AATSR Validation:

LST Validation Protocol

ESA Contract Number: 19054/05/NL/FF

Written by: P. Schneider[1] …….....
D. Ghent[2] …….....
G. Corlett[2] …….....
F. Prata[1] …….....
J.Remedios[2] …….....

[1] NILU – Norwegian Institute for Air Research
[2] Earth Observation Science, Space Research Centre, Department of Physics & Astronomy, University of Leicester

Approved by: P. Goryl …….....

Document Change Record

Version	Date of Issue	Comments
Issue 0 Revision 0	04/04/2012	First version sent to UoL
Issue 1 Revision 0	16/04/2012	First release to ESA

Summary

This document provides a classification and protocol for various methods of validating Land Surface Temperature (LST) derived from spaceborne thermal infrared instruments. Four categories of LST validation are established, namely (A) validation with in situ data, (B) radiance-based validation, (C) multi-sensor intercomparison, and (D) time series analysis. Each category is further subdivided into several classes, which approximately reflect the validation accuracy that can be achieved by the different approaches, as well as the complexity involved with each method. Advice on best practices is given for methodology common to all categories. Each category further gives recommendations on specific methodology that has proven to be valuable for each approach. Selection criteria used for distinguishing the accuracy classes are established for each category. Examples are provided for validation classes in each category, where available.

Table of Contents

1. Introduction

Land surface temperature (LST) and emissivity are important parameters for environmental monitoring and earth system modelling. They have important implications for various energy fluxes between the Earth surface and the atmosphere and are a very valuable indicator for the changing state of the land surface. LST has been observed from spaceborne instruments for several decades, including moderate-resolution sensors such as the Along-Track Scanning Radiometer (ATSR) series of instruments, the Moderate Resolution Imaging Spectroradiometer (MODIS), the Advanced Very High Resolution Radiometer (AVHRR), and higher resolution thermal infrared sensors such as the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER). Instruments in geostationary orbit such as the Spinning Enhanced Visible and Infrared Imager (SEVIRI) complement the LST data collected from satellites in low earth orbit. Upcoming instruments such as the Sea and Land Surface Temperature Radiometer (SLSTR) will continue the existing archive of global LST measurements.

Validation of LST is of crucial importance for estimating the accuracy of the operational products and understanding the potential and limitations of satellite observations of LST.

1.1. Scope and format of this document

This document is intended to suggest an organizational structure for the various LST validation techniques, to describe hierarchical quality/complexity class within each category, and to provide a description of the recommended methodology for performing LST validation using the various approaches. The recommendations are based on previous experiences of the LST validation community. The document is intended to be generic and applicable to all existing and future satellite instruments used for observing LST.

The format of this document is as follows: Section 2 gives definitions for various terms that are used throughout this document. Section 3 provides a summary of the background and the state of the art of LST validation. Section 4 subsequently recommends 4 major categories for LST validation approaches and suggests several hierarchical quality classes for each. Sections 5 through 8 then describe each category in detail, providing advice on methodology and outlining a protocol for each of the quality classes. Finally, Section 9 summarizes recommendations for the Quality Assurance Framework for Earth Observation (QA4EO).

1.2. Objectives

The primary objectives of this document are:

- To establish several categories classifying the primary approaches to validation of satellite-based LST.
- To establish several quality levels for each validation category.
- To describe a recommended LST validation methodology for each category and level.

2. Terminology and Definitions

A variety of definitions exists for terms associated with validation and uncertainty assessment, with different fields often adopting slightly different meanings of the same term. In the following we provide a brief list of the terms most often encountered for LST validation. For more general and statistical definitions of many related concepts it is recommended to consult the "Evaluation of measurement data - Guide to the expression of uncertainty in measurement" (GUM) (Joint Committee for Guides in Metrology, 2008), from which several of the following definitions are adopted.

