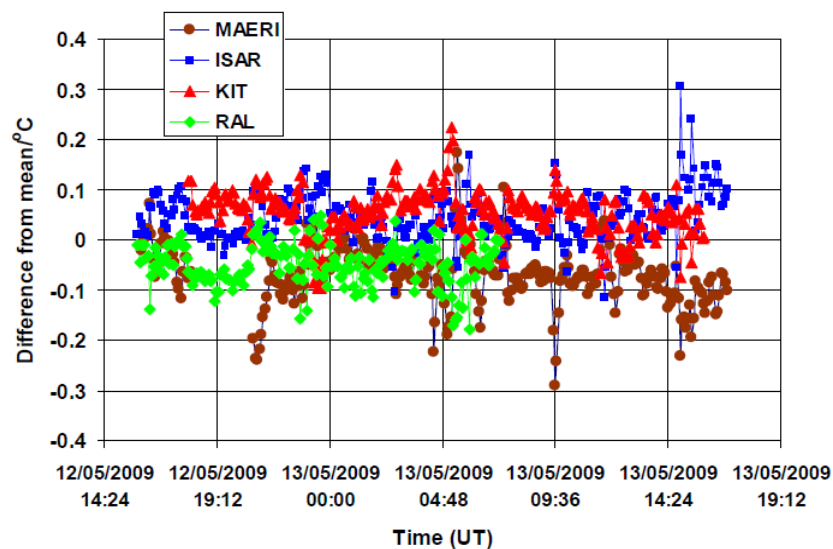


Figure 4-9: Difference of SST measured by M-AERI, ISAR, KIT and SISTeR from their mean. Taken from Theocharus et al (2010; RD.242), their Figure 3.11.9.

Currently, uncertainties are not available for each individual radiometer.

4.2.5 Near-surface temperature measurements from Argo

4.2.5.1 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 Met Office Hadley Centre 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 4-10 shows how the coverage of the array has increased since 2001.

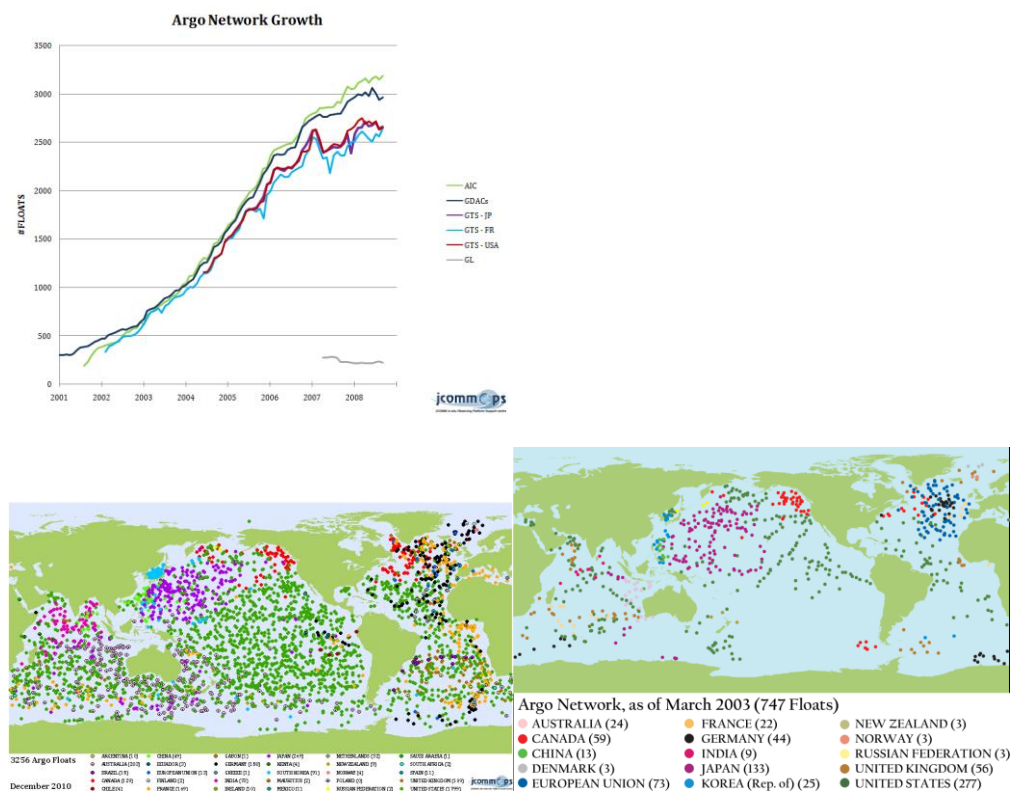


Figure 4-10: The evolution of the Argo array. Top: the number of active floats, 2001-2008. Bottom: snapshots of the Argo network in March 2003 (left) and December 2010 (right). Source: <http://wo.jcommops.org/cgi-bin/WebObjects/JCOMMOPS>.

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 4-11.

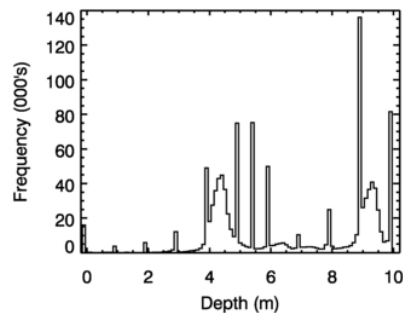


Figure 4-11: Distribution of sampling depths over the upper 10m for Argo profiling floats, from the EN3 data set for the period 2000-2009.

4.2.5.2 Accuracy

Coincident near-surface measurements from drifting buoys and Argo from 2000-2009 were examined (Figure 4-12, Rob Smith, personal communication). 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.

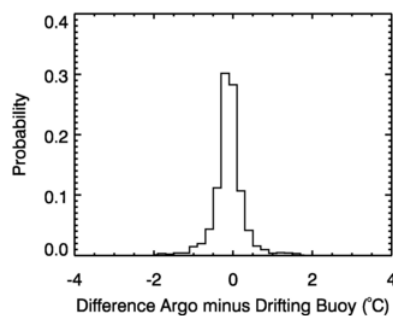


Figure 4-12: Distribution of Argo-minus-drifting buoy differences from co-incident (within 10 km and 1 hour) observations (4-6 m, shallowest selected) 2000-2009.

Figure 4-12 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). 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.

4.2.5.3 Stability

In addition to the pressure bias issues noted above, 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.

4.2.6 SST at various depths in HadSST3

Unlike the previous reference data sources, which are collections of individual observations, HadSST3 is a gridded data set of average SST anomaly relative to 1961-90. It comprises quality controlled, bias adjusted, gridded monthly averages of ICOADS v2.5 SST measurements in 1 or 5° latitude by longitude grid boxes. It is not independent from the data sources outlined in Sections 10.2.1-10.2.3, but has been adjusted to reduce the effects of changing relative biases between measurement types (Kennedy et al, 2011a, RD.210, <http://www.metoffice.gov.uk/hadobs/hadsst3/>).

HadSST3 is the only currently available SST data set to contain adjustments for the effects of changing measurement types through the whole record since 1850. Other SST data sets currently available do not adjust for the effects of database changes after 1941. In particular, the period since 1991 has been a time of revolution in the in situ observing array and needs careful adjustment.

Each grid box average is presented with uncertainty information. HadSST3 is presented as an ensemble of many equally-likely realisations which span the uncertainty in the assumptions made when deriving the bias adjustments. The random, uncorrelated sampling and measurement uncertainty is presented separately from the random, correlated measurement uncertainty which arises from residual biases of individual ships relative to the adjustments calculated for each grid box average (Kennedy et al 2011b; RD.211).

4.3 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, lack of 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.

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. If insufficient buoys are available, VOS using buckets are selected, followed by VOS using ERI.

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.

The following figures show random months at different times between 1991 and 2010 and are intended to provide an indication only of the numbers of in situ measurements available at those times on the 100km scale. They highlight the scarcity of drifting buoy measurements in the early 1990s. They are an over-estimate of possible matchup availability, because they neglect the need for coincident times of measurement.

We could add monthly HadSST3 anomalies as a fifth choice in our hierarchy on a 1° grid. This ignores the numbers of observations available and the measurement method and assumes that the HadSST3 bias adjustments have been effective and the uncertainty estimates are good. Then, we get the following set of figures.

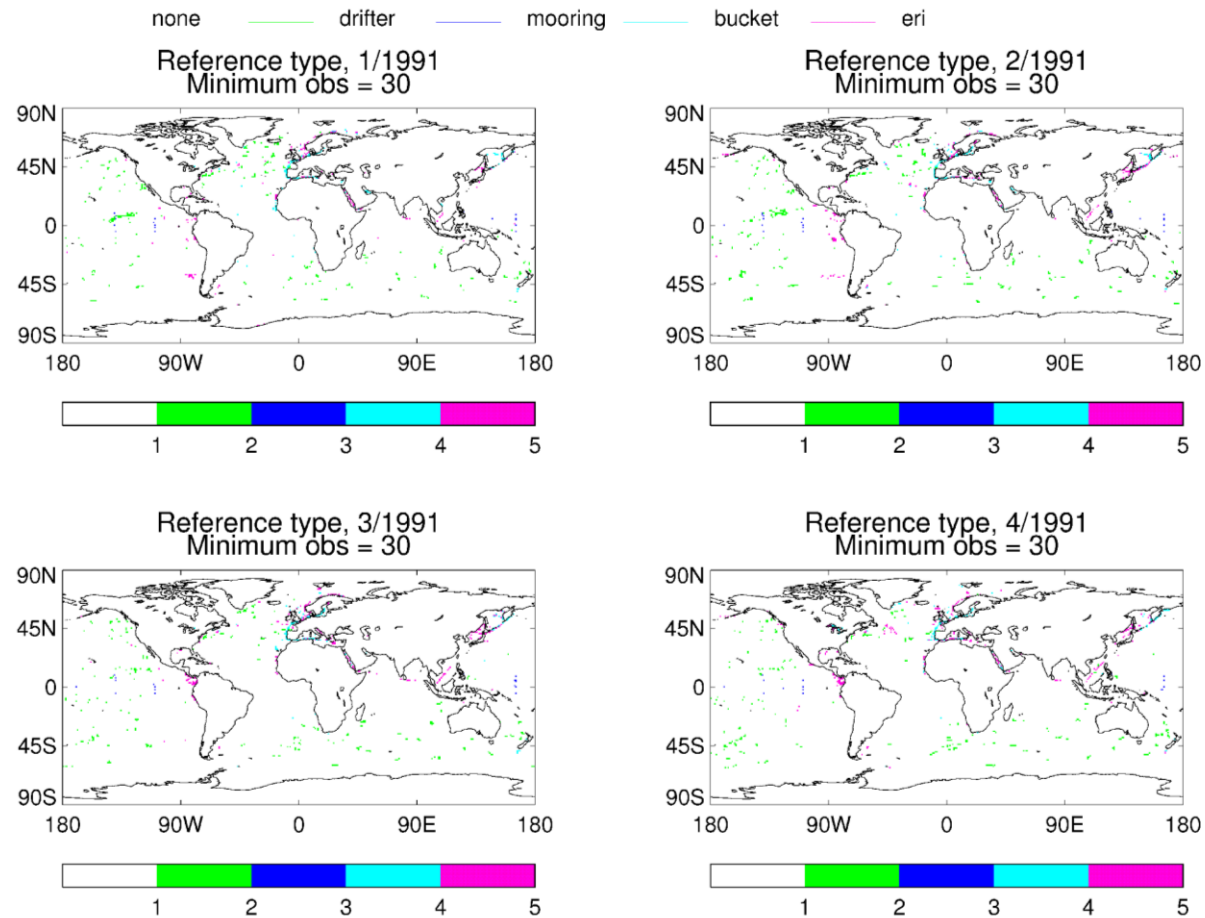


Figure 4-13: Reference types available with greater than 30 measurements in each month in 1° grid boxes. Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan) and VOS with ERI (pink). Months are: 1/1991, 2/1991, 3/1991 and 4/1991.

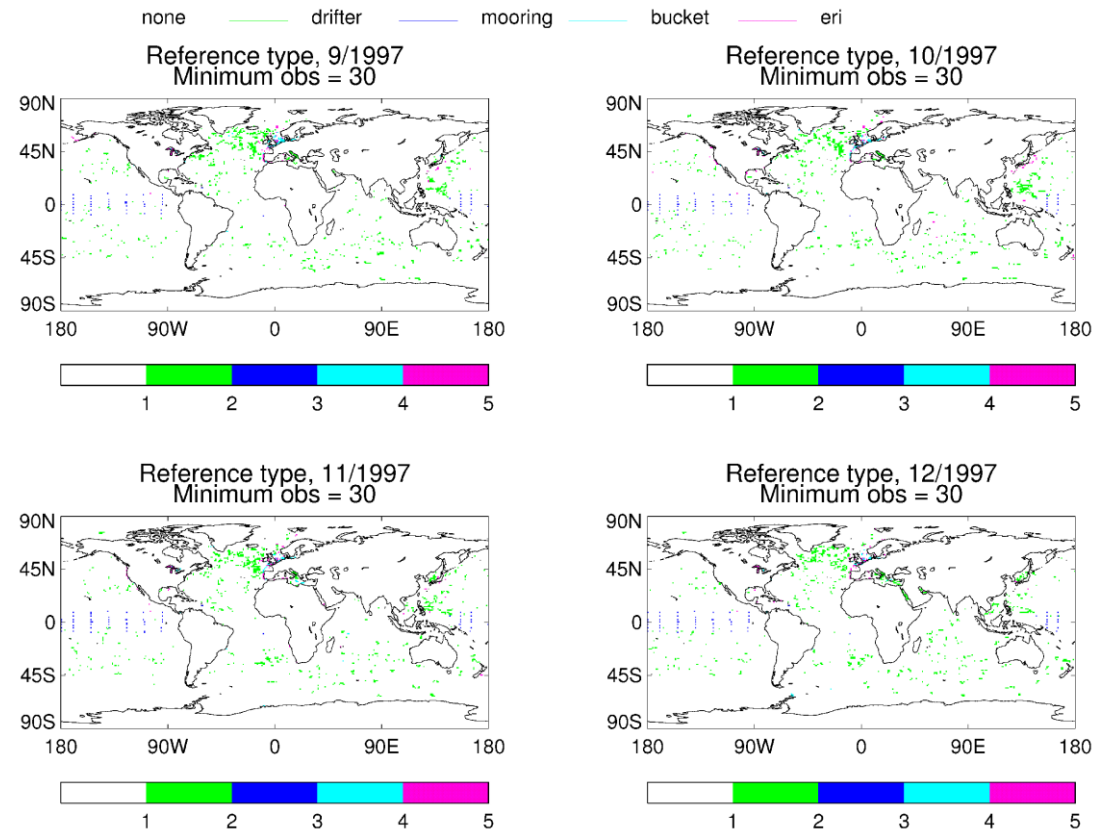


Figure 4-14: Reference types available with greater than 30 measurements in each month in 1° grid boxes. Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan) and VOS with ERI (pink). Months are: 9/1997, 10/1997, 11/1997 and 12/1997.

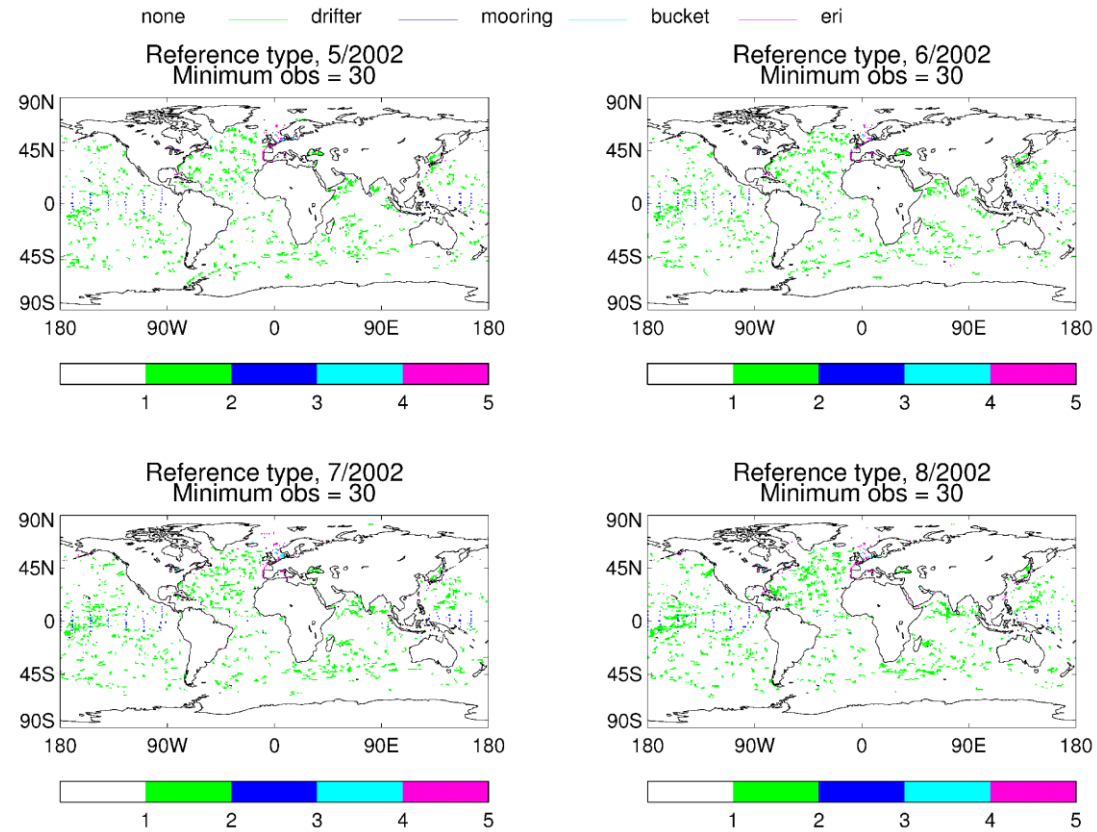


Figure 4-15: Reference types available with greater than 30 measurements in each month in 1° grid boxes. Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan) and VOS with ERI (pink). Months are: 5/2002, 6/2002, 7/2002 and 8/2002.

