



## Climate Data Assessment Framework

GHRSSST Document Reference CDR-TAG\_CDAF v 1.0.5

4 June 2014

Prepared by

Christopher J. Merchant, Jonathan Mittaz and Gary K. Corlett

## Document Change Record

Author	Modification	Issue	Rev	Date
CJM	Original draft	0.1	1	6 February 2013
JM & GKC	Comments on draft implemented	0.2	1	Feb 2013
CJM	Inclusion of Argo and possible extensions	0.3	1	Feb 2013
CJM	Update to make language consistent with language of “system maturity metrics” and “climate data assessments” adopted by CEOS WG Climate. Also Bates matrix removed, to preserve the above distinction.	0.4	1	22 Feb 2013
JM & GKC	Final comments prior to TAG consultation	0.5	1	23 Feb 2013
CJM	Updates in response to comments received during TAG consultation at <a href="http://podaac.jpl.nasa.gov/forum/node/54">http://podaac.jpl.nasa.gov/forum/node/54</a>	1.0	1	6 June 2013
CJM	Further comments received directly	1.0	2, 3	16 June 2013
CJM	Updates in light of comments from Sasha Ignatov: 1. Proposed to remove word “approval” throughout. Action: clarification that it is assessments, not datasets that are approved. 2. Proposed that strict independence of data used in evaluations should not be required. Action: explanatory note to be added; however, strict independence is required. 3. Proposed different time scales of evaluation. Action: none, since these are already listed in possible extensions. 4. Queried names in Section 4.3. Action: all names reviewed and improved. 5. Proposed to make some evaluations option. Not agreed, no action taken. 6. Proposed additional ways to present stability results. Action: added to list of possible extensions. 7. Proposed replicating all measures with conventional and robust statistics. Action: over-all histogram of differences added to metrics to give an impression of the degree of outliers and supra-Gaussian tails to data distribution.	1.0	4	4 November 2013
CJM	Updates following trial CDAF application as discussed in CDR-TAG session at GHRSSST-XV 1. More guidance on “valid data fraction”. 2. Clarify instructions for “dispersion relative to drifting buoys” 3. Clarify treatment of products with multiple SSTs 4. Clarify that high resolution GTMBA data are required	1.0	5	4 June 2014

## Table of Contents

<b>1</b>	<b>REQUIREMENT FOR CLIMATE DATA ASSESSMENT FRAMEWORK (CDAF)</b>	<b>4</b>
<b>2</b>	<b>PURPOSE AND SCOPE</b>	<b>4</b>
<b>3</b>	<b>OPERATION OF CDAF</b>	<b>4</b>
3.1	BASIC SCREE	5
3.2	GENERATE ASSESSMENT INFORMATION AND SUBMIT TO CDR-TAG	6
3.3	CDR-TAG REVIEW	6
3.4	APPROVAL AND PUBLICATION	7
<b>4</b>	<b>ASSESSMENT INFORMATION</b>	<b>7</b>
4.1	OVERVIEW INFORMATION	7
4.1.1	<i>Status of Assessment</i>	7
4.1.2	<i>Dataset name and version</i>	8
4.1.3	<i>Lead investigator and/or agency</i>	8
4.1.4	<i>Principal strengths of data set</i>	8
4.1.5	<i>Principal recommended applications</i>	8
4.2	KEY DESCRIPTIVE FEATURES	8
4.2.1	<i>Period covered</i>	8
4.2.2	<i>Geographic range</i>	8
4.2.3	<i>Spatial resolution</i>	9
4.2.4	<i>Temporal resolution</i>	9
4.2.5	<i>Timeliness of new data</i>	9
4.2.6	<i>Volume of dataset</i>	9
4.2.7	<i>Valid data fraction</i>	9
4.2.8	<i>Observation technology</i>	9
4.2.9	<i>Dependence on other data</i>	9
4.2.10	<i>Type(s) of SST</i>	10
4.2.11	<i>Traceability</i>	10
4.2.12	<i>Uncertainty info in product</i>	10
4.3	QUANTITATIVE MEASURES	10
4.3.1	<i>Systematic effects</i>	11
4.3.2	<i>Non-systematic uncertainty</i>	14
4.3.3	<i>Stability</i>	15
4.3.4	<i>SST sensitivity</i>	16
4.4	AVAILABILITY, DOCUMENTATION, FEEDBACK	18
4.4.1	<i>Data URL / ftp / DOI</i>	18
4.4.2	<i>Primary peer-reviewed reference</i>	18
4.4.3	<i>Dataset restrictions</i>	18
4.4.4	<i>Facility for user feedback</i>	18
4.4.5	<i>Other documentation</i>	18
4.5	CONFORMANCE TO GCOS MONITORING PRINCIPLES	18
<b>5</b>	<b>TEMPLATE FOR ASSESSMENT INFORMATION</b>	<b>20</b>
<b>6</b>	<b>POSSIBLE EXTENSIONS</b>	<b>23</b>
<b>7</b>	<b>LIMITATIONS</b>	<b>23</b>
<b>8</b>	<b>REFERENCES</b>	<b>24</b>

## 1 Requirement for Climate Data Assessment Framework (CDAF)

Within the Group for High Resolution Sea Surface Temperature (GHRSSST), the Climate Data Record Technical Advisory Group (CDR-TAG) accepts responsibilities relating to long-term, stable and accurate SST data sets. The primary responsibilities relevant to this Climate Data Assessment Framework (CDAF) are (CDR-TAG Terms of Reference v1.1):

1. Develop, regularly review and revise the GHRSSST community consensus on the requirements that must be met by products intended to be Climate Data Records (CDRs).
2. Define, document, maintain and improve the Climate Data Assessment Framework (CDAF) in conjunction with relevant international bodies.
3. Review, revise and approve assessments made of GHRSSST data sets that are proposed as CDRs. Maintain the authoritative list of assessment results, indicating which assessments have CDR-TAG approval.
4. Maintain CDR-TAG documents and information on the GHRSSST web site (<http://www.ghrsst.org>), including:
  - climate data assessment framework
  - authoritative source of CDAF outcomes

## 2 Purpose and scope

A CDR is "a time series of measurements of sufficient length, consistency and continuity to determine climate variability and change" (NRC, 2004), ideally traceable to SI standards.