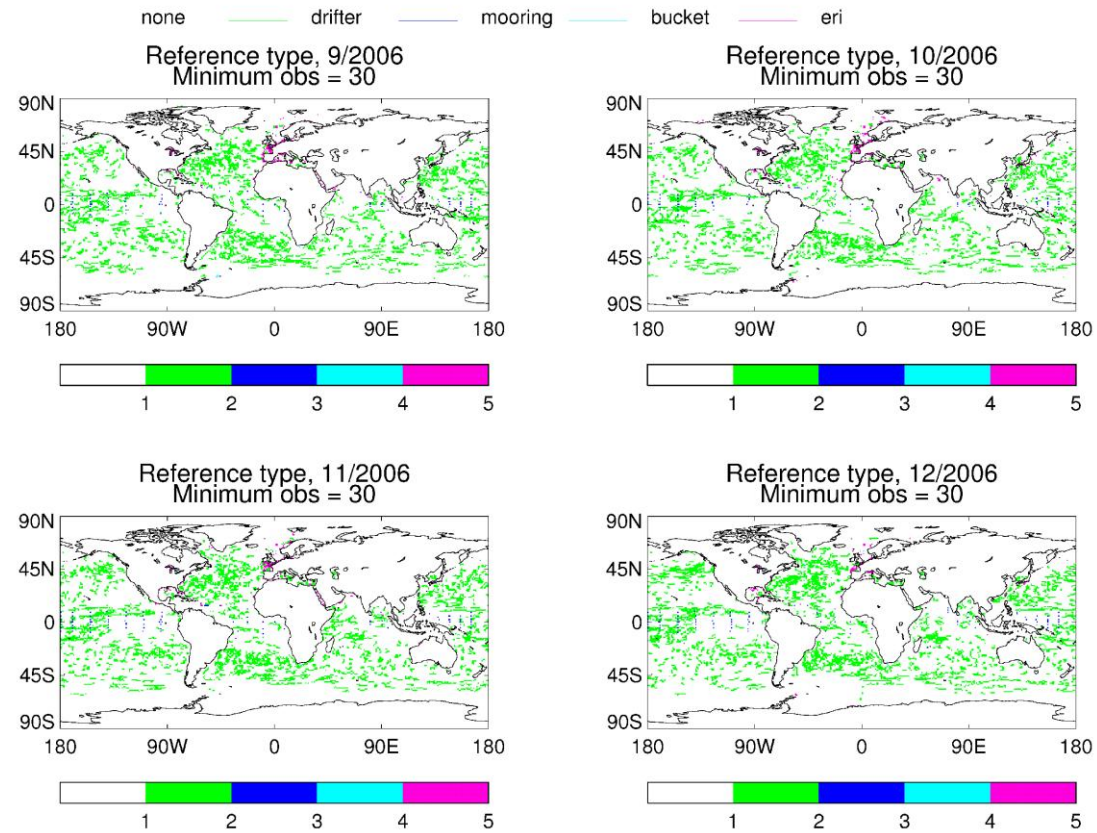


Figure 4-16: Reference types available with greater than 30 measurements in each month in 1° grid boxes. Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan) and VOS with ERI (pink). Months are: 9/2006, 10/2006, 11/2006 and 12/2006.

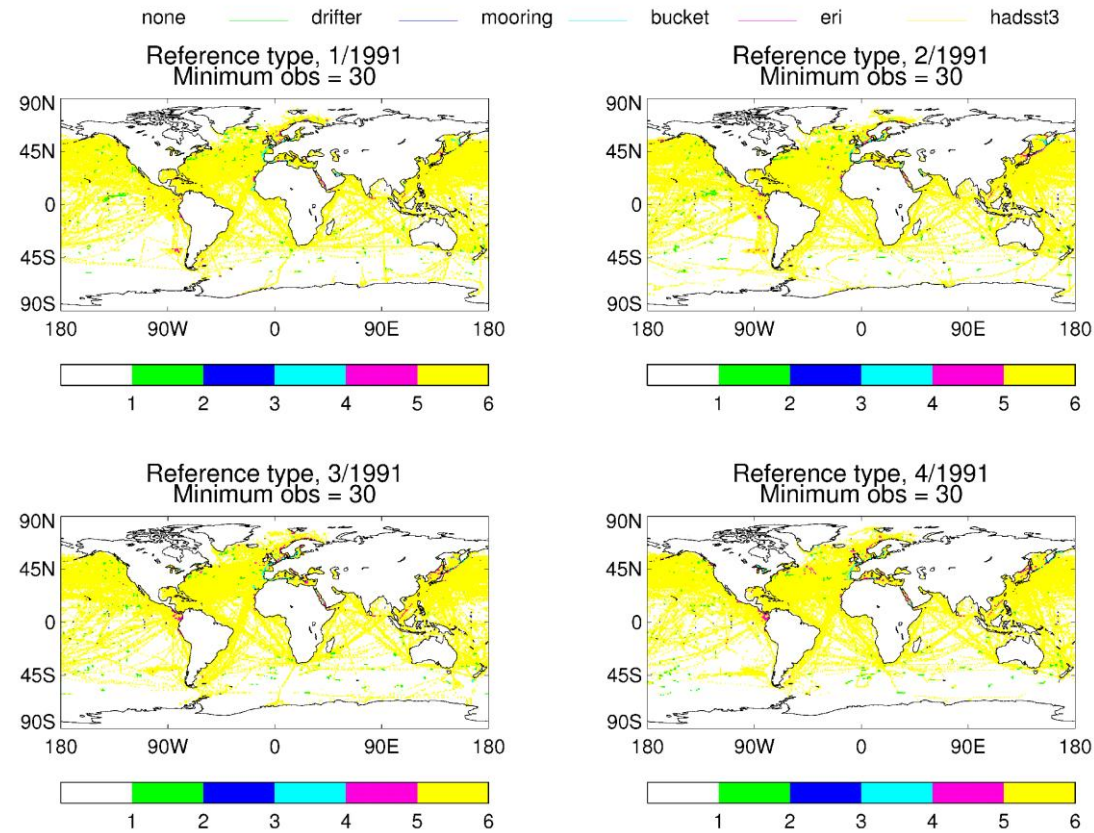


Figure 4-17: Reference types available with greater than 30 measurements in each month in 1° grid boxes (except HadSST3, where no minimum number is required). Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan); VOS with ERI (pink) and HadSST3 (yellow). Months are: 1/1991, 2/1991, 3/1991 and 4/1991.

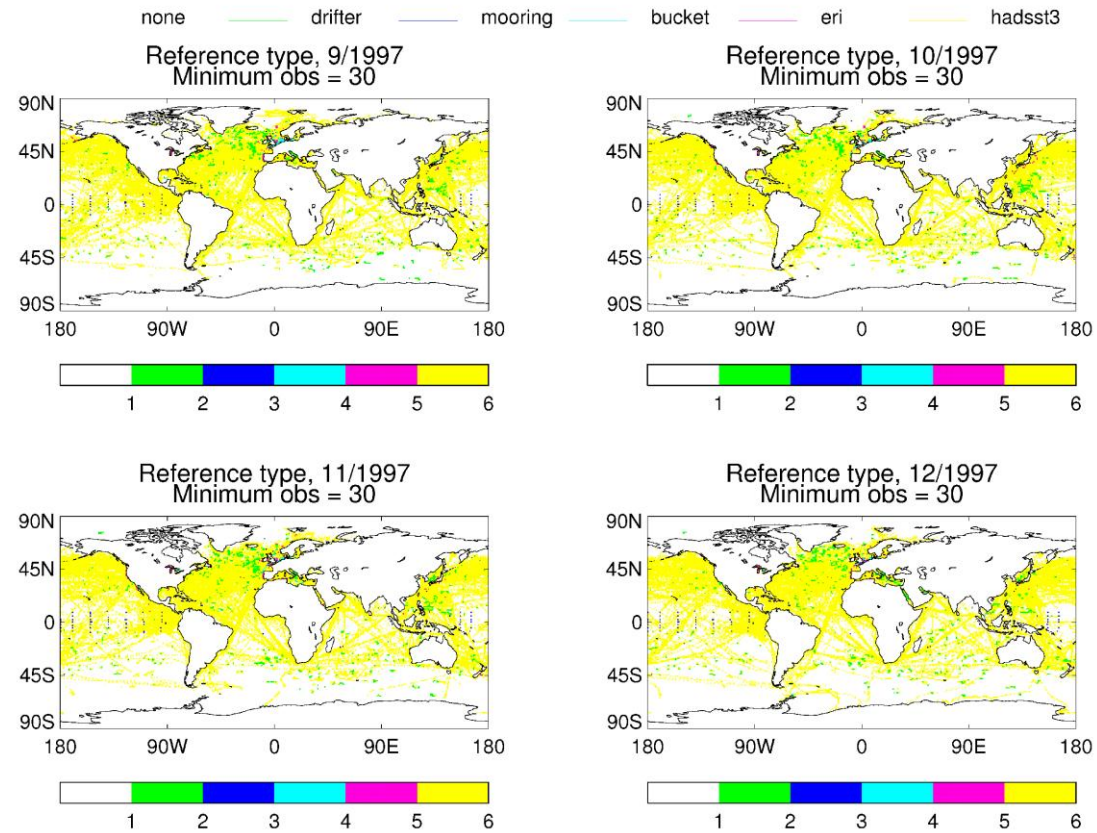


Figure 4-18: Reference types available with greater than 30 measurements in each month in 1° grid boxes (except HadSST3, where no minimum number is required). Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan); VOS with ERI (pink) and HadSST3 (yellow). Months are: 9/1997, 10/1997, 11/1997 and 12/1997.

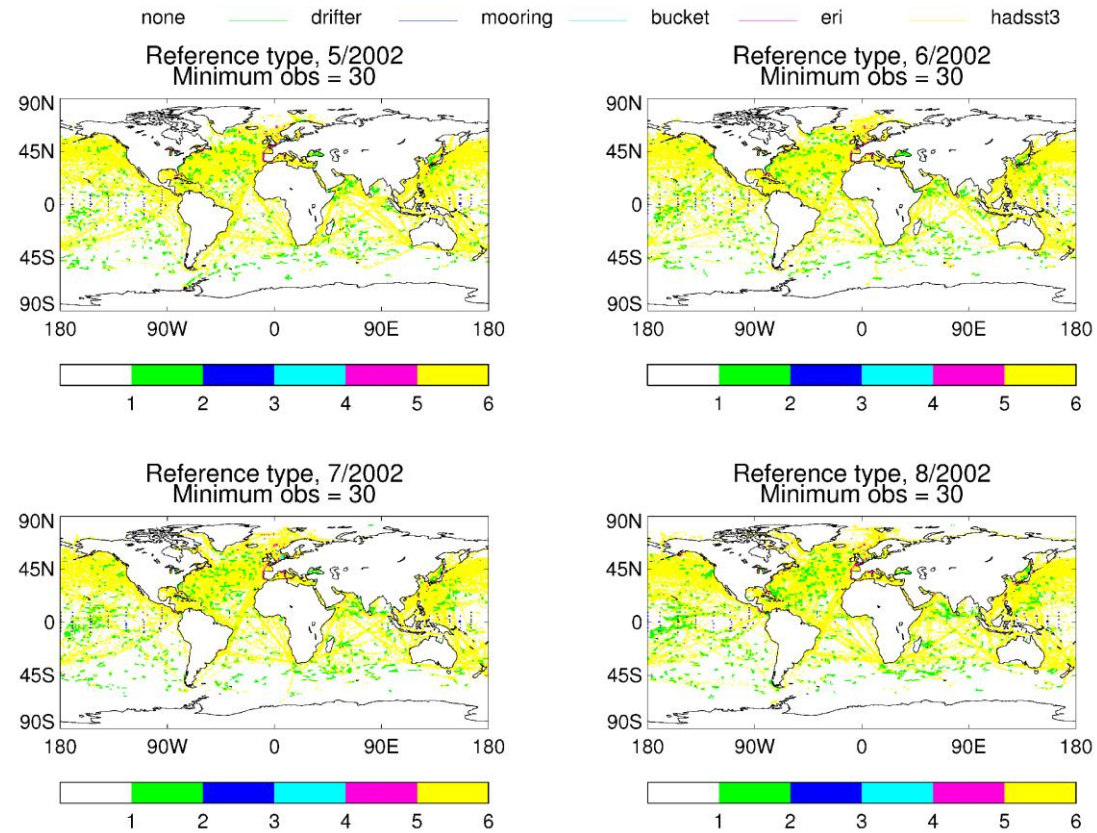


Figure 4-19: Reference types available with greater than 30 measurements in each month in 1° grid boxes (except HadSST3, where no minimum number is required). Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan); VOS with ERI (pink) and HadSST3 (yellow). Months are: 5/2002, 6/2002, 7/2002 and 8/2002.

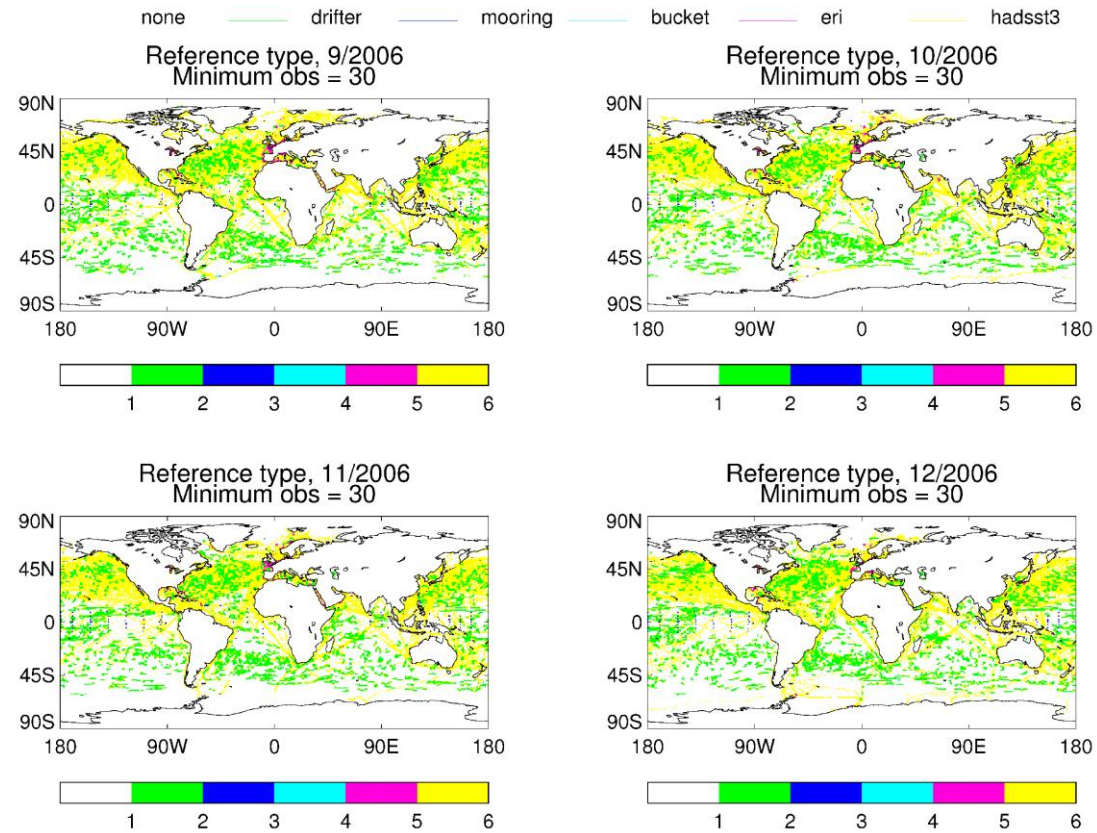


Figure 4-20: Reference types available with greater than 30 measurements in each month in 1° grid boxes (except HadSST3, where no minimum number is required). Reference types are: drifting buoy (green); tropical moored buoy (blue); VOS with bucket (cyan); VOS with ERI (pink) and HadSST3 (yellow). Months are: 9/2006, 10/2006, 11/2006 and 12/2006.

Clearly, including HadSST3 in our reference dataset increases our ability to perform validation on this spatial scale because, despite the lack of observations, it comes with quantified uncertainty estimates. Even including HadSST3, there are large regions without in situ coverage on a 1° grid.

Note that, since we propose to use a subset of drifting buoys and VOS that have been checked for gross biases and drifts, this will decrease the numbers of reference data available compared to the numbers depicted in Figure 4-13 to Figure 4-16.

4.4 Content of Reference Dataset

The content of the reference dataset at the time of writing this report is given in Table 4-1. Ongoing work in external projects to refine uncertainties and knowledge of in situ data will occur in parallel with the SST_CCI project and so the final content of the reference dataset will not be defined until the start of the product validation in February 2014.

Data type	Time period	Coverage	Comment
Ship-borne IR radiometers	2000 - 2010	Caribbean Sea; Bay of Biscay	Independent SSTskin
Argo floats	2000 - 2010	Global [#]	Independent SSTdepth
GT MBA	1991 - 2010	Tropics	Independent SSTdepth
Drifting buoys	2008-2010* 1991-2010	Global [#]	Independent and pseudo-independent SSTdepth
HadSST3	1991 - 2010	Global [#]	Pseudo-independent SSTdepth Gridded

Table 4-1: Content of SST_CCI reference dataset

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

[#] Data are not truly “global” but cover majority of Earth’s oceans.

5. MULTI-SENSOR MATCH-UP DATASET

Traditional SST retrieval/estimation algorithm development and validation has relied on a single sensor match-up dataset (MD), where the retrieved SST from a single satellite instrument is matched to a single in situ measurement. To support the development and validation activities in the SST_CCI project, we require a multi-sensor match-up dataset (MMD) 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 (such as a drifting buoy).

The MMD approach offers the new capability to cross reference sensors (in this case, to use the ATSR series as a reference for AVHRR) at the level of the retrieval process (rather than as a post-retrieval SST bias correction). Moreover, the MMD is designed to provide improved information for interpretation of the diurnal cycle and the consequent effects of time differences between various satellite measurements, as well as a partition of the uncertainty budget into components for each individual sensor (including in situ). Each multi-sensor match-up (MM) is supplemented with auxiliary data providing estimates of the atmospheric and surface state at the time of the satellite observation.

This section provides an overview of the MMD concept and how they will be produced within this initial demonstration phase. The system for producing MMD files, the Multi-Sensor Match-Up System, MMS, is described in the System Requirements Document (SRD; RD.258) and the System Specification Document (SSD; RD.259).

5.1 From concept to reality

The MMD approach described in this document uses as its basis pre-existing MD datasets for ATSR, AVHRR GAC, METOP Full Resolution Area Coverage (FRAC) and SEVIRI. The reason for using these files is that long-term archives of METOP FRAC and SEVIRI data were not available at the start of the project. A future operational system would not use the exact approach described in this document; instead it would ideally have the capability to create multi-sensor match-ups from scratch. This sub-section summarises the main differences between the two methods. We start by considering the ideal case.

As noted earlier, the traditional way of creating single sensor MDs for algorithm development and validation compares a single satellite measurement to a single in situ measurement. In this approach the in situ data are assumed to be the reference dataset and the nearest-in-space-and-time (using predefined criteria) match-up pair between the in situ measurement and the satellite dataset is selected in each case. The current predefined criterion recommended by the GHRSSST ST-VAL group for satellite SST validation is that the in situ measurement should be located within the satellite pixel within 2 hours of the satellite overpass (<https://www.ghrsst.org/ghrsst-science/science-team-groups/stval-wg/sses-common-principles/>) based on previous work of Minnett (1991; RD.234). Validation match-up criteria is an active area of research and the limits will be reviewed prior to the product validation work which does not start until February 2014.

In developing our MMD a slightly different approach is required in that a satellite dataset is chosen as the primary reference for a particular multi-sensor match, to which all other datasets (including in situ) are matched. For the non-primary matches, the match-up rule on temporal coincidence between satellite datasets is relaxed to within 12 hours, to allow for the multiple overpass times of the various satellites. The strict criterion on spatial overlap is retained, but in this case the centre of the second satellite pixel must reside within the boundary of the first satellite pixel. This process is somewhat simplified if a priority is assigned to each satellite dataset in order to define the primary sensor.

To define the priority for each sensor we use the best available knowledge of the estimated performance of each sensor relative to each other sensor. We also need to consider the full list of sensors being evaluated within the project. The SST_CCI project will create two categories of data products:

1. A long-term record that combines the ATSR and AVHRR series
2. A short-term demonstration product that combines ATSR, METOP, SEVIRI, AMSR-E and TMI.

Within both categories the ATSR series is to be used to bias correct the other infrared sensors. Consequently in any multi-sensor match-up the ATSR will always be the primary sensor. Subsequently one would add METOP FRAC, AVHRR GAC, followed by SEVIRI and then the PMW sensors.

To support the many activities proposed within the SST_CCI the MMD will not contain information on a single satellite pixel but will contain image extracts from each sensor covering roughly the same spatial area in total. In addition, each MMD record will have an in situ history covering the match-up window where available (e.g. a drifting buoy record covering 12 hours each side of the primary satellite sensor overpass time), and will also contain auxiliary information on the atmospheric and surface state from NWP models, aerosol forecasts and sea-ice analyses.

A more complete description of the MMD specification for SST_CCI Phase 1 can be found in the SST CCI MMD specification, RD.232.

5.2 Match-up rules

The SST_CCI Phase 1 MMD shall be built using the following spatial and temporal match-up rules:

- Spatial: Each MM record shall have a central reference location using the priority of ATSR > METOP > AVHRR GAC. The centre of each additional image extract from subsequent sensors added to the MM record shall overlap the central reference location.
- Temporal: All match-ups shall be within a maximum time window of +/- 12 hours.

5.3 Segregation of drifting buoy match-ups

All drifting buoy match-ups within the MMD are split into four categories for use within the project. This segregation is done once using a random number generator on one year's worth of match-ups at a time for each of the three reference sensors (ATSR; METOP; AVHRR GAC).

For data between and 1991 and 2007 the data is split:

- Training – 40%; test – 20%; selection – 40%

For data between and 2008 and 2010 the data is split:

- Training – 40%; test – 10%; selection – 40%; validation – 10%

The segregation ratios were chosen based on the previous experience of the SST_CCI project team to ensure sufficient match-ups are available throughout the time series for training and selection. The limited amount of data for validation is not seen as an issue as the SST_CCI project will provide uncertainties with each product and will not rely on validation for uncertainty estimates merely for confirmation that the uncertainties are realistic. However, some of the drifting buoy match-ups used for training, test and selection will be used as pseudo-independent data as part of the validation (see Section 7.8 for further details).

5.4 MMD Input data

The following input datasets are required to produce MMD files.

- DS1: ATSR MD files – single sensor match-ups in monthly NetCDF files
- DS2: METOP MD files – single sensor match-ups in daily NetCDF files
- DS3: SEVIRI MD files – single sensor match-ups in daily NetCDF files
- DS4: AVHRR MD files – single sensor match-ups in monthly NetCDF files
- DS5: ATSR Level 1b files – single orbit files in Envisat format
- DS6: AVHRR GAC Level 1b files – single orbit files in NOAA KLM format
- DS7: AMSR-E Level 2 files – single orbit files in GHRSSST L2P format (NetCDF)
- DS8: TMI Level 2 files – single orbit files in GHRSSST L2P format (NetCDF)
- DS9: Drifter files – history of drifting buoy measurements in NetCDF
- DS10: Sea Ice files – daily files of sea ice concentration (one for each hemisphere) in OSI-SAF HDF and NetCDF format
- DS11: TOMS-type aerosol – daily files of TOMS/GOME-1/OMI/GOME-2 aerosol absorbing index (AAI) in NetCDF
- DS12: ECMWF ERA-interim reanalysis forecast and analysis fields in GRIB and NetCDF

Further details on each dataset can be found in the Data Access Requirements Document (DARD; RD.172).

5.5 MMD output format

The SST_CCI MMD files will be output in NetCDF format using the specification given in RD.232.

6. SELECTION OF ALGORITHMS TO BE IMPLEMENTED IN SST_CCI

Once contributions to the Round Robin exercise have been received, an algorithm selection process will take place.

6.1 Purpose and Scope of Algorithm Selection in SST CCI

6.1.1 Purpose and Definition of Potential Scope

The role of algorithm selection in the SST CCI is to ensure that the most suitable algorithms are selected for creation of the long-term and short-term SST CDRs and are specified in the SST CCI processor. The algorithm selection process will be open in the sense that algorithms for the SST CCI processor are not predefined, but will be selected via the algorithm selection process defined in Section 6.3.

Algorithms considered via the algorithm selection process will include, on an equal basis, those entered into the process by parties other than the SST CCI project team.

The scope for algorithm selection in principle could cover the following categories of algorithms:

- Observation classification (e.g. ice, rain, cloud, aerosol, land, RFI).
- SST estimation.
- SST uncertainty estimation.
- SST product confidence assignment.
- SST-skin to SST-subskin to SST-depth adjustment.
- SST time adjustment to 1030 or 2230 h.

Not all categories are relevant to every sensor, either for technical/scientific reasons, or because of limitations to the overall SST CCI project scope in the light of the available resources. Table 6-1 defines the sensor and category combinations for which a formal algorithm selection exercise within the SST CCI was considered. These can be divided into four groups:

- areas of funded effort in algorithm development and/or testing of more than one algorithm within the SST CCI leading to the need for selection between algorithms;
- areas of funded effort where one algorithm is developed and/or tested within the SST CCI, leading to the need for a selection exercise only if other algorithms are submitted into the process;
- areas where no development or testing is feasible within the SST CCI.

Sensor	Classification	Estimation	Uncertainty	Confidence	Depth	Time
ATSR	✓§	✓§	✓§	✓&	✓§	✓§
AVHRR	✓§	✓§	✓§	✓&	✓§	✓§
Metop		✓§	✓§			
SEVIRI		§	§			
TMI			&	&		
AMSRE			&	&		

Table 6-1: Categories of algorithm relevant to sensors within the SST CCI.

§ = competing algorithms are being developed and/or tested within the funded SST CCI project in this area.

& = a single algorithm is being developed and/or tested within the funded SST CCI project.

✓ = external parties can submit algorithm results for comparison and potential selection.

It was not expected or essential that multiple algorithms will “compete” for every category-sensor combination, although, subject to the restrictions in Table 6-1, the openness of the process allowed for that possibility. Note also that the categories of algorithms for depth and time adjustments are generic to all sensors and the restriction to ATSR and AVHRR simply arises because these adjustment estimates will be used only within the long-term SST processor¹, which uses only the ATSR and AVHRR GAC.

6.1.1.1 Observation classification

Satellite radiances observed from space reflect surface properties (surface type, temperature, emissivity and/or reflectance) and the state of the intervening atmosphere (absorption, scattering and emission of radiance in the field of view, by gases, aerosols and/or clouds).

Not all SST estimation methods are equally valid for every state of the surface and atmosphere. There is therefore a need to identify which pixels are observations of surface or atmospheric states that render the estimation method invalid or of lesser quality.

Classifications of pixel state include (with relevance depending on the sensor):

- surface type contributing to pixel (e.g., water, ice, land or mixed).
- atmospheric state (clear-sky with negligible aerosol, clear-sky with significant aerosol, cloud-affected, cloud-filled, precipitation-affected).
- radiance conditions (sun-glint, Radio Frequency Interference (RFI)).

¹ The project provides for development of a long-term processor and climate data record based on ATSR and AVHRR GAC. In addition, a system for including a wider range of sensors will be demonstrated for a 6 month period, which will include METOP 0.05 degree data, SEVIRI, TMI and AMSRE.

Classification can be undertaken in a manner that yields a deterministic result (with certain rates of misclassification) or a probability for each of a set of outcomes. Note that the most common preference expressed by climate users of SST in the SST CCI URD [RD.171] is for a probability that a given pixel is valid.

6.1.1.2 SST estimation

Also known as retrieval or inversion, SST estimation is the process of inferring a value for SST from radiances (usually expressed as brightness temperatures).

6.1.1.3 SST uncertainty estimation

SST uncertainty estimation is the reasoned attribution of uncertainty information to an estimate of SST.

The total uncertainty in a single SST estimate reflects:

- the propagation of radiometric noise in the observed radiances through the estimation algorithm.
- the effect of algorithmic limitations, such as prior error and non-linearity error.
- the propagation of uncertainty in any ancillary information exploited in the estimation algorithm.
- the effect of classification errors and/or of undetected sub-pixel variations in the state.
- the uncertainty in radiance calibration propagated through the estimation algorithm.
- the uncertainty in true spatial location of the field(s) of view relative to the nominal geolocation.

An SST uncertainty estimate should quantify and combine at least the dominant sources of uncertainty.

Where composites or averages of individual SST observations are made:

- sampling (representativity) errors then affect the uncertainty in the composite or average SST (when interpreted as the SST for a specified area and time interval).
- some errors tend to average out whereas others do not, and the net effect on the uncertainty to attribute to the composite or average needs to account for these behaviours.

6.1.1.4 SST product confidence assignment

SST outputs within SST CCI will be GHRSSST-compliant and will therefore associate product confidence levels with the SST data. Ideally, this information should have some

intuitive consistency across the products of different sensors (so that data of a certain confidence level from different sensors can be sensibly combined).

For example, confidence could reflect the probability that the observations belong to a class for which the estimated SST is valid, and/or the uncertainty in the SST given that classification.

A common confidence *scale* has been defined within GHRSSST, but consistency between sensors regarding the qualitative meaning of points on the scale is not required. There are therefore no GHRSSST criteria at present for assessing the validity of one set of confidence flags compared to another for a given sensor.

6.1.1.5 SST-skin to SST-subskin to SST-depth adjustment

SST-skin to SST-subskin adjustment involves the estimation of the magnitude of the ocean thermal skin effect under the prevailing conditions on the SST-skin observation.

SST-subskin to SST-depth adjustment involves the estimation of the stratification of the near-surface ocean between the subskin and the SST-depth target depth.

6.1.1.6 SST time adjustment to 1030 or 2230

SST time adjustment is the estimation of the temperature difference (for SST-skin, subskin or depth) between the time of observation and the standardized time for the long-term SST product. The standardized times in the SST CCI project are 1030 and 2230 h local mean solar time. These times are chosen both because they are close to the local time of observation for the ATSRs, and therefore are the local times near which, at most latitudes, the most stable satellite observations are available.

6.1.2 Algorithm types covered in Algorithm Selection

A consultation with potential external algorithm “competitors” was undertaken in April 2011 advertised via GHRSSST, the Science Leader’s blog and by direct e-mails. The content for the call for interest is recorded in the blog (<http://sst-cci.blogspot.com/2011/04/preparing-for-round-robin.html>). There was external interest only in participating in a formal Algorithm Selection for “Estimation” and (where this is already defined within the SST retrieval process) “Uncertainty” algorithm categories.

All algorithms in the categories covered within the Algorithm Selection process will be documented in the SST CCI “Algorithm Theoretical Basis Document v0” [RD.225].

6.2 Organisation and responsibilities

6.2.1 Pre-Selection Engagement

Responsible: Science Leader

Prior to the announcement of the availability of the Round Robin Data Package (start of July 2011) the Science Leader has ensured that the international satellite SST community is alerted to the nature of the Algorithm Selection process via the NASA Science Team meeting (November 2010), the SST CCI and ERNESST web pages, the GHRSSST joint Working Group meeting (February 2011), and the GHRSSST XII Estimation and Retrievals Working Group workshop.

6.2.2 Putting the SST CCI development algorithm outputs in the MMD / RRDP

Responsible: EO Science Team WP Leaders

There are WPs within the SST CCI project involved in algorithm development. In order to compare and select algorithms, the results of these will be added to the MMD / RRDP prior to the beginning of the algorithm selection work package.

The relevant WPs are:

20410 - IR SST algorithm improvement (Pumphrey).

20414 - IR SST algorithm improvement - SEVIRI (Le Borgne).

20415 - IR SST algorithm improvement – Hi Lat (Hoyer).

20510 - IR cloud detection improvement (Merchant).

20516 - IR cloud detection improvement – Hi Lat (Eastwood).

20660 - Ice detection improvement (Eastwood).

20661 - Ice detection improvement – Bayesian classifier (Merchant).

20710 - Diurnal variability estimation (Merchant).

20713 - Diurnal variability estimation – Fairall (Rayner).

20810 - Passive microwave uncertainty characterisation (Pumphrey).

6.2.3 Solicit and receive extensions to MMD

Responsible: Validation Science Team Leader

Define a mechanism by which the MMD can be extended to include a broader range of sensors. Announce opportunity and solicit data inputs. Receive and verify inputs. Extend MMD using inputs.