The CDR-TAG is tasked to support users of sea surface temperature (SST) datasets to understand the suitability of GHRSSST datasets for use as Climate Data Records (CDRs).

This CDAF lays out how the CDR-TAG will discharge this responsibility by providing authoritative, comparable information about GHRSSST datasets that will allow users to make their own judgment about use of the datasets as CDRs for their application.

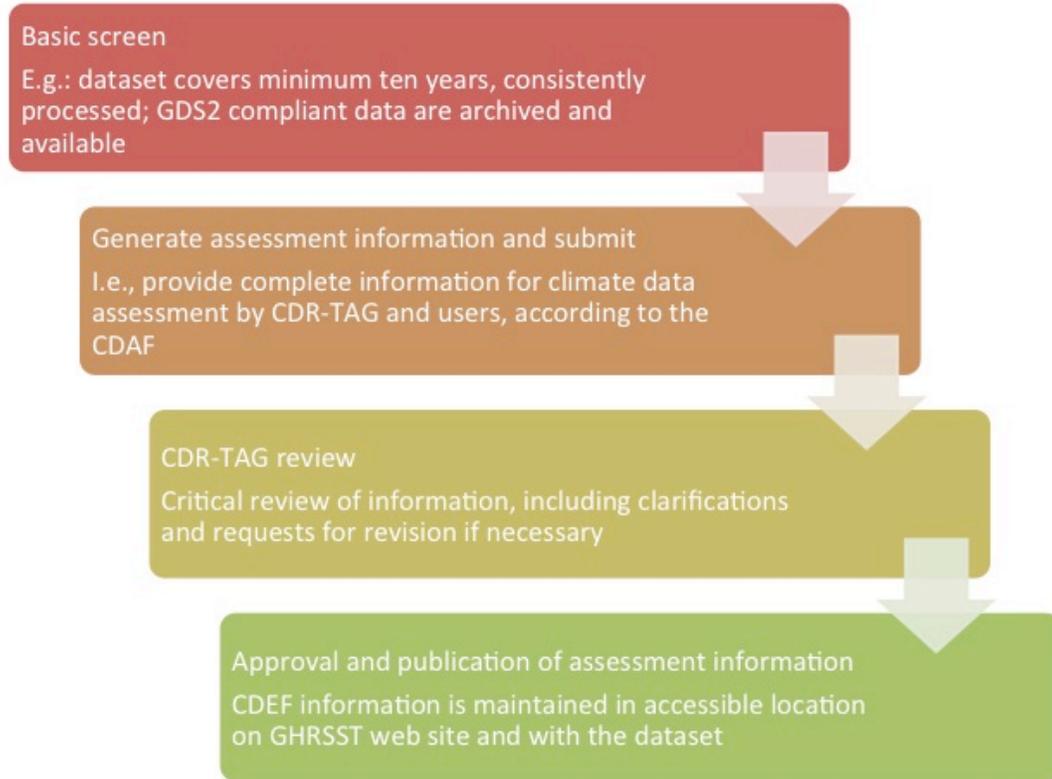
The datasets to which the CDAF is applicable are those derived largely or wholly from satellite SST estimates. **This present version of the CDAF is intended to be applicable to L2P and L3 satellite SSTs.** Blended L4 products will be addressed in a future version of the CDAF after experience is gained with swath and gridded data.

## 3 Operation of CDAF

This section describes the process by which the CDR-TAG will collectively review, revise and approve assessments GHRSSST data sets that are proposed as CDRs.

The operation of the CDAF is shown in Figure 1.

## Climate Data Assessment Framework



**Figure 1. Process for providing climate-user oriented assessment information for GHRSSST datasets within the Climate Data Assessment Framework (CDAF)**

### 3.1 Basic screen

A dataset here means a coherent collection of SST products, from one or multiple sensors.

Dataset producers identify whether their SST dataset passes some basic screening criteria. This requires affirmative answers to all of the following questions:

1. Is the dataset >10 years in length?
2. Is the dataset in its entirety freely available from the Long Term Stewardship and Re-analysis Facility (LTSRF) or from a sustained archive linked/discoverable from LTSRF pages?
3. Where multiple missions/sensors contribute to the dataset, have the data been harmonized? (This means: have relative SST biases between sensors been minimized, by exploiting overlap periods or by some other means?)

Datasets producers then inform the GHRSSST Project Office (GPO) and CDR-TAG chair of the existence of a dataset appropriate for assessment as a climate data record, and proceed to the next step, namely, self-assessment.

### 3.2 Generate assessment information and submit to CDR-TAG

The dataset must include the complete set of information specified in this CDAF, hereafter the “assessment information”, which is defined in detail in §4. The information overlaps with, and in certain cases exploits, criteria and indices developed outside of GHRSSST. This reduces the documentation burden on the dataset producer, since information can be reused. However, some of the assessment information may need to be developed specifically for consideration within the CDAF. Climate user requirements (e.g, Good and Rayner, 2011) justify the effort required to generate this information.

An important aspect for users will be confidence in the comparability of assessment information, particularly of quantitative measures of quality. The CDR-TAG has noted that a multi-sensor match-up system covering the GHRSSST constellation as a whole would provide such comparability, since measures could then be calculated using identical approaches. At present, a community-wide multi-sensor match-up system does not exist. At present, only a certain level of comparability can be achieved, by specifying relatively detailed principles for calculation of quantitative information within this CDAF.