This activity does not feed the Algorithm Selection process directly, but is undertaken to foster new links and to build future scientific exploitation.

6.2.4 Announcement and dissemination of RRD

Responsible: Validation Science Team Leader

Announce and disseminate RRD (start Sep 2011), informing external collaborators of protocols and timescales. A major mechanism was use of the GHRSS XII conference (28 June to 2 July, Edinburgh, UK). Seek and monitor engagement during time window for contributions (September 2011 to January 2012).

6.2.5 Round-robin data package consultation (WP 21220)

Responsible: Validation Science Team Leader

Co-ordinate inputs from external collaborators via RRD protocol. Write summary for external RRD submissions as contribution to Product Validation and Algorithm Selection Report. (August to October 2011).

6.2.6 Algorithm comparison and selection

Responsible: Science Leader and EO Science Team (WP leaders).

Table 6-2 summarises the responsibilities of the participating organisations for algorithm selection, presented by sensor and algorithm classification. The following work packages apply to this stage of development:

21310 - Algorithm comparison and selection (Merchant).

21314 - Algorithm comparison and selection – Metop & SEVIRI (Le Borgne).

21316 - Algorithm comparison and selection – Hi Lat (Eastwood).

These WPs are scheduled for April 2012 to May 2012.

Sensor	Classification	Estimation	Uncertainty	Confidence	Depth	Time
ATSR	UoE & met.no	UoE & met.no	UoE	UoE	UoE	UoE
AVHRR	UoE & met.no	UoE & met.no	UoE	UoE	UoE	UoE
Metop	<i>MF</i>	UoE & MF	UoE & MF	<i>MF</i>	NA	NA
SEVIRI	<i>MF</i>	MF	MF	<i>MF</i>	NA	NA
TMI	<i>RSS</i>	<i>RSS</i>	UoE	UoE	NA	NA
AMSRE	<i>RSS</i>	<i>RSS</i>	UoE	UoE	NA	NA

Table 6-2: Responsibilities for algorithm development and selection, by sensor and category of algorithm. Black text indicates responsibility both for development/testing internal to the project and for algorithm comparison and selection including externally submitted algorithm results (i.e., the scope of the “Round Robin”). Grey upright text indicates responsibility for development/testing internal to the project only. Italicised grey text means there is no development, testing or selection within the SST CCI project.

6.2.7 Write report and journal paper on algorithm selection

Responsible: EO Science Team

Write deliverable (Algorithm Selection Report [RD.226]) and draft a scientific paper covering the most important and/or innovative results. The Algorithm Selection Report is scheduled to be delivered at the end of May 2012.

6.3 Selection criteria and process

The selection of algorithms to be implemented in the SST CCI processor will be made by assessing algorithms against several criteria that depend on the algorithm category in question. To make comparisons fair, algorithm results need to be compared on common sets of matches within the RRDP, requiring complete submission of results to have been made in accordance with the protocol given in Appendix C.

6.3.1 Over-arching principles

Algorithms will be compared on a fair basis by standardisation of the approach:

- Competing algorithms:
 - will be developed using identified training data within the RRDP (if necessary);
 - can be objectively assessed by developers internally using identified test data; and
 - will be compared on the basis of results when applied to identified “blind” data.
- All algorithm developers including the project EO team will have access to the same data in the training and test categories (including in situ validation data), and to the same blind data (in situ data withheld).
- It is recommended that developers do not use test data in algorithm development/training, in order to maintain their own objective assessment of performance without tuning.
- The blind data (with in situ withheld) will be distributed towards the end of the exercise, in time for developers to apply their algorithms and submit the results to the Validation Science Team leader.
- Common metrics describing the results (detailed below) will be used for each type of algorithm to facilitate comparison of performance.

Algorithm selection requires joint assessment of a range of metrics and wider considerations. Not all properties of interest are quantifiable as metrics. Among measures that are quantifiable in principle, it may not always be feasible to undertake proper quantification within the scope of the project, and thus a qualitative approach may still be necessary.

For algorithm selection purposes, the validation data to be used are the matches flagged as “blind data” in the final (complete) release of the RRDP. Certain test data will also be flagged as “high latitude” and “coastal” cases, and some metrics will be evaluated for performance separately using these subsets.

6.3.2 Definition of common metrics

6.3.2.1 Bias

Bias is systematic difference from the truth, and is assessed via systematic differences from validation data. Since validation data may also have both bias and random error, the assessment can only be interpreted to the level of systematic uncertainty in the validation data and accounting for the statistical power of the available validation.

A difference between a retrieved quantity and the matched validation value is referred to as a **discrepancy**. This neutral choice of word emphasises that the difference arises from error in both quantities.

Bias is, therefore, assessed by looking at “systematic discrepancy”, bearing in mind the likelihood of bias being present in both satellite and validation data.

Bias is estimable where there are objective/independent data available for “validation”. It is an applicable concept to quantities that can sensibly be averaged.

The specific bias metrics that will be calculated across test data for different categories of algorithm are defined in Table 6-3.

Category of algorithm (categories as introduced in Table 6-2s)	Bias metrics
SST estimation	<p>For each sensor's test dataset:</p> <p>Calculate the mean and median SST discrepancy with respect to each class type of validation data. Here and below, the discrepancy is the difference between the validation data and the <i>individual satellite observation</i> ("single pixel") identified as the reference pixel in the RRDP.</p> <p>Map the mean and median SST discrepancy against drifting buoys. It is necessary that the map resolution is chosen such that the statistical uncertainty in the bias in each cell is adequately small, and this will depend on information, such as the number of matches per cell, that is not presently available. Therefore, the resolution will be chosen such that a mean 0.1 K discrepancy will be statistically significant for at least 90% of grid cells. The same resolution will be used for all algorithms applied to a given sensor. Statistical significance will be assessed assuming drifting buoys with different buoy IDs have biases drawn from a Gaussian distribution of standard deviation 0.2 K and no random uncertainty².</p> <p>Where sensors use different channel combinations and/or different algorithms in different situations, the above should be repeated for each situation.</p> <p>"Less biased" algorithms will give mean and/or median values closer to zero³ globally, and will have a narrower distribution of grid-cell mean and/or median values.</p>
SST uncertainty estimation	<p>For each sensor's test dataset:</p> <p>Calculate the chi-squared statistic, which measures the goodness of fit between the actual and estimated uncertainties of SST estimates and validation values, defined by:</p> $\chi^2 = \frac{1}{n} \sum_{i=1}^n \frac{(x_i - v_i)^2}{\delta x_i^2 + \delta v_i^2}$ <p>where n is the number of discrepancies, i is an index, x is the SST estimate, v means validation value, δx is the SST uncertainty estimate, and δv is the uncertainty attributed to the validation value.</p> <p>The expected value for χ² is unity. A value lower than this indicates the uncertainties attributed to the measurements or the validation values or both are too high. A value greater than unity indicates the uncertainties attributed to the measurements or the validation values or both are too low. Since the uncertainty to attribute to the validation values is itself uncertainty, the result for χ² should not be over-interpreted, and confidence intervals in the result will be estimated in order to assess the significance of differences in χ².</p> <p>Where sensors use different channel combinations and/or</p>

² The values assumed here are provisional, and may be updated if better understanding is obtained of drifting buoy errors in the interim.

³ In some circumstances, the "correct" mean or median discrepancy is a value other than zero, such as when comparing SST-skin and buoy SST.

different algorithms in different situations, the above should be repeated for each situation. Where the uncertainty estimate is known to be a function of some independent parameter (latitude, NWP TCWV, etc.), the statistic should also be calculated for bins of this parameter to assess the validity of that dependence.

Table 6-3: Bias metrics for categories of algorithm.

6.3.2.2 Non-systematic uncertainty (precision)

Observations generally differ from the truth according to a distribution that has a spread (or “dispersion”). The concept of precision is to characterise that dispersion, with a precise observation having a narrow spread. Precision is often said to describe the “random error” but this is an over-simplification. First, it is a matter of choice as to how certain errors are partitioned between “bias” or “precision”. An example is error that has a systematic geographical variation. This is partitioned into bias only if a geographically resolved approach is taken; in a global approach, this error variance appears as dispersion. Second, non-systematic error components are not necessarily truly random, although some are. For example, algorithmic errors in SST estimation are often well correlated within an image (e.g., on synoptic space-time scales) while they may be uncorrelated between images sufficiently separated in time or between parts of an image sufficiently separated in space.

For these reasons, the metric used within this SST CCI algorithm selection related to the dispersion of errors will be referred to as “non-systematic uncertainty”.

Non-systematic uncertainty is estimable where there are objective/independent data available for “validation”. It is an applicable concept to quantities whose standard deviation can sensibly be computed.

The specific metrics of non-systematic uncertainty that will be calculated across test data for different categories of algorithm are defined in Table 6-4.

Category of algorithm	Metrics of non-systematic uncertainty
SST estimation	<p>For each sensor’s test dataset:</p> <p>At the same grid cell resolution used above to map bias, map the standard deviation of discrepancy in each cell. Also map the RSD of discrepancy in each cell.</p> <p>Calculate the mean of the cell standard deviations for the statistically significant cells, and likewise the median of the cell RSDs. “More precise” algorithms will have maps with smaller values of the above metrics than “less precise” algorithms.</p> <p>Where sensors use different channel combinations and/or different algorithms in different situations, the above should be repeated for each situation.</p>
SST uncertainty estimation	<p>NA</p> <p>(While the uncertainty of an uncertainty estimate can be computed in principle, this will not be practically applied within the SST CCI project.)</p>

Table 6-4: Metric of non-systematic uncertainty for categories of algorithm.

6.3.2.3 Stability

Stability is constancy of bias in time. Stability of observation is critical when looking at differences between observations (i.e., changes of SST over time).

Category of algorithm	Metrics of non-systematic uncertainty
SST estimation	<p>In each sensor's blind dataset, the following need to be done for each channel-set/sensor combination relevant.</p> <p>Stability with respect to long-term trends: Put the N discrepancies for the stability subset in time order. From the standard deviation of discrepancy, calculate the number of points necessary for an average to have standard error ~ 0.01 K: $n \sim (SD / [0.01 K])^2$. Divide the series into N/n periods, and calculate the average discrepancy and average time for each. Fit a linear (or low order polynomial, if appropriate) to these points, reporting the slope(s) of the fit as a stability estimate.</p> <p>Stability with respect to seasonality: Divide the N discrepancies into latitude bands (south of 15 S, 15 S to 15 N, north of 15 N). For each band, bin all discrepancies by month. Calculate mean and standard error for each latitude-month bin, and inspect the means for evidence of any annual cycle that is significant compared to the standard error.</p> <p>Stability with respect to diurnal cycle: calculate the mean discrepancy and standard error for day and night subsets, and check for significant differences. (Since some algorithms apply only at night, check for significant differences between any proposed pairs of algorithms for day and night use, also.) Repeat the trend procedure for day and night separately, to check for the long-term stability of any diurnal bias.</p> <p>Day is the period between sunrise and sunset – i.e., when the solar zenith angle is less than or equal to 90 degrees. Night is the period between sunset and sunrise when the solar zenith angle exceeds 90 degrees.</p>
SST uncertainty estimation	<p>NA</p> <p>(While the stability of an uncertainty estimate can be computed in principle, this will not be practically applied within the project.)</p>

Table 6-5: Metric of stability for categories of algorithm.

6.3.2.4 Independence from in situ SST

SST retrievals can be based on either empirical correlations to in situ observations, or on radiative transfer modelling. For applications where satellite SSTs are required to complement, enhance or test in situ observations, independence from in situ SSTs is an advantage (and in some cases a necessity). While it is not guaranteed to be achievable, independence needs to be considered in algorithm selection.

The degree of independence from in situ SST will be evaluated and categorized as high, medium or low.

Category of algorithm	Metric of independence
SST estimation	Describe any usage of in situ observations in <i>defining</i> (not validating) algorithm.
SST uncertainty estimation	Describe any usage of in situ observations in <i>defining</i> (not validating) algorithm.

Table 6-6: Metric of independence from in situ SST

6.3.2.5 SST sensitivity

For the ideal SST estimate, changes in true SST are (on average) wholly reflected in changes in the estimated SST. Thus, $\frac{\hat{\alpha}}{\alpha}$ needs to be evaluated, where x is true SST and \hat{x} is the estimated SST. We refer to $\frac{\hat{\alpha}}{\alpha}$ as “SST sensitivity”⁴ for the SST estimate.

Some algorithms naturally provide $\frac{\hat{\alpha}}{\alpha}$ as an output.

In the RRDP, the rate of change in brightness temperature in the i^{th} channel with respect to change in SST ($\frac{\hat{\alpha}_i}{\alpha}$) will be provided based on RTTOV forward modelling and ECMWF NWP for each match. (It is assumed all algorithms operate with brightness temperatures.) Where an algorithm is differentiable (at least locally) in terms of these derivatives, the provided partial derivatives should be used to calculate the SST sensitivity. Where an algorithm is not differentiable, the partial derivatives provided should be used to calculate a perturbation of input brightness temperatures corresponding to a small increment in true SST, allowing an estimate by perturbation of the SST sensitivity.

Category of algorithm	Metric of SST sensitivity
SST estimation	Calculate SST sensitivity for each match. Map the mean SST sensitivity for drifting buoy matches on the same grid cells as used for mapping SST bias. Where sensors use different channel combinations and/or different algorithms in different situations, the above should be repeated for each situation.
SST uncertainty estimation	N/A

Table 6-7: Metric of SST sensitivity

⁴ Readers familiar with retrieval theory will recognize this as the element of the trace of the averaging kernel corresponding to the SST element of the state vector. We are not aware of a standard name in the literature for a single element extracted individually like this, and use of “averaging kernel” language is not intuitive in cases other than sounding (retrieval of profiles). Hence, we introduce the “SST sensitivity” term.

6.3.2.6 Generality

Part of the SST CCI project is to specify a system for creation of climate-quality SST products in the future. It is also the case that sensors can partially fail, in such a way that the number of channels decreases, but SST information is still retrievable.

In this context, a relevant factor for algorithm selection is the degree to which it is adaptable to other sensors and/or channel combinations, including future missions. An algorithm whose structure and functioning is independent of the number of channels or precise spectral characteristics are more general, and, other things being equal, this would be preferred. It may be advantageous to use an algorithm that, for some reason, is dedicated to a single sensor, but as part of a bigger system, this brings an overhead in terms of specification and maintenance.

The approach to be taken to assess generality is to create a list for each algorithm that states:

- to what sensors and/or channel-combinations and/or situations it applies.
- the degree and nature of adaptation required to apply the algorithm to a new sensor/channel-combination/situation.

On the basis of the comments, each algorithm's generality will be categorized as high, medium or low. This is illustrated (with invented content) for the case of an ATSR-series SST retrieval algorithm in Table 6-8:.

Applicability	Algorithm 1	Algorithm 2
Sensors	ATSR-1 and ATSR-2 and AATSR	ATSR-2 and AATSR
Views	Nadir only and Forward only and Dual	Nadir only and Dual
Thermal channel combinations	11 and 12 μm or 3.7, 11 and 12 μm or 3.7 and 11 μm	3.7, 11 and 12 μm
Reflectance channel combinations	All (reflectance channels not required)	0.87 and 1.6 μm or 0.55 and 1.6 μm
Scene illumination	Day and night	Day (solar zenith angle < 75°)
Swath	Full swath	AATSR: Full swath ATSR-2: Restricted

Stratospheric aerosol conditions	All (robust)	Background levels (non-robust)
SUMMARY	High, applying to all sensors in series, and all useful instrumental and environmental conditions – i.e., applicable to approximately 100% of input data	Low, applying only to specific channel combinations that represent approximately 30% of input data.

Table 6-8: Hypothetical example of table comparing the generality of application of different SST retrieval algorithms.

6.3.2.7 Improvability

It is not clear what the mode of operation of the future CCI processors will be, but one possible model is of “Continuous Development and Operations”. The principle is that while operations are ongoing, insight is gained into the outcomes of algorithms that permit their improvement, and there is a cyclical effort to upgrade the operations accordingly.

The degree to which an algorithm can be further refined is therefore of long-term relevance. For each algorithm, a statement will be requested/provided specifying potential methodological improvements and the degree to which they are likely to improve accuracy, uncertainty or stability.

It is recognised that this aspect can only be assessed subjectively (since the actual improvements and results are not available). On the basis of the information available, improvability will be categorized as high, medium or low.

High will mean that the algorithm provider has clearly identified three or more routes for further improvement, each with credible expectation of reduced uncertainty. Medium will mean that the algorithm developer has clearly identified one or two routes for further improvement, each with credible expectation of reduced uncertainty. Low will mean that no routes to algorithm improvement have been identified.

6.3.2.8 Difficulty of implementation

For each algorithm, the following will be listed:

- The size and nature of any static auxiliary files required for the algorithm (e.g., look up tables).
- The size and nature of any dynamic auxiliary files required for the algorithm (e.g., NWP).
- The main steps of the algorithm, identifying where significant computation, use of external models, etc, is involved.

On the basis of the information available, difficulty of implementation will be categorized as high, medium or low.

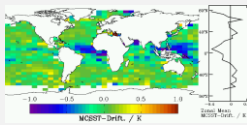
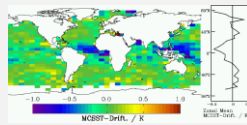
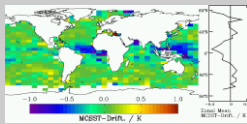
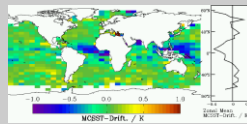
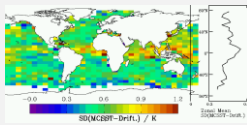
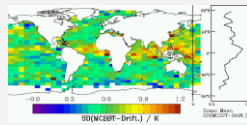
6.3.3 Selection of SST_CCI algorithms to be implemented

For each algorithm, all relevant metrics will be generated and documented. The relevance of different metrics to categories of algorithm is summarized in Table 6-9.

	Estimation	Uncertainty	Importance Weighting
Bias	✓	✓	Very High
Precision	✓		Medium
Stability	✓		Very High
Independence	✓	✓	High
Sensitivity	✓		High
Generality	✓	✓	Medium
Improvability	✓	✓	Medium
Difficulty	✓	✓	Low

Table 6-9: Relevance of different metrics to categories of algorithm for systematic consideration in the algorithm selection, all relevant metrics need to be provided and complied on a common basis.

To facilitate comparison between “competing” algorithms, the results will be compiled in a tabular form allowing ready comparison. This is illustrated (with invented content) in Table 6-10. Entries that indicate a likely conflict with user requirements will be highlighted in red.

	Algorithm 1	Algorithm 2	Weight
Bias (mean discrepancy)	0.04 K	-0.11 K	Very High
Bias (median discrepancy)	0.03 K	-0.08 K	Very High
Bias (mean discrepancy map)			Very High
Bias (median discrepancy map)			Very High
Precision map of (SD discrepancy)			Medium

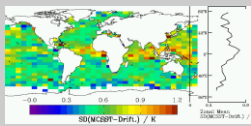
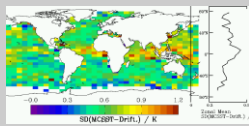
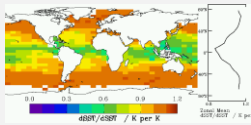
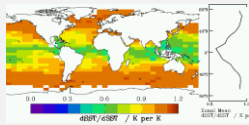
Precision map of discrepancy (RSD)			Medium
Precision (mean of cell SDs)	0.43 K	0.39 K	Medium
Precision (median of cell RSDs)	0.33 K	0.34 K	Medium
Stability with respect to trend	0.004 ± 0.003 K/yr	0.013 ± 0.005 K/yr	Very High
Stability with respect to season (amplitude of cycle)	North Not significant Equator Not significant South 0.2 K	North 0.3 K Equator Not significant South 0.5 K	Medium
Stability between day and night	0.13 ± 0.02 K day-night. No significant trend.	0.06 ± 0.03 K day-night. Trend in day-night of 0.02 K/yr.	Medium
Independence from in situ	High	Low	High
Map SST sensitivity			High
Generality	High	Low	Medium
Improvability	High	High	Medium
Difficulty	High	Low	Low

Table 6-10: Illustrative table for side-by-side comparison of two hypothetical SST estimation algorithms using standardized metrics.

The selection of algorithm requires a reasoned analysis of the relative merits of the algorithms. Part of the process is to assess, for a particular selection decision, which metrics should receive more consideration. This assessment will be documented in the summary table (it is the “Weight” column in Table 6-10) and justified in the discussion text. Where metrics give an indication of whether the algorithm is adequate to meet user requirements, this will be explicitly discussed using information from the URD. Ultimately, we can expect competing algorithms to have strengths in different areas, and the decision will consist of weighing the relative strengths and weaknesses. This selection process maps directly onto the trade-off analysis outlined in the statement of work [SST-TR-34]. Table 6-11 shows how each of the trade-off analysis criteria are met by the proposed metrics. The reasoned analysis leading to the algorithm selection will be documented carefully for each selection, in the algorithm selection document.

Trade-off Analysis Criteria	Algorithm Selection Metric
Global retrieval accuracy	Bias
Degree of residual cloud contamination.	None applicable (no classification in RR)
Degree of residual aerosol contamination	None applicable (no classification in RR)

Performance of products in the MIZ.	Bias and non-systematic uncertainty assessed on high-latitude subset of test data
Performance of products in the coastal zone.	Bias and non-systematic uncertainty assessed on coastal subset of test data
Performance of products with respect to diurnal variability.	Stability
Ability to meet user and GCOS requirements.	Selection Process
Potential for further algorithm improvement to achieve ECV accuracy.	Improvability

Table 6-11: Match-up between trade-off selection criteria [SST-TR-34] and algorithm selection metrics.

7. VALIDATION OF SST_CCI PRODUCTS

The outputs from the SST ECV demonstration period and long-term SST record will be verified and validated against independent and pseudo independent reference data (see Section 7.8 and Table 4-1). Uncertainties in the output products will be taken account of along with known uncertainties in the independent reference data.

A key requirement in the SoW [RD.164] is 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 will have no involvement in algorithm development and selection, achieving the independence required.

The output of product validation and user assessment will be the Product Validation and Inter-comparison Report (PVIR). In addition, all results and findings will be published in the scientific literature in peer review journal articles. Publication of peer reviewed journal articles is seen as the key step in ensuring scientific acceptance of the SST_CCI outputs.

7.1 Endorsement of methods

This section on product validation is written using the best available knowledge at the date of issue of this document. It is expected that improved knowledge of in situ uncertainties, validation methods and procedures will be available prior to the actual product validation work taking place later in the project. Consequently this section of the PVP will be reviewed and updated prior to the actual product validation with the latest knowledge at that particular moment in time.

The endorsement of methods and procedures used will be sought through a number of approaches. In the first instance, agreed principles for satellite SST validation approved by the GHRSSST Science Team will be adopted. These methods ensure conformance with the guidelines for calibration and validation provided under the QA4EO framework developed by the CEOS-WGCV. It is expected that the methods defined in this document will form the basis for a community best practice protocol on SST validation developed in conjunction with GHRSSST.

7.2 Definitions

We propose to adopt 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.

7.3 SST validation

In an ideal case, the derivation of satellite SST uncertainties would require:

1. A complete traceable characterisation to agreed national standards of the satellite instrument and SST retrieval algorithm, at all times throughout the lifetime of the mission, and
2. A suite of global traceable reference data points that preserve the nature of the satellite SST, are of an accuracy and precision that is higher than the satellite sensor, and are provided throughout the mission.

In addition, if the sensor is part of a series, there should be a sufficient overlap period between successive sensors to allow for a robust characterisation of the inter-sensor period, using the same traceable reference data points for both sensors.

The traditional approach to SST validation has been to compare the satellite products to in situ data (with the in situ assumed to be 'truth') and to use the resulting bias and standard deviation as an indicator of the accuracy and precision of the satellite dataset. Note: Here *bias* is used as the in situ data is taken as truth within GHRSSST; within SST_CCI the term *discrepancy* will be used as the in situ data is not taken as truth (see Section 2 for further details on the definition of bias and discrepancy).

This approach is currently used widely in the SST community and is the basis for deriving the Single Sensor Error Statistics (SSES) provided in GHRSSST L2P products. The bias and standard deviation calculated from the comparisons to the in situ dataset are derived from a database of match-up coincidences produced within predefined spatial and temporal limits; the current GHRSSST match-up limits for SSES are nearest pixel (i.e. the satellite pixel containing the in situ location at the time of the overpass) and within 2 hours of the satellite overpass time.

Of course the bias and standard deviation calculated from such a comparison do not provide the uncertainty of each dataset individually, but are simply the mean difference and combined uncertainty of a two dataset comparison. Consequently, the resulting statistics are often dominated by real changes in the SST that can occur within the predefined spatial and temporal limits in addition to uncertainties from both datasets. In the SST_CCI project the products will be provided with associated uncertainties, derived from our understanding of uncertainties inherent in the retrieval, and we will validate both the product and its uncertainties using measurements from the reference dataset.

Recently, a new method of multi-sensor match-up processing has been proposed that aims to deduce the uncertainty of an individual dataset, providing it is bias free (O'Carroll et al., 2008). This approach to uncertainty estimation using multi-sensor match-ups will be fully exploited within the SST_CCI project using the MMD.

7.4 Reference data

The product validation will use a variety of reference data including, amongst others, drifting buoys, Argo floats and ship-borne radiometers. Full details of the reference dataset for the SST_CCI project are given in Section 4.

7.5 Rules and responsibilities for objective independent product validation

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

- The overall validation will be led by UoL (Corlett), who will also lead the validation using in situ and other reference data
- The Met Office (Rayner and Martin) will lead the inter-comparison and other user assessment activities
- DMI (Hoyer) will focus on high latitude validation
- No other team members participate in product validation aside from the development of tools (Brockmann Consult)
- A set of in situ data will be reserved solely for validation and will not have been used (previously) for algorithm selection

7.6 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 shall adopt the current GHRSSST limits such that the reference data are within the satellite pixel within 2 hours of the satellite overpass. 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.

7.7 Validation confirmation levels

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 it is important that users use the uncertainty information provided in the product and do not rely on comparisons to other datasets. Therefore we propose to provide maps to indicate the *degree of confirmation* that the validation provides taking into account the uncertainty and availability of the reference data.

We will provide degree of confirmation maps at 15° resolution that:

- Are for each SST_CCI output product
- Are provided annually or monthly where sufficient validation data is available
- 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.

7.8 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% from 2008 onwards)
 - GTMBA
 - VOS and other ships
 - Argo
 - Ship-borne radiometers
2. Pseudo-independent data: Use all drifter match-ups. We do not expect to use SST algorithms that are tuned to drifting buoys, and in this case these matched data remain statistically independent, although not independent of the algorithm development and selection process.
 - Allows improved regional validation

Clearly the degree of confirmation 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 SST_CCI L4, which does not directly use the drifter data.

7.9 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 dramatically improves the match-up coverage (both temporally and spatially). Also, as this type of comparison uses 'average' in situ data there is 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.

7.10 Analysis procedures

All SST ECV system outputs will be validated using the three types of validation data detailed in Section 7.9 noting the degree of independence detailed in Section 7.8. Discrepancies and uncertainties will be derived using both robust and non-robust statistical methods for each type of reference data, and where sufficient match-ups allow, by a least squares fit of a Gaussian distribution function to the histogram of the discrepancies. Uncertainties will be provided for a confidence level of 68% (the “one-sigma” level). All validation will be 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 discrepancy and uncertainty will be provided for each SST_CCI dataset, as well as any dependence on auxiliary data in the MMD (e.g. wind speed), proximity to nearest cloudy pixel and satellite and solar zenith angles. The stability of each SST_CCI product relative to the reference dataset will be determined by looking at the relative change in discrepancy from month to month.

The results from the independent validation will be compared to the products uncertainties to identify areas where they are self-consistent. All results will contribute to the degree of confirmation maps detailed in Section 7.7.

7.11 Review process and decision sequence

The review process and decision sequence will be:

1. Ingest SST_CCI products in MMS
2. Extract multi-sensor match-up records from MMS
3. Determine discrepancy and uncertainty of each coincident satellite/reference match-up pair using criteria defined in Section 7.6
4. Plot discrepancy and uncertainty using methods defined in Section 7.10
5. Produce degree of confirmation maps for each product using methods defined in Section 7.7
6. Write PVIR and journal article
7. Release products

7.12 Re-validation of upgrades

The entire validation process will be repeated after each future upgrade of the SST_CCI products. This involves re-evaluating the composition of the reference dataset as well as the validation methods and procedures.

8. SST_CCI PRODUCT INTERCOMPARISON WITHIN THE GMPE

We will compare the long-term ECV on a 0.25° latitude by longitude daily grid in the context of the near-real-time (NRT) GHRSSST Multi Product Ensemble (GMPE, WP 40130, SST-TR-30, SST-TR-40). Currently, results from the NRT system are displayed on a web-page:

http://ghrsst-pp.metoffice.com/pages/latest_analysis/sst_monitor/daily/ens/index.html

Each analysis in the ensemble is interpolated onto a regular 0.25° grid. For each point on the grid, the ensemble median and standard deviations are calculated, and the total number of contributing values is accumulated.

A peer reviewed paper will be drafted, documenting the long-term comparisons. In addition, given the available information, we will document the strengths and weaknesses of the analyses and their design to enable potential users of the SST_CCI products to make an informed choice between analyses. Users need different types of products, depending on their application (see RD.171).

8.1 Long-term product

We will develop the NRT GMPE system to enable longer-period comparisons. There are currently very few high resolution analyses which span the 1991-2010 period, compared to the number available for the NRT system. We will perform comparisons to the analyses listed in **Error! Reference source not found.**1, and will compare these analyses to the reference data set (Section 4).

Product name	Version	Reference
MyOcean OSTIA reanalysis	Version 1.0	Roberts-Jones, J., E. Fiedler and M. Martin, 2011: Daily, global, high-resolution SST and sea-ice reanalysis for 1985-2007 using the OSTIA system, J. Climate, 25, 6215-6232.
CMC		Brasnett, B. (2012). A 20-year reanalysis of sea surface temperature, CMC. Brasnett, B. (2008). The impact of satellite retrievals in a global sea-surface-temperature analysis. Q. J. R. Meteorol. Soc., 134: 1745-1760. DOI: 10:1002/qj.319.
NOAA Optimum Interpolation 1/4 Degree Daily Sea Surface Temperature Analysis - AVHRR OI	Version 2	Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S., Schlax, M.G., 2007, Daily High-Resolution-Blended Analyses for Sea Surface Temperature. J. Climate, 20, 5473–5496. doi: 10.1175/2007JCLI1824.1; http://www.emc.ncep.noaa.gov/research/cmb/sst_analysis/
HadISST2	Version 2	Under development
JMA MGDSST_re		Kurihara, Y., T. Sakurai, and T. Kuragano (2006), Global daily sea surface temperature analysis using

Product name	Version	Reference
		data from satellite microwave radiometer, satellite infrared radiometer and in-situ observations. Weather Bulletin, 73. s1-s18 (in Japanese).

Table 8-1: Analyses used to compare to the long-term and demo ECV in the GMPE.

In particular, we will assess their standard deviation from the GMPE median and their feature resolution, e.g. by assessing gradients or undertaking spectral analysis. We will also assess the stability of these analyses, e.g. by comparison to tropical moorings.

8.2 Demonstration product

We will entrain the 1st demonstration product (JJA 2007) into a demonstration version of the long-term GMPE system, including it in calculations of a new GMPE median. We will also compare the demonstration ECV and other analyses within the GMPE system (see **Error! Reference source not found.**) to Argo, since they provide an independent reference. This will be done on monthly resolution, for different ocean regions.

9. THE SST_CCI CLIMATE ASSESSMENT REPORT

In the Climate Assessment Report (CAR), the Climate Data Research Package (CDRP) will be used to assess:

- the long term behaviour of the SST_CCI products on global and regional scales;
- the impact of the SST_CCI products on the output of climate models and/or the impact of the SST_CCI products on the assessment of the output of climate models; and
- the consistency of SST_CCI products with other related CCI ECV products.

In this section, we discuss plans for each of these activities. We detail how activities undertaken within the SST_CCI will be brought together with voluntary contributions from engaged users outside of the SST_CCI project to compile the CAR in August/September 2013. Where appropriate, results will also be used to draft papers for peer reviewed publication.

9.1 Assessment of long-term behaviour of SST_CCI products

Our aim is to independently assess the SST_CCI products' quality and fitness for purpose. To that end, we will characterise the products' stability and verify their estimated uncertainties through comparison to the reference data set and other SST data sets and analyses, at both high and low spatio-temporal resolution. We will:

1. incorporate the results from comparisons using individual matchups from the MMD (WP 40220, see Section 7) and
2. produce global, hemispheric and regional averages and standard indices, used to summarise modes of variability, from the SST_CCI products and the reference and other data sets and assess trends and variability in these (WP 40430, see Section 9.1.1).

The latter activity will also allow us to assess the long-term consistency of the SST_CCI products with these other established SST data sets and analyses.

We will perform some analysis ourselves, described in Sections 9.1.1, 9.2.1 and 9.3.1 and invite and encourage independent international experts to perform complementary analysis (outlined in Sections 9.2.2 and 9.3.2).

9.1.1 Our analysis

In this section, we detail our plans to assess the behaviour of the long-term ECV over multi-annual and decadal time scales. Here, we will compare the long-term ECV product to established, multi-decadal, lower resolution SST data sets and analyses, as used by the IPCC and others. We will explore consistency with these other data sets and analyses with a view to understanding any differences found (see later in this section). Our focus here is to assess the large-scale features of the SST_CCI products. Many of the longer term data sets and analyses are available only on monthly and 1 or 5° latitude resolution. Therefore, when comparing them to the long-term ECV products, we will have to degrade the resolution of the SST_CCI products to be the same as that of the comparison data.