Creating a community multi-sensor match-up system is a GHRSSST objective agreed at GHRSSST 2012. Within such a system, the exact *in situ* datasets used as reference points for metrics in this document could be controlled between different assessments for minimum effort on the part of dataset providers.

### 3.3 CDR-TAG review

The CDR-TAG *will* review the assessment information provided. The TAG will approve the assessment information as a fair and accurate representation to climate users of the nature of the dataset.

The CDR-TAG will consider:

1. Is the information complete?
2. Do we have confidence that quantitative measures are fair summaries of data quality?  
This may require technical review of the means of determination of these measures.
3. Is qualitative information given fair and accurate?
4. Is the information consistent and comparable with previously approved cases? This is arguably the paramount consideration, and perhaps the most difficult judgment required.

The CDR-TAG *will not* review or approve the dataset itself. Approving datasets is considered problematic. However, as experience with and confidence in the CDAF builds, the GHRSSST Science Team may in future reconsider whether it wishes to task the CDR-TAG to identify some

datasets as GHRSSST CDRs using the assessment information in comparison to climate user requirements.

Reviews will be carried out on a rolling basis by e-mail correspondence, to address any detailed questions and revisions effectively. Review decisions will be formally concluded at meetings of the CDR-TAG, usually at annual GHRSSST meetings. It is expected that formal review decisions will be made by consensus of the CDR-TAG. CDR-TAG meetings are open to all attendees of GHRSSST, however, only listed CDR-TAG members not involved in creation of the dataset under review will participate in voting on review decisions, if any voting is required.

### **3.4 Approval and publication**

After the CDR-TAG has approved the assessment information, the GHRSSST web site will add the dataset to a maintained list of GHRSSST datasets that have undergone climate data assessment. The list will link to the approved assessment information and to the data set.

Since it may be useful to climate users, candidate datasets will also be listed and the assessment information linked. However, it will be made clear that the information has not yet been CDR-TAG approved, and that therefore the information is not verified as comparable with information for other datasets.

## **4 Assessment Information**

This section describes the information required for assessment within the CDAF. Some headings are self-explanatory and the sub-section contains no further text.

### **4.1 Overview information**

#### ***4.1.1 Status of Assessment***

This indicates the status of the CDR-TAG's consideration of a particular GHRSSST dataset and its assessment information.

The status can be either:

“Dataset producer’s self-assessment against CDAF version 1.0, which is not yet verified or approved by GHRSSST.”

or

“Assessment information has been verified against CDAF version 1.0 and has been approved by GHRSSST.”

#### **4.1.2 Dataset name and version**

Long and short official names of datasets, recommended for users to use when presenting use of the dataset in presentations, papers, etc.

#### **4.1.3 Lead investigator and agency**

Need a specific, up-to-date named contact.

#### **4.1.4 Principal strengths of data set**

This should be a brief headline statement of what makes this dataset worthwhile creating for use in climate. For example:

“35 year continuous record, harmonized across missions by exploiting overlap periods.”

“12 year record with fully resolved diurnal cycle.”

Avoid claims relative to other datasets (e.g., “best available precision of any SST record”). Even if justified, such a claim may be superseded, and users can compare the assessment information to form such judgments themselves.

#### **4.1.5 Principal recommended applications**

A brief headline statement of the particular uses for which this dataset is suited. For example:

“Coastal zone applications requiring high resolution data with good long-term stability.”

### **4.2 Key descriptive features**

#### **4.2.1 Period covered**

Give start and end dates in standard form (UTC). For example:

“06-11-1994 to 29-02-2004”

or “26-01-1999 to present (dataset extended monthly)”

If there are any significant (>4 week) data gaps, add a footnote stating what these periods are.

#### **4.2.2 Geographic range**

For example:

“Global (between 82.5°N and 82.5°S)”

### **4.2.3 Spatial resolution**

In the simplest case, the spacing of pixels/data in the dataset and the true resolution of the pixels/data correspond, and this could be:

“0.05°, regular lat-lon grid”

Swath data are often more complicated. Be as simple as possible without misleading, e.g.:

“Regridded to a 1 km grid, underlying pixel resolution between 1 and 5 km”

rather than

“1 km”.

### **4.2.4 Temporal resolution**

E.g., “Daily coverage (typically 95% of ice-free ocean is viewed once or more in each 24 h period).”

### **4.2.5 Timeliness of new data**

Is the dataset extended regularly in time as new satellite observations are made? Within what time is the LTSRF generally populated with new observations? (If there are preliminary and final versions made available with different latencies, briefly describe this.)

### **4.2.6 Volume of dataset**

Quote the data volume as stored on LTSRF, plus additional volume per year if relevant.

### **4.2.7 Valid data fraction**

For a swath (L2) product, what fraction of ocean pixels in a typical file are flagged as giving valid SST. E.g., “SST from clear-sky observations only, yielding valid SST in ~X% of ice-free ocean pixels on average.”

For a gridded (L3) product, what fraction of ocean grid cells in a typical file have valid SST data.

### **4.2.8 Observation technology**

Nature of the sensor(s) used, and of major algorithms applied in dataset creation.

### **4.2.9 Dependence on other data**

All satellite processing chains have dependencies on data beyond the satellite observations themselves. Here, the point is to identify any direct dependencies in the SST estimation algorithm, such as that the SST algorithm is regressed to drifting buoys.

#### **4.2.10 Type(s) of SST**

Use GHRSSST nomenclature. Point out any subtleties in a footnote: e.g., “The retrieval algorithm is regressed to drifting buoys, therefore, on average the product nominally represents SST( $z=-20\text{cm}$ ). However, the satellite is in reality sensitive to SST-skin, which has somewhat different geophysical variability.”

#### **4.2.11 Traceability**

Brief statement of the degree to which the SSTs are traceable to SI standards. (Further guidance on how to assess this will be added to a future CDAF revision.)