Hence, we will not be able to calculate these diagnostics on the full resolution of the long-term ECV products.

However, as previously described in Section 8, we will exploit the GHRSSST Multi-Product Ensemble (GMPE) system to perform comparisons to higher resolution analyses for the last two decades at higher resolution. Our focus here is to assess the information provided by the high resolution of the SST_CCI products.

We will also incorporate the assessment of the stability of the SST_CCI products (SST-TR-46, Section 7) by comparison to the reference data set (described in Section 4).

Many data sets and analyses of SST designed for use in climate research are produced at much lower spatial and temporal resolution than the SST_CCI long-term ECV. For example, there are very few daily SST analyses which have been produced for more than a few years. More commonly, SST analyses used in climate research represent monthly, or perhaps weekly, averages. Frequently too, SST data sets and analyses are presented on 1 or 5° latitude by longitude grids, or at best on a 0.25° grid. This means that the SST_CCI long-term ECV will have to be averaged from its 0.05° grid to a lower-resolution in order to make meaningful comparisons with many other climate SST products.

The long-term behaviour of a data set or analysis can be described through:

- the analysis of linear trends both globally and regionally (SST-TR-47) for regions summarised in
- the exploration of known inter- or multi-annual modes of variability, through calculation of standard indices (shown schematically in Figure 9-1, lower):
 - Niño 1+2 [0-10°South, 90°West-80°West]
 - Niño 3 [5°North-5°South, 150°West-90°West]
 - Niño 4 [5°North-5°South, 160°East-150°West]
 - Niño 3.4 [5°North-5°South, 170-120°West]
 - Dipole Mode Index (DMI) [SSTA differences in the western tropical Indian Ocean and south-eastern tropical Indian Ocean (Saji et al., 1999; RD.241)]
 - tropical Atlantic meridional SST gradient (TAMG) (a mode of climatic variability known to be largely associated with abnormal rainfall regimes in South America and West Africa) [difference between area indices of sea surface temperature anomaly north and south of the meteorological equator (~5°N)]
- comparison of multi-annual or decadal averages
 - 1991-1995; 1996-2000; 2001-2005; 2006-2010
 - 1991-2000; 2001-2010; and
- the calculation of autocorrelations in time.

We will produce these diagnostics for the long-term ECV, the gridded reference data set and the other SST data sets and analyses (see **Table 9-2**, WP 40430). This will allow a detailed comparison to be made and any relative drifts or biases will be highlighted.

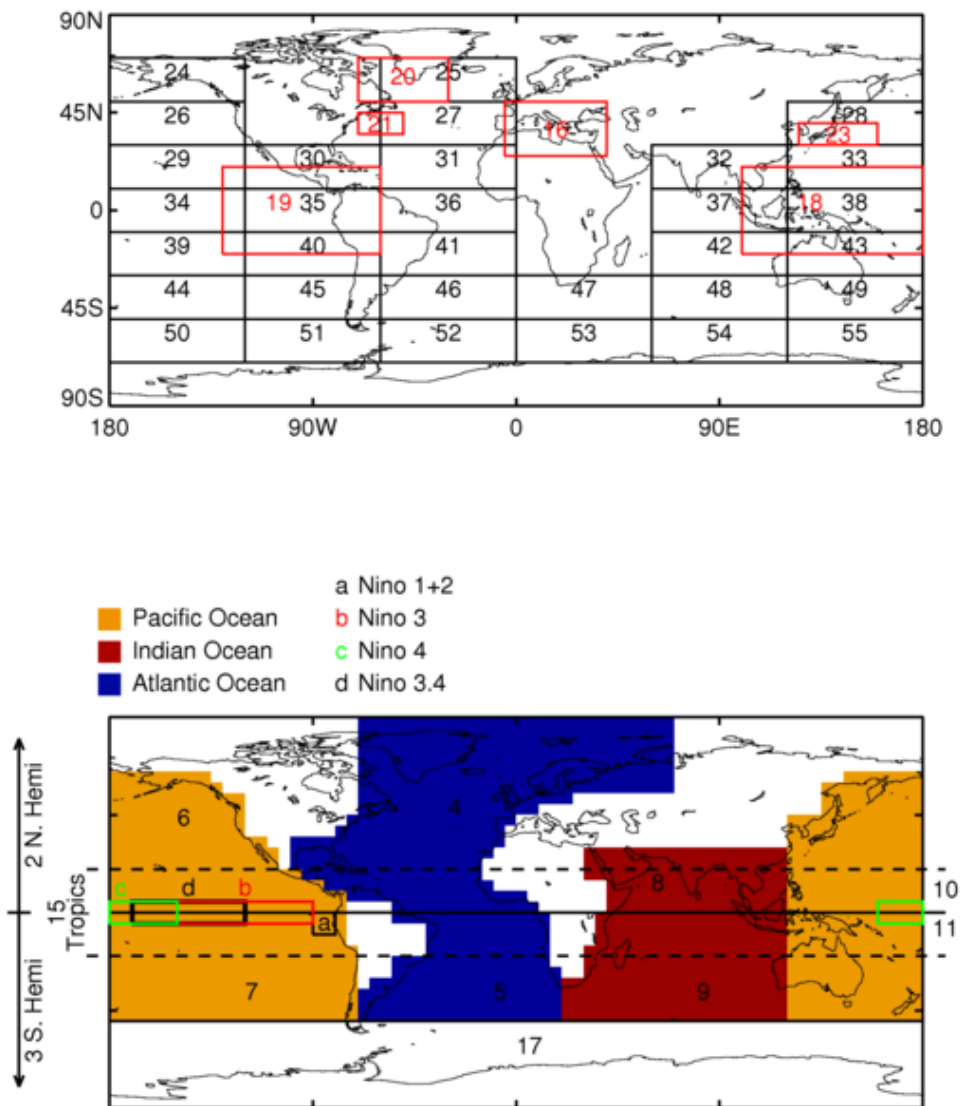


Figure 9-1: (Upper) Map indicating the regions for the analysis of linear trends summarised in Table 9-1 and (lower) regions for the exploration of known inter- or multi-annual modes of variability, through calculation of standard indices as listed in the text.

Globe	1
northern hemisphere	2
southern hemisphere	3
north atlantic ocean	4
south atlantic to 50s	5
north pacific ocean	6

south pacific to 50s	7
north indian ocean	8
s.indian ocean to 50s	9
northern tropics	10
southern tropics	11
atlantic ocean to 50s	12
pacific ocean to 50s	13
indian ocean to 50s	14
tropics (20n - 20s)	15
mediterranean	16
southern ocean, 50s southwards	17
w. tropical pacific	18
e. tropical pacific	19
greenland 50-70n 30-70w	20
gulfstream 35-45n 50-70w	21
s.hem & n.ind-rest n.h.	22
kuroshio 30-40n 125-160e	23
area 50-70n 180-120w	24
area 50-70n 60-0w	25
area 30-50n 180-120w	26
area 30-50n 60-0w	27
area 30-50n 120-180e	28
area 10-30n 180-120w	29
area 10-30n 120-60w	30
area 10-30n 60-0w	31

area 10-30n 60-120e	32
area 10-30n 120-180e	33
area 10n-10s 180-120w	34
area 10n-10s 120-60w	35
area 10n-10s 60-0w	36
area 10n-10s 60-120e	37
area 10n-10s 120-180e	38
area 10-30s 180-120w	39
area 10-30s 120-60w	40
area 10-30s 60-0w	41
area 10-30s 60-120e	42
area 10-30s 120-180e	43
area 30-50s 180-120w	44
area 30-50s 120-60w	45
area 30-50s 60-0w	46
area 30-50s 0-60e	47
area 30-50s 60-120e	48
area 30-50s 120-180e	49
area 50-70s 180-120w	50
area 50-70s 120-60w	51
area 50-70s 60-0w	52
area 50-70s 0-60e	53
area 50-70s 60-120e	54
area 50-70s 120-180e	55

Table 9-1: List of regions for the analysis of linear trends in the SST_CCI time series. The regions are shown schematically in Figure 9-1 (upper).

Product name	Version	Reference
A gridded version of the reference data set, excluding the radiometer measurements (described in Section 10)		Section 4
HadSST3	HadSST3 is version 3 of HadSST	Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011). Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 1: measurement and sampling errors. J. Geophys. Res., 116, D14103, doi:10.1029/2010JD015218 Kennedy J.J., Rayner, N.A., Smith, R.O., Saunby, M. and Parker, D.E. (2011). Reassessing biases and other uncertainties in sea-surface temperature observations since 1850 part 2: biases and homogenisation. J. Geophys. Res., 116, D14104, doi:10.1029/2010JD015220 http://www.metoffice.gov.uk/hadobs/hadst3/
AVHRR Pathfinder	Version 5.2	http://www.nodc.noaa.gov/SatelliteData/pathfinder4km/
HadISST	Version 1 (version 2 will also be used, if it is available)	Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., Kaplan, A., 2003, Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century J. Geophys. Res. Vol. 108, No. D14, 4407 10.1029/2002JD002670; http://www.metoffice.gov.uk/hadobs/hadisst
ERSSTv3	Version 3	Smith, T., R. Reynolds, T. Peterson, and J. Lawrimore (2008), Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006), Journal of Climate, 21 (10), 2283–2296, doi:10.1175/2007JCLI2100.1. http://www.ncdc.noaa.gov/oa/climate/research/sst/ersstv3.php
Kaplan	Version 2	Kaplan, A., Cane, M.A., Kushnir, Y., Clement, A.C., Blumenthal, M.B., Rajagopalan, B., Analyses of global sea surface temperature 1856-1991, Journal of Geophysical Research-Oceans, 103, C9, 18567-18539, 1998
Cobe SST		Ishii, M., Shouji, A., Sugimoto, S., Matsumoto, T., 2005, Objective Analyses of Sea-Surface Temperature and Marine Meteorological Variables for the 20th Century using ICOADS and the KOBE Collection. Int. J. Climatol., 25, 865-879.
NOCS Surface Flux Dataset v2.0	Version 2.0	Berry, D. I., Kent, E.C., 2009, A New Air–Sea Interaction Gridded Dataset from ICOADS With Uncertainty Estimates. Bull. Amer. Meteor. Soc., 90, 645–656; http://www.noc.soton.ac.uk/ooc/CLIMATOLOGY/noc2.php
Karspeck		Karspeck, A. et al 2011: Bayesian modelling and ensemble reconstruction of mid-scale variability in North Atlantic SSTs for 1850-2008 QJRMSS doi:10.1002/qj.900
OI.v2		Reynolds, R. W., Rayner, N.A., Smith, T.M., Stokes D.C., Wang, W., 2002, An improved in situ and satellite SST analysis for climate. J. Climate, 15, 1609-1625; http://www.emc.ncep.noaa.gov/research/cmb/sst_analysis/
MyOcean OSTIA reanalysis	Version 1	Roberts-Jones, J., E. Fiedler and M. Martin, 2012: Daily, global, high-resolution SST and sea-ice reanalysis for 1985-2007 using the OSTIA system, submitted to J. Climate
NOAA Optimum Interpolation 1/4 Degree Daily Sea Surface Temperature Analysis - AVHRR and AMSRE OI	Version 2	Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S., Schlax, M.G., 2007, Daily High-Resolution-Blended Analyses for Sea Surface Temperature. J. Climate, 20, 5473–5496. doi: 10.1175/2007JCLI1824.1;

Product name	Version	Reference
		http://www.emc.ncep.noaa.gov/research/cmb/sst_analysis/
NOAA Optimum Interpolation 1/4 Degree Daily Sea Surface Temperature Analysis - AVHRR OI	Version 2	Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S., Schlax, M.G., 2007, Daily High-Resolution-Blended Analyses for Sea Surface Temperature. <i>J. Climate</i> , 20, 5473–5496. doi: 10.1175/2007JCLI1824.1; http://www.emc.ncep.noaa.gov/research/cmb/sst_analysis/

Table 9-2: Data sets and analyses used to compare to the long-term ECV at low resolution.

The above sets of diagnostics are calculated using SST anomalies. These are aggregated deviations of each SST value from a reference climatology. We will use a common reference climatology and use it to calculate SST anomalies from the long-term ECV and the comparison data sets to ensure that we find any differences between them. The common reference climatology will be that for 1985-2007 calculated from the 0.05°, daily OSTIA Reanalysis, version 1; unless a new option becomes available in the meantime, prompting us to re-evaluate our choice.

When comparing diagnostics, it is important to recognise that SST data sets and analyses contain inherent uncertainties due to measurement errors, inadequate sampling of variability, adjustments applied to reduce the effects of relative biases, analysis methodologies, etc. Hence, at best, they can be expected to agree only to within the limits set by these uncertainties. Therefore, where possible, we will use the estimates of uncertainty provided with the comparison data sets and analyses to assess uncertainties in our calculated trends and other diagnostics. It is often the case, however, that uncertainty estimates can be underestimated. We will ameliorate the effect of this through the use of our ensemble of comparison data sets and analyses. The spread of the results obtained from the ensemble can represent the uncertainties that are neglected in the assessment of uncertainties in one data set or analysis. However, this is not always the case, particularly if all data sets hitherto have neglected the effect of a particular uncertainty, but it does represent the combined understanding of the whole community.

The long-term ECV will have length of 19 years 4 months. In the context of global and regional climate and climate change, this is a relatively short period. Therefore, we will exploit the over-150 year SST record provided by data sets and analyses that incorporate measurements of SST made in situ to set the trends and variability seen in the long-term ECV in their longer term context.

9.1.2 Engagement of others

We will involve independent experts from the GCOS SST Working Group and the GHRSSST Reanalysis, Intercomparison and Validation Technical Advisory Groups in the comparison of the long-term ECV with other SST data sets and analyses. It is likely that they will compare the long-term ECV with the same data sets and analyses that we do, but their approach will be different from ours. Indeed, we will ensure that there is no duplication of effort and encourage alternative approaches. It is not possible at this stage to detail exactly what their approach will be.

In addition, many of the potential users of the long-term ECV are experts in the analysis of climate variability. We will invite them to use the CDRP in their own analyses and feed back their findings.

The timetable for engagement is as follows:

- MARCDAT-III [May 2011]: engage GCOS SST WG and other workshop attendees for independent validation of the products [completed]
- GHRSSST-XII [June 2011]: engage RANTAG/ST-VAL for independent validation of the products [completed]
- GHRSSST-XIII [June 2012]: plan how RANTAG/ST-VAL will provide feedback on validation plan of the products
- Consultation with small subset of the best-engaged users. Obtain feedback on a preview of the products [October 2012]
- CLIMAR-IV: gather feedback from GCOS SST WG and others on independent validation of the products (if CLIMAR-IV not available, then a specific discussion will be convened) [May 2013]
- GHRSSST-XIV [June 2013]: gather feedback on independent validation of the products from RANTAG/ST-VAL

9.2 Assessment of the impact of SST_CCI products on climate model simulations

SST plays a crucial role in the modelling of climate, as it describes an important aspect of the interface between the ocean and the atmosphere. To achieve an effective simulation of global climate, either in the context of a shorter term seasonal prediction, or a medium term decadal prediction or a longer term 50-100+ year climate projection, it is vital to correctly model the fluxes of heat and moisture between the atmosphere and the ocean. SST gives an indication of the upper ocean temperature, which partly controls and partly reflects these fluxes.

SST observations are used in different ways, depending on the particular climate research activity:

- They are assimilated into a time-varying analysis of ocean state that provides the initialisation of seasonal and decadal forecasts, both operationally and in assessment, or hindcast, mode;
- They are used as a source of “truth” in the assessment of coupled ocean/atmosphere model simulations, which do not involve data assimilation. We will directly test the impact of SST_CCI products in this activity within this project (see below);
- They are used in climate change detection and attribution analyses, where observed changes (here in SST) are attributed to natural or man-made causes using an assessment of the extent to which the signal of man-made or natural effects can be seen to emerge from natural variability;
- Level 4 analyses of SST are used as lower boundary conditions for atmosphere-only simulations or Reanalyses (i.e. including atmospheric data assimilation), prescribing actual SST changes, to allow a close representation of atmospheric variability and change over the period of the SST analysis;
- They are used in climate change impact assessments, e.g. in studies on coral bleaching or other marine ecosystem changes;

- They are used to explore and understand large and small scale climate phenomena through statistical analysis etc.

In the following sections, we will detail how we plan to assess the impact of the SST_CCI products in a climate modelling framework and how we will involve others to enable an assessment of impact over the range of possible applications.

9.2.1 Our analysis

As new coupled climate models are developed, they are tested to assess the realism of the simulated climate. This is done in a standard framework to facilitate the comparison of different model versions and sub-versions. The realism of the SST and its variability is a key aspect of the realism of the climate simulation as a whole.

The current development version of the Met Office Hadley Centre climate model is HadGEM3. The validity of the SST as simulated by HadGEM3 is currently assessed by comparison to the HadISST1 and Reynolds et al daily OI analyses. The SST analysis is re-gridded to the model spatio-temporal grid to facilitate comparison. Currently, the ocean component of HadGEM3 is run on two different resolutions: 0.25° latitude by longitude and 1° latitude by longitude. The simulations are performed using a 360-day calendar, so the SST observations are first converted to this calendar.

Basic assessment is done by comparing both the long-term means and variability of the observations and a “control simulation”. The control simulation provides a measure of the internal variability of the climate model and is forced using a fixed greenhouse gas concentration and recent start up conditions, but is then left to simulate internal modes of variability for decades, hundreds, or thousands of years depending on the resolution of the model and the associated cost. The control simulation is assessed to determine whether there are any model biases and to assess the various modes of variability, by comparison to their equivalents in the real world, as measured by the observations. Currently, long-term means of the Reynolds et al. (2007; RD.076) Daily OI have been used. Assessments of long term variability have been performed with reference to HadISST1 (Rayner et al., 2003; RD.074). Standard assessment methods have not yet used information on observational uncertainty. Simulated and observed SST are currently compared using a system of diagnostics, detailed in Sections 9.2.1.1 and 9.2.1.2 below.

We will utilise 20 years of simulated daily SST from the latest control simulation of HadGEM3 to explore the daily variability simulated by the model; comparing it to the level 4 analysis of the long term SST_CCI product. Good representation of daily variability is important, for example, in areas affected by the Asian Monsoon and in order to explore fast moving phenomena such as the Madden Julian Oscillation. By including information on the uncertainty in the long term level 4 analysis, we will be able to determine whether differences seen between the simulated and observed SST are statistically significant. Currently, this information is not used and it will be informative to see whether inclusion of this information changes any conclusions made about the efficacy of the model simulation, based on use of current SST analyses.

We detail below the diagnostics we will use as a baseline. We will also consider any further approaches that become possible in the meantime, resources permitting. We will compare the simulations to the SST_CCI long term level 4 analysis and also to the Reynolds et al Daily OI and HadISST1 (as appropriate) to determine the impact of the SST_CCI products on the assessment.

9.2.1.1 Time means

It is important to assess to what extent the time mean of the simulation agrees with a mean of the observations over an equivalent time period. As it is the control simulation that is assessed, there are no exact equivalents to the period of the observations used, but it is important to compare 20-year means of the observations to 20-year means of the simulation.

We will examine both calendar monthly and annual 20-year means, averaged over different regions (detailed in Table 9-3). The results are presented in diagrams, such as the example illustrated in Figure 9-2, which relates to an early development version of HadGEM3. The simulation under consideration is compared to a reference simulation, to determine whether or not the model version of interest is an improvement upon what has gone before.

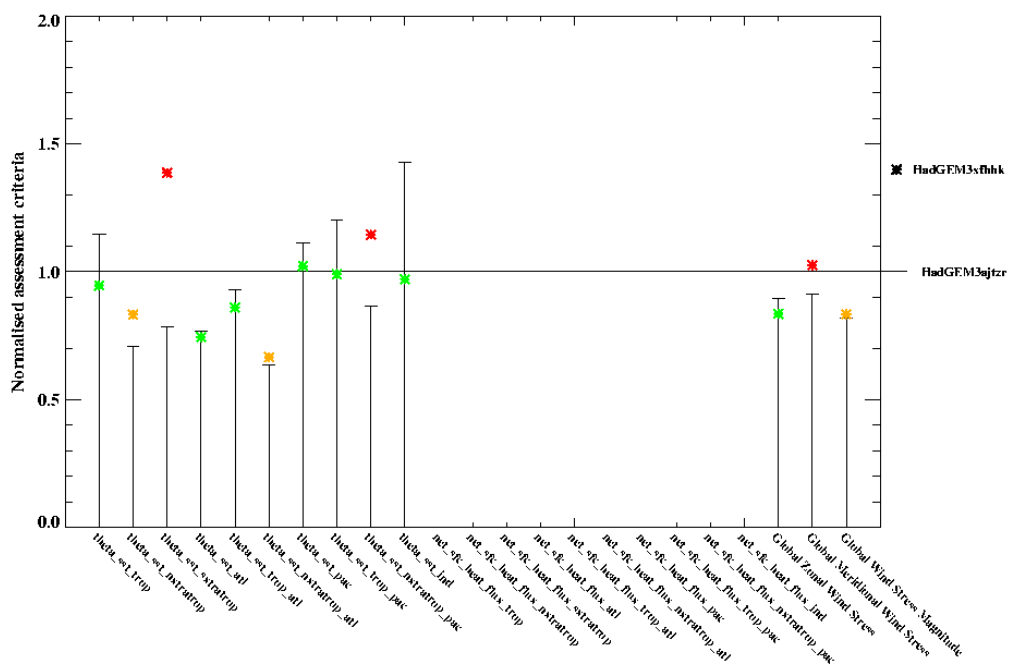


Figure 9-2: Example climate model assessment diagram, relating to an early development version of HadGEM3. Definitions of quantities are found in Table 9-3.

Area for 20-year mean	Metric name
Global SST	theta_sst
Tropical SST	theta_sst_trop
Northern extra-tropical SST	theta_sst_nxtratrop
Southern extra-tropical SST	theta_sst_sxtratrop
Atlantic Ocean SST	theta_sst_atl
Tropical Atlantic SST	theta_sst_trop_atl
Northern extra-tropical Atlantic SST	theta_sst_nxtratrop_atl
Pacific Ocean SST	theta_sst_pac

Area for 20-year mean	Metric name
Tropical Pacific SST	theta_sst_trop_pac
Northern extra-tropical Pacific SST	theta_sst_nxtratrop_pac
Indian Ocean SST	theta_sst_ind

Table 9-3: Metrics used to assess mean SST field, corresponding to labels in Figure 13.1

An additional large-scale comparison against mean fields is made to assess the model against its own requirements:

Metric	Current comparison	Model is fit for purpose if
Global SST	diff. with HadISST /EN3/Reynolds	At least 70% of area within 1degC of climatology
Atlantic SST	diff. with HadISST /EN3/Reynolds	Ditto
Indian SST	diff. with HadISST /EN3/Reynolds	Ditto
Pacific SST	diff. with HadISST /EN3/Reynolds	Ditto
SO SST	diff. with HadISST /EN3/Reynolds	Ditto

Table 9-4: Comparison to long-term means, plus assessment of fitness for purpose

This is the basic component of the assessment. From here we will go on to assess the simulated variability.

9.2.1.2 Variability

Clearly, it is not sufficient that a model should simulate the mean climate well. Large and small scale phenomena, such as ENSO, affect the downstream ability of the model to make effective seasonal forecasts, regional climate predictions, etc. In the current model assessment framework, the following diagnostics are calculated:

Metric	How it is calculated
Atlantic SST variability	RMS of interannual SD of SST, rel. to HadISST. Calculate fraction of ocean surface where SD from model is 70-130% of observed SD, integrate across 4 seasons. To include subjective assessment of SST-covariability/EOFs.
Indian SST variability	ditto
Pacific SST variability	ditto
SO SST variability	ditto
ENSO index	SD of Nino 3 index

Metric	How it is calculated
ENSO index	Composites of DJF seasons where nino3 index exceeds 1.5sigma
Atlantic SST dipole index	Annual mn. StDev of the dipole index

Table 9-5: Metrics used to assess SST variability

However, the 20 years of data in the long-term products is insufficient to assess the ENSO variability in the simulation in a robust manner, so the ENSO diagnostics will not be assessed.

9.2.2 Engagement of others

Many of the potential users of the long-term ECV are climate modellers, as are the CMUG. We will invite them to use the CDRP in their own analyses and feed back their findings. If necessary, we will also ask the CMUG to help us identify potential users of the products who might be willing to assess their impact in their applications.

As the resources within the project for assessment of the impact of the products in a modelling framework are small, we rely on voluntary assessment of the products by engaged potential users. We will use resources set aside for user engagement to facilitate this process and try to ensure that test users come from a wide range of different applications.

During our user requirements gathering, we discussed with various users what they might be able to do with the SST_CCI products. The suggestions given below are adapted from Appendix B of the URD [RD.171]. These do not, at this stage, represent commitments from the potential users concerned, nor do they provide an exhaustive list of what could be done, but they give an idea of the kinds of things that could be explored:

- Investigation of air/sea coupling
- Seasonal hindcast for 1991-2010.
- Force latest AGCMs and compare to latest CGCMs. AMIP run using higher resolution SSTs.
- Use as a baseline for assessing SST biases in high resolution ocean model
- Start to validate models in currently data sparse regions, e.g. the Arctic and west tropical Pacific
- Look at areas without in situ, e.g. Arctic, Southern Ocean, west tropical Pacific. Blend with high resolution land temperature data and use for monitoring reports.
- Test IR vs IR/PM data sets under cloud. Radiation balance under clouds.
- Case studies, e.g. tropical cyclones. Could use 6-month prototype for this - testing model or validate from cold wake in the data set.
- High resolution AMIP runs (although question mark about what would be used for the sea ice), at 10-20 km resolution.

- Look at the how the data are different in diagnostics such as global temperature compared e.g. to reanalyses.
- Explore sensitivity of marine life to temperature
- Produce detailed maps of temperature around the UK.

The time table for engagement is as follows:

- Consult with CMUG at their meeting 1 [March 2011]
- Consult with CMUG at co-location meeting 2 [October 2011]
- WCRP conference [October 2011]: engage climate research community via Climate Data Records session. [Completed]
- Consult with CMUG at their meeting 2 [March 2012*]
- Consult with CMUG at co-location meeting 3 [October 2012]
- Identification of climate modellers who are interested in testing the products [November 2012]
- Dissemination of the CDRP and associated web-based presentations/interactions, when the products are presented to potential users and commitments obtained for suitability and impact testing [April 2013]
- Consult with CMUG at their meeting 3 [March 2013*]
- Series of web-based presentations/interactions with climate research users who have taken up CDRP. Here they can present their work and we can gather their findings [May 2013]. Identification of material for peer reviewed publication.
- Present findings from Climate Assessment Report at CMUG meeting 4 [July 2013*]

* Estimated dates of future meetings as of the date of this document.

9.3 Assessment of the consistency of SST_CCI products with other CCI ECVs

SST has strong links with several of the ECVs to be produced as part of the CCI, namely ocean colour, clouds, aerosol, sea level and sea ice. In some cases, the best way to determine consistency between these ECVs is to assimilate them in a dynamic ocean reanalysis, such as those planned for MyOcean2. As the sea ice project is running more than one year behind the SST_CCI, we will not have an opportunity for intercomparison of SST and sea ice products. However, we do plan to assess the consistency of SST and ocean colour and invite others to assess the consistency of SST and other ECVs.

9.3.1 Our analysis

SST and ocean colour have strong linkages. In frontal regions, ocean productivity is high, leading to areas of high chlorophyll and high ocean colour. Comparisons between chlorophyll-a and SST fronts on a global scale are not appropriate as a means to assess consistency between SST and ocean colour products due to the differing physical and biological forcing mechanisms that govern their evolution. However, there are a number of regional cases where it is appropriate.

The first example is in the case of many shelf break fronts, where the fronts arise from the meeting of two water masses at a shelf edge, which have quite different thermal and biological properties. He et al. (2010, RD.260) describe an inverse relationship between chlorophyll-a concentration and SST at such a front.

A second example would be in the case of fronts in the Antarctic Circumpolar Current in the Southern Ocean (i.e. Polar Front, sub-Antarctic Front, Subtropical Front). In this example, the fronts interact with shallow topography/islands, generating vertical motions in the water column (a process known as topographic upwelling), which leads to the consistent development of ocean colour fronts in the form of bands of enhanced productivity, along the SST fronts and downstream of the topographic feature (Sokolov and Rintoul, 2007, RD261).

We will assess the co-location of ocean colour and SST fronts in case studies from such regions, as a means to examine the consistency of the two sets of products, over a short period common to both the demonstration and long-term products. This will be interesting because the demonstration product will incorporate lower resolution passive microwave retrievals, which might affect the analysis significantly in this respect. The inclusion of passive microwave retrievals might also permit us to detect fronts in persistently cloudy regions, which might not be as well observed in the long term products. We will use published frontal detection methods (e.g. Ullman and Cornillon, 2000, RD.247; Miller, 2009, RD.240) to identify the positions of case study fronts using SST and ocean colour separately. We will then assess, for these case studies:

- Whether there are fronts which are indicated in the ocean colour, but have not been captured in the SST products;
- the positions of those fronts, as determined using SST and ocean colour and report on any discrepancies; and
- The difference in representation of these fronts between the demonstration and long term products

9.3.2 Engagement of others

We will continue to discuss other potential areas of inter-comparison between ECVs with the other CCI teams and the CMUG at six-monthly CMUG and co-location meetings. In particular, we will identify inter-comparison activities planned by other CCI teams, which may have a bearing on the Climate Assessment Report for SST at co-location meeting 3 in October 2012. We will then discuss our initial results with the relevant CCI teams at co-location meeting 4 (currently planned for June 2013).

We will encourage potential users of the SST_CCI products who work in a multi-variable framework, either identified during our User Requirements gathering or with whom we have communicated as the project progresses to consider the SST_CCI products in conjunction with other CCI ECVs and feedback their findings on consistency between products to us via interactive web-based sessions, as discussed in Section 9.2.2.

APPENDIX A 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 algorithm selection, product validation, intercomparison, and climate assessment is given in Table A-1 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 algorithm selection are not included here if 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 PVP
Bias, precision, drift			
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 bias metric defined for algorithm selection will quantify bias across spatial scales determined by the statistical power of available validation data (see Section 6.3.2.1).
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 precision metric defined for algorithm selection will quantify precision across spatial scales determined by the statistical power of the available validation data (see Section 6.3.2.2).
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 stability metric defined for algorithm selection will quantify stability across spatiotemporal scales determined by the statistical power of the available validation data (see Section 6.3.2.3).

Requirement identifier	Requirement	Comments (from URD)	How we have addressed this UR in the PVP
SST_CCI-UR-QUF-51	At the threshold, breakthrough and objective requirement levels, the most common response for the acceptable drift in 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.	However, many users have stricter requirements.	The stability metric defined for algorithm selection will quantify day-night differences in bias across spatial scales determined by the statistical power of the available validation data (see Section 6.3.2.3).
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.		The stability metric defined for algorithm selection will quantify seasonal differences in bias across spatial scales determined by the statistical power of the available validation data (see Section 6.3.2.3).
Uncertainty information			
SST_CCI-UR-REF-4 / SST_CCI-UR-QUE-31	Uncertainties need to be characterised fully.	Characterisation of uncertainties needs to be improved relative to current datasets. This should include the full error budget of the translation from the input data to the products. [RD-3, RD-15]	The availability and validity of an SST uncertainty model for a given SST retrieval method is an algorithm selection criterion (see Section 6.3.2.1).
SST_CCI-UR-REF-7	Uncertainty characteristics should be verified by comparison against independent observations.	[RD-3]	Uncertainty estimates in products will be validated against independent measurements as part of the product validation (see Section 7.7).

Requirement identifier	Requirement	Comments (from URD)	How we have addressed this UR in the PVP
Requirements for features of the data			
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 7.4).
SST_CCI-UR-QUF-85	Independence from in situ measurements.	Classed as essential or preferable by 61% of respondents.	Degree of independence is a criterion in the algorithm selection (see Section 6.3.2.4)
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 7.5)
SST_CCI-UR-DIS-129	Backwards compatibility with older data is required	This would satisfy the needs of users who want to be able to use data from before the satellite era but also want to take advantage of the SST_CCI products, i.e. it is important that the two are consistent	The extent to which this is the case will be assessed in the Climate Assessment Report (see Section 9).

Table A-1: Summary of SST_CCI user requirements relevant to algorithm selection, product validation, intercomparison, and climate assessment.

APPENDIX B ADHERENCE TO CCI PROJECT GUIDELINES

The first collocation meeting of the ESA CCI was held at ESA ESRI, 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 B-1 summarises these recommendations and explains how each one has been addressed within the SST_CCI project.

Number	Recommendation	Adhered to in SST_CCI	Comment where required
RR1	The meaning of the 'best' algorithm and of how to select it (evaluation protocol) has to be defined before the start of the Round Robin exercise. The definition of 'best' and the scope of the Round Robin exercise have to be specified in the Product Validation Plan (PVP).	Yes	See Section 6.
RR2	(a) The Round Robin should be made at the beginning of the project based on objective criteria. (b) There should be one or more iterations to show algorithm improvement throughout the project. (c) The most objective algorithm selection would be based on blind testing to avoid any bias.	Yes to (a) and (c)	(b) does not fit the SST CCI approach (the duration of phase 1 is extended and includes all algorithm development) See Appendix C for (a) and (c)
RR3	Every CCI project has to perform a Round Robin exercise. In the exceptional case that a final algorithm has been pre-selected, separate modules need to be tested also for this preselected algorithm. Furthermore, the pre-selection criteria should be in line with the CCI objectives.	Yes	See Appendix C
RR4	The same auxiliary and Level 1 data should be used in the processing, as well as the same reference data.	Yes	The RR algorithm selection and the validation use a common MMD (see Section 5)
RR5	The round robin results need to be open and the algorithm must be well-documented and public, but the actual code does not need to be public.	Yes	All RR results will be made public and published in the PVASR [RD.226] and the ATBD [RD.225].