#### **4.2.12 Uncertainty information in product**

Brief description of uncertainty information provided *in the product* (and not uncertainty information published elsewhere, for example). Indicate whether the uncertainty information is provided per observation (distinguishing high and low uncertainty observations), or generically for the product as a whole.

The “standard deviation” element of the GHRSSST “sensor specific error statistic” is an acceptable estimate of uncertainty that is presently provided in GHRSSST products (see GDS2.0), e.g.:

“Standard deviation of differences from matched drifting buoys, stratified by product confidence level.”

The ‘standard uncertainty’ quantifies the uncertainty associated with a given effect (a given source of errors) as the standard deviation of the estimated error distribution. If other quantifications of uncertainty are provided (e.g., confidence intervals), this will need detailed description in a footnote.

The “SSES bias” is not uncertainty information, nor, in the sense intended here, are quality flags uncertainty information.

The form of uncertainty information provided is to be stated here, not the value(s).

### **4.3 Quantitative measures**

One of the main motivations for development of the CDAF, rather than adoption of frameworks developed elsewhere, is the importance of specific quantitative measures relevant to the climate quality aspects of SST datasets. These quantitative measures need to be comparable between different sets of assessment information. As discussed above, the cleanest way to achieve this is to have a community-wide multi-sensor system from which to derive such measures in a common approach. In the absence of this, it is necessary to describe and circumscribe the way in which quantitative measures are derived. The measures proposed are doubtless imperfect and should be further developed. Nonetheless, they are intended to represent progress in providing quantitative, comparable measures of the climate quality of SST datasets.

### 4.3.1 Systematic effects

The intention here is to provide users with indications of the degree to which the SST in the product at a given location may differ from the truth on average, i.e., representing the uncertainty associated with effects that are systematic.

This intention is problematic for three reasons:

1. We have no globally distributed reference data with negligible errors against which satellite SST biases can be tested, and must accommodate use of what validation data are available.
2. Any averaging of SST must assume some relevant space and time scale over which to average, so a choice of space and time scale(s) must be made.
3. For comparability of assessment information derived for different datasets, strict independence of satellite and *in situ* measurements is needed.

Point 3. may exclude some metrics for some GHRSSST products in the assessment information, where products are tied to *in situ* data and therefore are not independent. The assessment information will then include an explanatory note to this effect.<sup>1</sup>

The Global Climate Observing System (GCOS, 2011) statement of requirements for satellite SST is that “accuracy” should attain “0.1 K over 100 km scales”, noting that “some ... datasets may approach 0.1 K accuracy on a global average basis but have biases >0.5 K for many important regions”. This GCOS accuracy requirement therefore appears to be a statement about the acceptable systematic effects. The relevant space-scale is 100 km, according to GCOS. The present density of validation values available (mainly drifting buoys) is insufficient to support assessment at this space scale. However, Merchant et al (2009) have argued that satellite SST biases can be assessed using drifting buoys on space scales of 1000 km. Assessments on coarser space scales can be made using other validation data sets. Particularly important here are the uppermost SST measurements (typically at ~4 m depth) from Argo profiling floats, which, following a decision of the ST-VAL at GHRSSST XII Science Team Meeting in 2011, are recommended to be reserved for assessment of GHRSSST products and not used in product development (e.g., for tuning algorithms or blended in products). This means there should always be at least one *in situ* assessment that can be made.

As for time scale, none is mentioned by GCOS, the emphasis apparently being on the acceptable degree of geographical variation in SST bias. Some datasets concatenate several satellite sensors to create harmonized data, and here the full period of the time series would seem to be implied. However, averaging over the subset of the full period relevant to each satellite sensor is also informative in this situation.

---

<sup>1</sup> A dataset producer wishing to tie products to *in situ* observations may elect to reserve some data specifically to enable a comparable, independent assessment. This is best practice and is to be encouraged where *in situ* matches are sufficiently plentiful. It is important that the reservation of data is done appropriately, e.g., by reserving all matches for certain buoy IDs for assessment purposes.

On the basis of the above considerations, we can define some measures of systematic differences intended to evaluate datasets against this GCOS requirement to the degree possible.

#### 4.3.1.1 Systematic differences relative to drifting buoys

First, the overall systematic difference from drifting buoys is reported, using the global median of the satellite minus drifting buoy SST difference, put in context regarding any geophysical difference in the type of SST, e.g.:

*Global median difference relative to drifting buoy SST. The satellite SSTs are  $SST_{skin}$  with no skin-effect adjustment, so a skin-effect difference of order -0.2 K is to be expected.*

The second measure is:

*Geographical variation in difference relative to drifting buoy SST, as described by the standard deviation of median satellite minus drifting buoy SST differences on space scales of ~1000 km*

We must account for the uncertainty in drifting buoy calibration, that is apparently ~0.2°C, such that the results are reasonably sound statistically. At least for missions since around 2005, this measure should be assessable for most of the global oceans at this spatial scale.

The steps involved are as follows:

1. Match satellite SSTs to drifting buoy observations<sup>2</sup>
2. Divide the dataset into subsets of 10° latitude by 10° longitude
3. Count the number of individual buoy IDs within each subset,  $n_{ID}$
4. For each subset, evaluate<sup>3</sup>  $\sigma_b = (0.2 \text{ K})/\sqrt{n_{ID}}$
5. Where  $2\sigma_b > 0.1 \text{ K}$ , the validation data are insufficient for a statistically sound difference to be calculated, so discard this subset<sup>4</sup>
6. For each retained subset, find  $\mu = \text{median}(x_{satellite} - x_{in-situ})$ , the median<sup>5</sup> of the satellite-validation SST differences.

---

<sup>2</sup> Criteria for matching should meet guidelines for satellite SST validation of the GHRSSST Satellite Sea Surface Temperature Validation TAG. The satellite match must be at full resolution: i.e., a single pixel/cell from the product rather a local average, etc.