Number	Recommendation	Adhered to in SST_CCI	Comment where required
RR6	The algorithm selection should be made by an independent team that is not directly involved in the algorithm development, although of course the members of that team should be experts. The selection shall be made based on a Round Robin evaluation protocol developed beforehand and providing objective criteria.	No.	The EO team will select the algorithm in SST CCI, as agreed and presented at the first collocation meeting. Transparent selection criteria and open publication of methods will avoid any potential bias.
RR7	The development of new tools should only be considered when really needed and no good tools for the purpose are available.	Yes	Existing tools will be reused.
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: 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. 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. 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.	Yes	1. Product validation will use the reference dataset, which was not used in production. 2. Auxiliary datasets used for validation have not been used in production. 3. Most validation activities are carried out by personnel not involved in algorithm selection or product generation.
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	Protocol has been presented at GHRSSST XII. This document will be circulated to GHRSSST through the RAN-TAG and ST-VAL groups once accepted by the agency.

Number	Recommendation	Adhered to in SST_CCI	Comment where required
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.
V5	The CCI programme should hold a dedicated session (or workshop) on common validation infrastructure during (or prior to) the next co-location 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 will be presented at the 2012 meeting of GHRSSST (Tokyo), giving the PVP international scrutiny.

Table B-1: Summary of recommendations relevant to the round robin and product validation from the first CCI collocation and adherence within the SST_CCI project

APPENDIX C ROUND-ROBIN PROTOCOL

This section summarises the protocol for the SST_CCI round robin algorithm selection exercise. The text matches that previously approved for distribution to external participants to inform them of the RR protocol in RD.218.

C.1 Participation

C.1.1 Who can participate?

The SST_CCI round robin algorithm selection exercise is open to anyone who can contribute to and/or benefit from the development of better SST algorithms.

C.1.2 What do I gain from participating?

By participating in the SST_CCI round robin algorithm selection exercise you will:

- See how results of your algorithm objectively compare with all participating algorithms
- Gain early use of an SST_CCI multi-sensor match-up dataset⁵
- Contribute to a major initiative for a SST climate data record
- Have opportunity to be a contributing co-author of a peer-reviewed paper
- Potentially provide the winning algorithm!

C.1.3 What am I expected to contribute?

All participants in the SST_CCI round robin algorithm selection exercise are required to contribute:

1. Retrieved satellite SSTs generated by your algorithm(s) based on the data provided by the project, associated uncertainties and sensitivities, for any sensor included in RRDP

The RRDP contains extracted satellite reflectances and brightness temperatures, data quality masks, cloud masks, NWP fields and RTTOV simulations, for multi-sensor match-ups between ATSR/AVHRR/SEVIRI and in situ data. You can provide retrieved SSTs, associated uncertainties and sensitivities, for one or all of the available sensors.

⁵ Within the SST CCI project, the multi-sensor matchup dataset (MMD) is a central capability for algorithm development, assessment and validation. MMD capability will be designed into the prototype system being specified and demonstrated for creating SST CDRs, and is intended to be a lasting innovation. The RRDP is a pre-prototype example of this concept.

2. A list of peer-reviewed references describing the algorithm for retrieval and uncertainty estimation, and a very brief technical note summarising the algorithm's
 - theoretical basis
 - degree of dependence on tuning to in situ data
 - generality
 - improvability
 - and difficulty of implementation (see below)

C.1.4 What commitment do I give?

All participants in the SST_CCI round robin algorithm selection exercise are required to:

1. Provide their own resources to cover their participation
2. Register their intention to participate with the SST_CCI team to gain access to the data
3. Agree to the SST_CCI Round Robin (RR) conditions of use⁶
 - To use the RR data package (RRDP) only for RR participation
 - To not redistribute the data to other parties without the permission of ESA or the original data provider, as appropriate
 - To acknowledge the assistance of the ESA CCI programme in any publication that is based upon the use of the SST_CCI Round Robin data.

Agreement to these conditions is implicit upon registration.

4. Download the RR data package and documentation from a dedicated download site
5. Deliver their contributions in the specified data format to a dedicated upload site by the date specified
6. Give permission for the SSTs, uncertainty estimates and calculated SST sensitivities to be made publicly available

In addition, any optional comments you wish to make regarding the RR exercise, the design and content of the RRDP, etc., will be welcomed.

⁶ For various data within the RRDP, the original data provider has given us permission to include and distribute the data within the RRDP only on these conditions

C.1.5 What happens next?

Once you submit your results the algorithm selection team led by the Science Leader will compare all submitted results on an equal basis using the pre-defined metrics described in the PVP (RD.216). The objective is to determine the preferred algorithms to implement for subsequent processing within the SST_CCI project to create two data records:

1. A long-term (1991-2010) dataset of ATSR and AVHRR to demonstrate a climate data record
2. A short-term dataset (six months within the period October 2010 to June 2011) of AMSR-E, ATSR, AVHRR, TMI and SEVIRI to demonstrate a climate service

All the results for different metrics and the outcome of the algorithm selection will be publicly available.

Information on how to submit your data to the exercise is given in Section C.4.4.

C.1.6 How will progress and results be reported?

Participants will receive periodic email updates about the progress of the Round Robin exercise.

The results of the SST_CCI algorithm selection exercise will be published in the Product Validation and Algorithm Selection Report (PVASR⁷). This document will be published on the SST_CCI website (<http://www.esa-sst-cci.org/>).

In addition, results will be submitted for publication in a peer-reviewed journal. We will contact you about your interest in co-authoring this publication.

C.1.7 Will the results and data be made public?

Yes, results will be made publicly available as follows:

- All results for algorithm selection metrics will be published in the PVASR on the SST_CCI website (<http://www.esa-sst-cci.org/>)
- Results will be prepared for peer-reviewed publication in consultation with participants
- The complete algorithm selection dataset (including the submitted SSTs, validation values, etc.) will be freely available online

⁷ PVASR is the mandated report title, but it is a misnomer: at this stage, the algorithms will not have been used to create products and no product validation will have taken place. The report will describe the results of algorithm selection metrics and the decision process for selecting algorithms on the basis of these results.

C.1.8 What if my sensor is not in the round robin data package?

The SST_CCI algorithm selection exercise is directed towards a set of specific sensors chosen for this initial project. If you wish to contribute data from a sensor not in the initial list then you may provide matched SSTs to the in situ locations provided in the RRDP.

Of course any additional sensors will not be included in the algorithm selection exercise but you will be able to compare your data against the sensors used in SST_CCI once the algorithm selection is complete and the dataset is made publically available.

C.2 Schedule

C.2.1 What are the time scales?

Seven milestones are defined for the SST_CCI round robin algorithm selection exercise. These are:

1. Launch of RRDP:
30th June 2011 during the GHRSSST-12 meeting (Edinburgh)
2. Release of training and test data:
23rd September 2011
3. Release of selection data:
15th December 2011
4. Submission of participant contributions:
31st January 2012
5. Start of algorithm selection:
1st March 2011
6. End of algorithm selection:
30th April 2012
7. Publication of results and data release:
1st July 2012

A Gantt chart summarising the SST_CCI round robin algorithm selection exercise schedule is shown in Figure C-1.

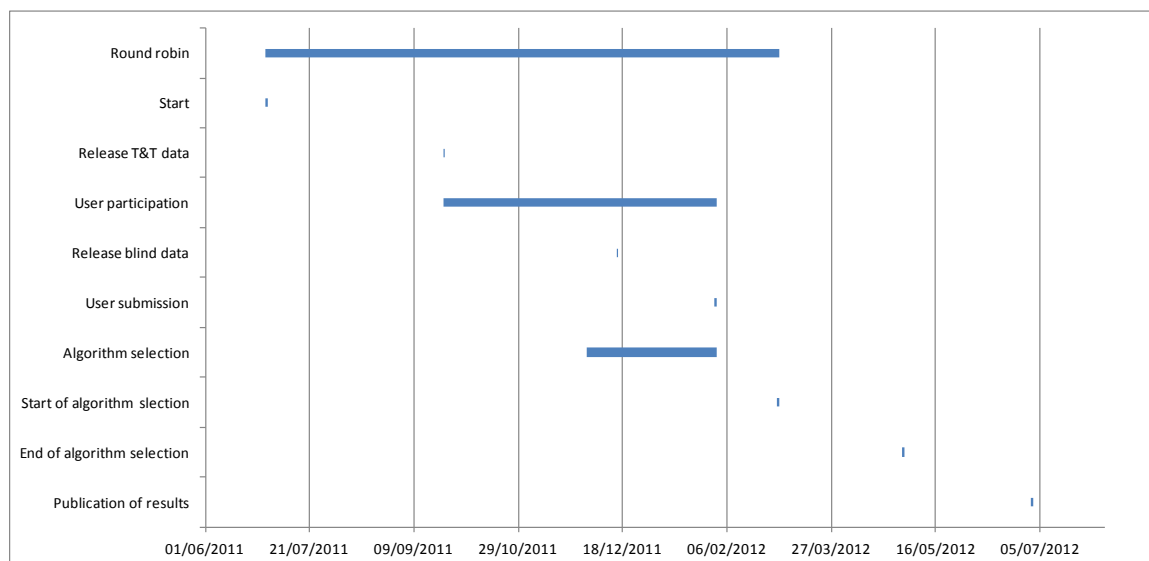


Figure C-1: Gantt chart showing schedule of RR

C.3 Experiment Design and Selection criteria

The experiment design and selection criteria for the Round Robin are explained in detail in Section 6 and are summarised below.

C.3.1 Experiment Design

Algorithms will be compared on a fair basis by standardisation of the approach:

- Competing algorithms will be developed using identified training data within the RRDP and will be compared by looking at their results when applied to test data within the RRDP. For fair comparison, test data must not be used at all in algorithm development; the test data will be reserved for use only after the algorithm is finalized. All participants for a given sensor and category must use the same training and test subsets in order to be considered within the exercise.
- Common metrics describing the results will be used for each type of algorithm to facilitate comparison of performance.
- Where objective/independent external data are available for validation, these will be used to compare performance, but in other situations, more subjective expert evaluation must be relied upon.

Algorithm selection requires joint assessment of a range of metrics and wider considerations. Not all properties of interest are quantifiable as metrics. Among measures that are quantifiable in principle, it may not always be feasible to undertake proper quantification within the scope of the project, and thus a qualitative approach may still be necessary.

For algorithm selection purposes, the validation data to be used are the matches flagged as “test data” in the RRDP. This will include at least two types of validation data (drifting buoys and moored buoys) for which results should be prepared separately. Certain test data will also be flagged as “high latitude” and “coastal” cases, and some metrics will be evaluated for performance separately using these subsets.

C.3.2 Selection Criteria

The selection criteria for the SST_CCI round robin algorithm selection exercise have been pre-defined before the start of the activity and are summarised below. All assessments will be carried out with reference to drifting buoys. Further details on each criterion can be found in Section 6 of this document.

For SST:

- Bias – the systematic difference from the truth, and is assessed via systematic differences from validation data
- Precision - observations generally differ from the truth according to a distribution that has a spread (or “dispersion”). The concept of precision is to characterise that dispersion, with a precise observation having a narrow spread
- Stability - constancy of bias in time. Stability of observation is critical when looking at differences between observations (i.e., changes of SST over time).
- Degree of independence - SST retrievals can be based on either empirical correlations to in situ observations, or on radiative transfer modelling. For applications where satellite SSTs are required to complement, enhance or test in situ observations, independence from in situ SSTs is an advantage (and in some cases a necessity).
- SST sensitivity - for the ideal SST estimate, changes in true SST are (on average) wholly reflected in changes in the estimated SST. Thus, $\partial \hat{x} / \partial x$ needs to be evaluated, where x is true SST and \hat{x} is the estimated SST. We refer to $\partial \hat{x} / \partial x$ as “SST sensitivity” for the SST estimate.
- Generality - the degree to which an algorithm is adaptable to other sensors and/or channel combinations, including future missions
- Improvability - the degree to which an algorithm can be further refined
- Difficulty of implementation – high, medium or low, based upon factors such as the use of external models and the size and nature of any required static or dynamic auxiliary files (e.g., look up tables, NWP, etc.)

For SST uncertainty, the following criteria will be used (definitions are as above):

- Bias
- Degree of independence
- Generality
- Improvability
- Difficulty of implementation

C.4 Data

C.4.1 What is in the round robin data package?

The SST_CCI RRDP contains the necessary satellite and auxiliary data to carry out SST_CCI round robin algorithm selection exercise. Three different subsets of data will be included:

1. 'Training': This is the subset that you should use for determining coefficients in empirically derived algorithms, for any other form of algorithm tuning, and/or as a training dataset in a supervised neural net optimisation (or similar). The training set is made available with validation (in situ) values. This subset of data is released at RRP milestone 2 (from 15th September 2011).
2. 'Test': This is the subset that you should use to get a (statistically) independent assessment of your algorithm developed using the training set -- it is "reserved data" for algorithm development, with validation values included. This subset of data is released at RRP milestone 2 (from 15th September 2011).

Ideally, you should use this type once any algorithm tuning is done; although if the test set performance is perceived as poor, it is recognised that this might prompt another cycle of training/test.

3. 'Selection': This subset will be distributed to participants (including the SST CCI developers) without validation values, and with fields sufficient only to derive SSTs, SST uncertainties and SST sensitivity for each "blind" matchup.

These derived quantities are the minimum set of data that you must submit for each sensor of interest.

This subset will be used to carry out the algorithm selection process and is released at RRP milestone 3 (from 15th December 2011).

In addition a fourth type of data, the 'validation' subset, is being retained to be used exclusively for product validation after system prototyping and product generation. (In ESA's SoW, it is the set referred to as "reference data".) No participants, including those responsible for algorithm selection, will have access to this subset of data prior to product generation.

C.4.2 How do I get the round robin data package?

The RRDP will be available for download from a secure FTP site. A detailed content specification of the RRDP can be found in RD.217.

To obtain login details for the RRDP download site you must email a request Gary Corlett (contact details given in Section C.5).

C.4.3 What data do I have to deliver?

All participants are required to deliver

1. Documentation: A brief technical note summarising the retrieval algorithm and uncertainty estimation, giving appropriate references.

The technical note should be submitted to the Science Leader (see Section C.5 for contact details) and can be submitted any time between the start of the RRDP (milestone 1) and the start of the algorithm selection exercise (milestone 5).

2. Data: Retrieved SSTs, uncertainties and SST sensitivities according to the specification given in Section 5.5 below.

Retrieved SST: This is the estimate of SST arising from your algorithm given the satellite observations provided

SST uncertainty: This is your estimate of the standard deviation of the distribution of error you expect for your retrievals. This may be a single estimate, or may vary between matches if you have a model for the variations in retrieval uncertainty.

SST sensitivity: This is an estimate of the responsiveness of your algorithm to a true change in SST, other factors being constant. Partial derivatives of brightness temperature with respect to SST are provided in the RRDP to facilitate consistent calculation of this estimate between participants.

C.4.4 How do I submit my data?

You will be required to submit your data to a secure FTP site. Further details will be provided as the exercise proceeds.

C.4.5 Format specification of participant contributions

All participant contributions must be in NetCDF version 3. As a minimum they must contain:

Variable name	Description
matchup.id	Unique MMD record number
<sensor_name>.<sen_id>.sea_surface_temperature	Retrieved SST

Optional fields are:

Variable name	Description
<sensor_name>.<sen_id>.sea_surface_temperature_uncertainty	Total uncertainty of retrieved SST#
<sensor_name>.<sen_id>.sea_surface_temperature_dSST_SST	Retrieved SST sensitivity to SST, i.e. dSST/SST

It is encouraged to provide uncorrelated (random), synoptically correlated (pseudo-random) and large-scale correlated (systematic) components of the SST uncertainty budget separately if they are known. In this case the fields should be named:

- sea_surface_temperature_uncorrelated_uncertainty
- sea_surface_temperature_synoptically_correlated_uncertainty

- sea_surface_temperature_large_scale_correlated_uncertainty

An example submission file will be provided to all participants.

Resource and time constraints will compel the SST_CCI round robin algorithm selection team to adopt a zero-tolerance policy to submissions with an incorrect format specification. Please contact the RR manager (see Section C.5 for contact details) if you have any questions regarding submission of data.

C.5 Important Contacts

The SST_CCI round-robin algorithm selection exercise is managed by the RR manager, Gary Corlett (gkc1@le.ac.uk). All technical enquiries should be directed to the RR manager in the first instance and copied to the project manager Paul Spinks (project.manager@esa-sst-cci.org).

The algorithm selection process will be led by the SST_CCI Science Leader, Chris Merchant (science.leader@esa-sst-cci.org). All scientific enquires should be directed to the SL in the first instance and copied to the Project Manager, Paul Spinks (project.manager@esa-sst-cci.org).

The Project Manager maintains a website (<http://www.esa-sst-cci.org/>) on which project documents are published.