<sup>3</sup> This quantity approximates the standard error in the mean of the validation data for the subset. It is valid under two assumptions: that inter-buoy biases are of order 0.2 K; and that there are many satellite matches per buoy ID (so that the standard error is dominated by the inter-buoy calibration differences). It neglects effects such as calibration drift over time for a given buoy.

<sup>4</sup> This is equivalent to requiring  $n_{ID} > 16$ . It is expressed as above to indicate the rationale: namely that the uncertainty in the mean of the validation data should be smaller than the GCOS target with a high degree of confidence (~95%).

7. Calculate the standard deviation of  $\mu$  across all subsets

Considering a full dataset built from a series of sensors (X, Y ...), the assessment information therefore may look like so:

Quantitative measure	Value	Comments
Difference relative to drifting buoys	-0.27 K	<i>Global median difference of satellite minus drifting buoy SST, across full dataset. The satellite SSTs are <math>SST_{skin}</math> with no skin-effect adjustment, so a skin-effect difference of order -0.2 K is to be expected.</i>
	-0.83 K	<i>As above, for contributing sensor X only. (Etc.)</i>
Geographical variation in difference relative to drifting buoys	0.32 K	<i>Geographical variation in bias, as described by the standard deviation of median satellite minus drifting buoy SST differences on space scales of ~1000 km, across the full dataset.</i>
	0.36 K	<i>As above, for contributing sensor X only. (Etc.)</i>

#### 4.3.1.2 Systematic differences relative to Argo measurements

This measure is equivalent to that described above for drifting buoys. Differences when using Argo measurements are: the density of matches is much less, which would tend to imply assessment on greater spatial scales; the Argo measurement calibration error is reportedly negligible for the purposes discussed here (S. Riser, GHRSSST X presentation, 2009); the deployment of floats ramped up between roughly 2001 and 2005, and therefore the time period for assessments is limited.

The steps involved are as follows:

1. Match satellite SSTs to Argo near-surface observations<sup>6</sup>
2. Divide the dataset into subsets of 20° latitude by 90° longitude
3. Count the number of matches within each subset,  $n$
4. For each subset, evaluate<sup>7</sup>  $\sigma = \text{stddev}(x_{\text{satellite}} - x_{\text{in-situ}})/\sqrt{n}$
5. Where  $2\sigma > 0.1$  K, the validation data are insufficient for a precise determination, so discard this subset

<sup>5</sup> The median rather than mean is recommended to minimize the effects of differences in matching criteria and in situ quality control between different dataset producers, since these should predominantly affect the tails of the distribution of differences.

<sup>6</sup> Criteria for matching should meet guidelines for satellite SST validation of the GHRSSST Satellite Sea Surface Temperature Validation TAG. The satellite match must be at full resolution: i.e., a single pixel/cell from the product rather a local average, etc. The shallowest available Argo SST should be used.

<sup>7</sup> This quantity approximates the standard error in the mean difference for the subset, which is dominated simply by the number of matches compared to the variability of the differences, as quantified by “stddev”, the standard deviation of the differences. (Note that use of  $n$  here implies that only one match to a given Argo profile should be included.)

6. For each retained subset, find  $\mu = \text{median}(x_{\text{satellite}} - x_{\text{in-situ}})$ , the median<sup>8</sup> of the satellite-validation SST differences.
7. Calculate the standard deviation of  $\mu$  across all subsets

Considering a full dataset built from a series of sensors (X, Y ...), the assessment information therefore may look like:

Quantitative measure	Value	Comments
Difference relative to Argo measurements	-0.27 K	<i>Global median difference of satellite minus upper Argo float SST, across full dataset. The satellite SSTs are <math>SST_{\text{skin}}</math> with no skin-effect adjustment, so a skin-effect difference of order -0.2 K is to be expected.</i>
	-0.83 K	<i>As above, for contributing sensor X only. (Etc.)</i>
Geographical variation in difference relative to Argo measurements	0.32 K	<i>Geographical variation in difference, as described by the standard deviation of median satellite minus upper Argo float SST differences on space scales of 20° latitude by 90° longitude, across the full dataset.</i>
	0.36 K	<i>As above, for contributing sensor X only. (Etc.)</i>

### 4.3.2 Non-systematic effects

The intention here is to provide climate users with information on the components of satellite-minus-*in situ* differences remaining after the systematic effects (whose distribution is quantified in §4.3.1 removed) are removed – i.e., information on the dispersion differences. Often, errors are considered as comprising systematic and random effects. The term “random” is avoided here since some satellite SST error components can be correlated across time and space because of effects associated with the state of the atmosphere.

The steps to estimate a measure of non-systematic uncertainty are as follows:

1. Using the same dataset(s) as in §4.3.1.1 and/or §4.3.1.2, subtract the appropriate  $\mu$  from each discrepancy:  $d' = (x_{\text{satellite}} - x_{\text{in-situ}}) - \mu$
2. Calculate the robust standard deviation of  $d'$  across the full dataset

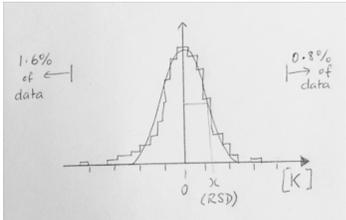
This standard deviation of differences over-estimates the uncertainty from non-systematic effects, because the measure includes effects arising from the imperfect *in situ*

<sup>8</sup> The median rather than mean is recommended to minimize the effects of differences in matching criteria and *in situ* quality control between different dataset producers, since these should predominantly affect the tails of the distribution of differences.

observations and true geophysical variability (differences from “point to pixel” comparisons, and difference in measurement times). The over-estimation is likely to be significant if the resulting value is comparable to 0.2 K (for drifting buoys) or 0.1 K (for Argo).

Use of the robust standard deviation is recommended to maximize comparability of measures between different sets of assessment information. The robust standard deviation is less affected by the outlier rate, which depends in turn on in situ quality control that varies between dataset producers. Users may also find it useful to be aware of the degree to which the distribution of differences is Gaussian. Thus the histogram of the differences should also be provided, over-plotted with the Gaussian curve associated with the robust standard deviation. Optionally, the conventional standard deviation and information about the outlier rate (differences exceeding 4 sigma) may be included on this plot.

The assessment information therefore may look like:

Quantitative measure	Value	Comments
Dispersion relative to drifting buoys	0.55 K 	<i>Spread of differences associated with non-systematic effects as quantified by a robust standard deviation of differences of satellite and drifting buoy data, after removing the geographical variations in differences quantified above</i>

### 4.3.3 Stability

Stability is the degree of invariance over time of the mean error from systematic effects in SST. In principle, it could be assessed by considering the time behavior of the  $\mu$  parameters from §4.3.1. This approach is supported by the GCOS statement that the stability requirement is “<0.03 K over 100 km scales”. (The time dimension is missing from this statement, but later text shows that absence of trend artifacts greater than 0.03 K per decade is the intended requirement (Ohring et al., 2005).)

However, the drifting buoy network is not known to be stable to this level, therefore this approach cannot be used with confidence. This arises because drifting buoys from various manufacturers are deployed in varying proportions over time, and are not generally traced to a common temperature calibration. Meanwhile, the Argo network is likely too sparse and covers too brief a period (at time of writing) to allow clear-cut assessments of decadal stability. Work is ongoing to assess the capability of using ship-borne radiometers in areas of repeat ship tracks for stability assessment looking at quantified uncertainties for the ship radiometer data and the length of ship record required to make a reliable stability assessment (Wimmer et al., 2012; Minnett and Corlett, 2012).

One assessment of stability capable of being informative at the level of the GCOS requirement has been published (Merchant et al., 2012). This assesses stability relative to the moored buoys of the Global Tropical Moored Buoy Array (GT MBA). These buoys are maintained, and are pre- and post-calibrated, to SI traceable standards (at least recently). Clearly, the stability assessment that is possible by this means only demonstrates stability in tropical regions. However, at present the GT MBA is the only reference available that provides reasonable confidence in its own stability and which is useful for datasets of more than a decade’s duration. Note that the GT MBA data required are of high temporal resolution ([http://www.pmel.noaa.gov/tao/data\\_deliv/deliv.html](http://www.pmel.noaa.gov/tao/data_deliv/deliv.html)); at time of writing, these are not fully present in ICOADs.

Stability assessment is rather complex. The published method uses step-change detection and auto-correlation assumptions that would be onerous to describe and demand of dataset providers. The steps below are simplified.

1. Identify GT MBA locations where buoy data (with post-calibration corrections applied) are available for >75% of the period to be assessed for stability<sup>9</sup>
2. Match satellite SSTs to these GT MBA measurements for the maximum possible number of complete years
3. Calculate the monthly median satellite-GT MBA difference for each location separately
4. For each month of the year and location, calculate the multi-year average of the monthly median satellite-GT MBA differences
5. Deseasonalize the monthly median differences at each location by subtracting the result of step 4 for the appropriate month of the year
6. Calculate the monthly mean difference across all locations. This results in a single satellite-GT MBA SST time series
7. Find the least squares fit to the time series of monthly mean differences
8. Quote the 95% confidence interval on the slope of the fit as the stability measure

The assessment information therefore may look like:

Quantitative measure	Value	Comments
Stability	-80 to +30 mK/yr	95% confidence interval for the relative multi-year trend between satellite SSTs and the Global Tropical Moored Buoy Array

#### 4.3.4 SST sensitivity

In general, remote sensing algorithms do not give results that are fully sensitive to true variations in the target geophysical variable, and SST retrieval is not exempt. The

<sup>9</sup> The longer the dataset, the fewer of the present GT MBA locations are available. For datasets that cover periods before 1986, the period prior to 1986 has to be excluded from consideration because of lack of mooring locations.

problem for a climate data record is that non-unity sensitivity indicates that part of the result actually derives from prior information and not from the satellite observations. Usually the degree of reliance on the prior information is variable across the product (as retrievals are more or less sensitive in different contexts) leading to complex and opaque prior error characteristics in the dataset.

SST sensitivity is calculable as:

$$\frac{d\hat{x}}{dx} = \sum_{c=1}^n \frac{\partial R}{\partial y_c} \frac{\partial y_c}{\partial x}$$

where:

Symbol	Meaning
$\hat{x}$	Estimated SST = $R(y_1 \dots y_n)$
$x$	True SST
$R$	The retrieval algorithm
$y_c$	The observation of the channel $c$ used for retrieval
$n$	The number of channels used for retrieval
$\frac{\partial R}{\partial y_c}$	The sensitivity of the retrieval process per unit variation in the observation in channel $c$ , which can be evaluated either by analytic partial differentiation of the retrieval algorithm or by numerical perturbation
$\frac{\partial y_c}{\partial x}$	The change in the observation in channel $c$ per unit change of true SST. This can only be estimated by radiative transfer calculation in practice.

The details of implementation must depend on the algorithm and on the radiative transfer simulation capability of the dataset producer. The principles are that sensitivity calculations should be undertaken for:

- A representative global sample of retrieval situations
- A representative range of viewing geometry
- Each algorithm used within the dataset
- Each parameter set (if algorithm parameters vary across the dataset)

The mean sensitivity across these calculations should then be reported.

The assessment information therefore may look like:

Quantitative measure	Value	Comments
SST sensitivity	87%	<i>Average weight of the satellite observations in determining SSTs in the dataset, the difference from 100% representing the weight of prior information in the SSTs</i>

## **4.4 Availability, documentation, feedback**

### **4.4.1 Data URL / ftp / DOI**

Provide each data web site, ftp site and/or doi information that is relevant. These should be hyperlinked directly from the document.

### **4.4.2 Primary peer-reviewed reference**

Give reference, and preferably hyperlink to open access paper or pre-print version.

### **4.4.3 Dataset restrictions**

Indicate any restrictions on use, license restrictions, prohibitions on further dissemination etc.

### **4.4.4 Facility for user feedback**

At minimum, the PI's contact e-mail address, or an official helpdesk, etc.

### **4.4.5 Other documentation**

Hyperlinks to Algorithm Theoretical Basis Document, Validation Report, Product Format Description, User Guide and similar documentation.

## **4.5 Conformance to GCOS monitoring principles**

It is assumed that datasets will largely conform to GCOS monitoring principles. Any discrepancies should be made clear to climate users.

The principles as adopted in December 2003 are:

1. The impact of new systems or changes to existing systems should be assessed prior to implementation.
2. A suitable period of overlap for new and old observing systems is required.
3. The details and history of local conditions, instruments, operating procedures, data processing algorithms and other factors pertinent to interpreting data (i.e., metadata) should be documented and treated with the same care as the data themselves.
4. The quality and homogeneity of data should be regularly assessed as a part of routine operations.
5. Consideration of the needs for environmental and climate-monitoring products and assessments, such as IPCC assessments, should be integrated into national, regional and global observing priorities.
6. Operation of historically-uninterrupted stations and observing systems should be maintained.
7. High priority for additional observations should be focused on data-poor regions, poorly observed parameters, regions sensitive to change, and key measurements with inadequate temporal resolution.
8. Long-term requirements, including appropriate sampling frequencies, should be specified to network designers, operators and instrument engineers at the outset of

- system design and implementation.
9. The conversion of research observing systems to long-term operations in a carefully-planned manner should be promoted.
  10. Data management systems that facilitate access, use and interpretation of data and products should be included as essential elements of climate monitoring systems.  
Furthermore, operators of satellite systems for monitoring climate need to:
    - (a) Take steps to make radiance calibration, calibration-monitoring and satellite-to-satellite cross-calibration of the full operational constellation a part of the operational satellite system; and
    - (b) Take steps to sample the Earth system in such a way that climate-relevant (diurnal, seasonal, and long-term inter-annual) changes can be resolved. Thus satellite systems for climate monitoring should adhere to the following specific principles:
  11. Constant sampling within the diurnal cycle (minimizing the effects of orbital decay and orbit drift) should be maintained.
  12. A suitable period of overlap for new and old satellite systems should be ensured for a period adequate to determine inter-satellite biases and maintain the homogeneity and consistency of time-series observations.
  13. Continuity of satellite measurements (i.e. elimination of gaps in the long-term record) through appropriate launch and orbital strategies should be ensured.
  14. Rigorous pre-launch instrument characterization and calibration, including radiance confirmation against an international radiance scale provided by a national metrology institute, should be ensured.
  15. On-board calibration adequate for climate system observations should be ensured and associated instrument characteristics monitored.
  16. Operational production of priority climate products should be sustained and peer-reviewed new products should be introduced as appropriate.
  17. Data systems needed to facilitate user access to climate products, metadata and raw data, including key data for delayed-mode analysis, should be established and maintained.
  18. Use of functioning baseline instruments that meet the calibration and stability requirements stated above should be maintained for as long as possible, even when these exist on decommissioned satellites.
  19. Complementary in situ baseline observations for satellite measurements should be maintained through appropriate activities and cooperation.
  20. Random errors and time-dependent biases in satellite observations and derived products should be identified.

Only those in the control of producers undertaking retrospective reprocessing of satellite datasets are appropriate to include in the CDAF. Conformance to GCOS principles is thus to be summarized in a short table, as per the following example:

GCOS monitoring principle	Comments
2. and 12. Overlaps between sensors exist and are exploited for harmonizing dataset	<i>No, the most recent sensor is used irrespective of overlaps.</i>
3. Detailed history of	<i>Yes, see <a href="http://www.sstremotesensing.com/history">www.sstremotesensing.com/history</a></i>

methods/algorithms is available	
11. Constant sampling within diurnal cycle	<i>Local time is constant to +/-1 hour in the raw observations, and variations are adjusted for</i>

## 5 Template for assessment information

The template for the assessment information is given overleaf. Note that the text in the “comments” column should be context specific, and text is given here to provide an example.

Some products contain more than one type of SST, or contain adjustments to transform a primary SST to other depths, times, etc. In these cases, the quantitative assessment metrics tables should be replicated the required number of times to give results for each relevant SST type, and explanation/annotation added as required summarizing the distinction between the different SSTs assessed.

## Information for Assessing [GHRSSST Dataset] as a Climate Data Record

Status of assessment:

Dataset name and version:

Lead Investigator and/or Agency:

Principal strengths of data set:

Principal recommended applications:

KEY DESCRIPTIVE FEATURES	INFORMATION
Period covered	
Geographic range	
Spatial resolution	
Temporal resolution	
Timeliness of new data	
Dataset volume	
Valid data fraction	
Data level / grid	
Observation technology	
Dependence on other data	
Type(s) of SST	
Traceability	
Uncertainty info in product	

QUANTITATIVE MEASURES	VALUE	COMMENTS
Difference relative to drifting buoys		<i>Global median difference of satellite minus drifting buoy SST, across full dataset. The satellite SSTs are SST<sub>skin</sub> with no skin-effect adjustment, so a skin-effect difference of order -0.2 K is to be expected.</i>
Difference relative to Argo		<i>Global median difference of satellite minus upper Argo float SST, across full dataset. The satellite SSTs are SST<sub>skin</sub> with no skin-effect adjustment, so a skin-effect difference of order -0.2 K is to be expected.</i>
Geographical variation in difference relative to drifting buoys		<i>Geographical variation in difference, as described by the standard deviation of median satellite minus drifting buoy SST differences on space scales of ~1000 km, across the full dataset.</i>
Geographical variation in difference relative to Argo measurements		<i>Geographical variation in bias, as described by the standard deviation of median satellite minus upper Argo float SST differences on space scales of 20° latitude by 90° longitude, across the full dataset.</i>
Dispersion relative to		<i>Spread of differences associated with non-</i>

<b>drifting buoys</b>		<i>systematic effects as quantified by a robust standard deviation of differences of satellite and drifting buoy data, after removing the geographical variations in differences quantified above</i>
<b>Stability</b>		<i>95% confidence interval for the relative multi-year trend between satellite SSTs and the Global Tropical Moored Buoy Array</i>
<b>Sensitivity to true SST</b>		<i>Average weight of the satellite observations in determining SSTs in the dataset, the difference from 100% representing the weight of prior information in the SSTs</i>

<b>AVAILABILITY, DOC'N, FEEDBACK</b>	
<b>Data URL / ftp / DOI</b>	
<b>Primary peer reviewed reference</b>	
<b>Source of technical documents</b>	
<b>Dataset restrictions</b>	
<b>Facility for user feedback</b>	
<b>Other documentation</b>	

<b>OTHER PRINCIPLES (GCOS)</b>	<b>COMMENTS</b>
<b>2. and 12. Overlaps between sensors exist and are exploited to harmonize the dataset</b>	
<b>3. Detailed history of methods/ algorithms is available</b>	
<b>11. Constant sampling within diurnal cycle</b>	

## 6 Possible extensions

The assessment information described in this version of the CDAF is intended as a useful overview for users to assess a dataset for use as a climate data record. The measures in particular are broad summaries of the quantitative characteristics of the dataset, and users may want to see more detail. These extensions in addition to the CDAF could be available online via ghrsst.org. The list below is of possible additions to the measures that may be useful to consider including in a future version of the CDAF:

- A set of informative plots “behind” the headline measures in the assessment table:
  - Maps of the systematic differences of satellite and drifting buoy SSTs (i.e., the spatial distribution behind the measure of §4.3.1.1).
  - Maps of the systematic differences of satellite and Argo SSTs (i.e., the spatial distribution behind the measure of §4.3.1.2).
  - Maps of the non-systematic uncertainty (i.e., the spatial distribution behind the measure of §4.3.2).
  - Dependence plots, e.g., of systematic effects against wind-speed, satellite zenith angle, etc
  - The time series underlying the stability confidence interval, showing the fitted line
- More statistical details “behind” the headline measures in the assessment table:
  - Histograms of distributions represented in table by one or a few numbers
  - Mean, standard deviation, minimum and maximum and 4-sigma outlier rates in addition to median and robust standard deviation
- Measures for additional aspects of stability
  - Measures of seasonal stability (mapped differences in systematic effects round the annual cycle).
  - Measures of day/night stability (mapped differences in systematic effects between day and night retrievals, especially important where different algorithms for SST are used in day and night).
- Additional description of sensitivity characteristics
  - Stratification of sensitivity by water vapour, zenith angle, etc.
  - Maps of sensitivity
- Measures for assessing the realism of product uncertainty information.
- Clearer principles for discussing the traceability issues of datasets.

## 7 Limitations

- Standardized tools for calculation of quantitative measures should be developed and available to the GHRSSST community
- Consistency of assessment information will be enhanced if a fully functional, all-sensor multi-sensor match-up capability is developed and used to source data for quantitative measures

- Research to characterize errors in validation datasets (drifting buoys, GTMBA and Argo) could allow more sophisticated treatment of the quantitative measures, including better backing-out of the in situ effects in the uncertainty estimates for the satellite dataset
- Experience in estimating SST stability and sensitivity are relatively limited across the GHRSSST community, so the advocated approaches should be considered provisional

## 8 References

- Good S and N Rayner, 2011, CCI Phase 1 (SST) User Requirements Document, [www.esa-sst-cci.org](http://www.esa-sst-cci.org), SST\_CCI-URD-UKMO-001.
- Global Climate Observing System, Systematic Observation Requirements for Satellite Based Data Products for Climate 2011 Update, GCOS 154, December 2011. <http://www.wmo.int/pages/prog/gcos/documents/SatelliteSupplement2011Update.pdf>
- Merchant, C. J., A. R. Harris, H. Roquet, and P. Le Borgne (2009), Retrieval characteristics of non-linear sea surface temperature from the Advanced Very High Resolution Radiometer, *Geophys. Res. Lett.*, 36, L17604, doi:10.1029/2009GL039843.
- Merchant, C. J., O. Embury, N. A. Rayner, D. I. Berry, G. Corlett, K. Lean, K. L. Veal, E. C. Kent, D. Llewellyn-Jones, J. J. Remedios, and R. Saunders (2012), A twenty-year independent record of sea surface temperature for climate from Along Track Scanning Radiometers, *J. Geophys. Res.*, 117, C12013, doi:10.1029/2012JC008400.
- Minnett, P. J. & Corlett, G. K. (2012). A pathway to generating Climate Data Records of sea-surface temperature from satellite measurements. *Deep Sea Research Part II: Topical Studies in Oceanography*, 77–80, 44-51.
- NRC (National Research Council) 2004. Climate Data Records from Environmental Satellites. Washington D.C.: National Academy Press. ISBN-10: 0-309-09168-3, [ISBN 978-0-309-09168-8](http://www.nrc.gov/reading_room/publ/publ/09168-3.pdf)
- Ohring, G., Wielicki, B., Spencer, R., Emery, B. & Datla, R. (2005). Satellite Instrument Calibration for Measuring Global Climate Change: Report of a Workshop. *Bulletin of the American Meteorological Society*, 86, 1303-1313
- Wimmer, W., Robinson, I. S. & Donlon, C. J. (2012). Long-term validation of AATSR SST data products using shipborne radiometry in the Bay of Biscay and English Channel. *Remote Sensing of Environment*, 116, 17-31.