Absolute bias	A systematic error between a measurement and the true value. This is of theoretical importance only here, as the exact true value of LST cannot be known due to measurement error.
Accuracy	Accuracy is defined as the degree of conformity of the measurement of a quantity and an accepted value or the "true" value.
Calibration	Calibration is the process of quantitatively defining the system response to known, controlled system inputs.
Discrepancy	Discrepancy described the lack of similarity between two measurements.
Emissivity	Emissivity describes a material's ability to emit the thermal energy which it has absorbed.
Error	Result of a measurement minus a true value of the measurand. Note that in ractice a true value cannot be determined and therefore a conventional true value is used instead (Joint Committee for Guides in Metrology, 2008).
Land Surf. Temp.	Land Surface Temperature (LST) is the radiative skin temperature of the land derived from solar radiation. It is a basic determinant of the terrestrial thermal behaviour, as it controls the effective radiating temperature of the Earth's surface.
Measurand	A measurand is the particular quantity subject to measurement.
Precision	Precision is the closeness of agreement between independent measurements of a quantity under the same conditions.
Protocol	A protocol describes a methodology used to carry out a specific operation such as a measurement, data comparisons, or data merging.
Relative error	The relative error is the error of measurement divided by a true value of the measurand.
Random error	Result of a measurement minus the mean that would result from an infinite number of measurements of the same measurand carried out under repeatability conditions (Joint Committee for Guides in Metrology, 2008).
Relative bias	A systematic error between measurements obtained from different data sources.
Reference Standard	Measurement standard designated for the calibration of other measurements standards for quantities of a given kind in a given organization or at a given location.
Sea Surf. Temp.	Sea Surface Temperature (SST) is defined as the water temperature of the ocean close to its surface. Satellite instruments measure the skin temperature in the uppermost micrometers versus buoys, which measure the bulk water temperature at various depths.

Systematic error	Mean that would result from an infinite number of measurements of the same measurand carried out under repeatability conditions minus a true value of the measurand (Joint Committee for Guides in Metrology, 2008).
True value	A true value is the value consistent with the definition of a given particular quantity.
Uncertainty	A parameter associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand, that is the value of the particular quantity to be measured (Joint Committee for Guides in Metrology, 2008).
Validation	The Committee on Earth Observation Satellites (CEOS) defines validation as the "process of assessing, by independent means, the quality of the data products derived from the system outputs".
Validation loop	The validation loop describes the iterative process between algorithm development and validation with the final goal of improving the output product.

3. Background

3.1. International Coordination Activities

A substantial amount of previous work has been done in the area of structuring and standardizing calibration and validation approaches within the field of Earth Observation. The Committe on Earth Observation Satellites (CEOS) coordinates civil spaceborne observations of the Earth, with the primary objective of optimizing the benefits of spaceborne Earth observations, serving as a focal point for international coordination of space-related Earth observation acitivities, and of exchanging policy and technical information to encourage the complementarity and compatibility of observation and data exchange system. Is it within the framework of these objectives that CEOS has established the Working Group on Calibration and Validation (WGCV), whose mission it is to ensure long-term confidence in the accuracy and quality of Earth Observation data and products, and to provide a forum for the exchange of information about calibration and validation activities.

Much of the technical work of WGCV is carried out by its six subgroups, two of which are the Infrared & Visible Optical Sensors Subgroup (IVOS) and the Land Product Validation Subgroup (LPV). Whereas the former focuses primarily on ensuring high quality calibration and validation of infrared and visible optical data and establishing an internationally harmonised calibration and validation strategy, the latter primarily focuses on quantitative validation of higher-level global land products acquired from Earth Observation satellites. The LPV further includes a Focus Group for Land Surface Temperature, which is particularly relevant for the purposes of this document.

The Quality Assurance Framework for Earth Observation (QA4EO) has been established by CEOS in order to contribute towards facilitating the vision of the Group on Earth Observations (GEO) for a Global Earth Observation System of System (GEOSS). The goal of QA4EO is to understand how the various datasets acquired by the Earth Observation community can work together and how the CEOS members can provide interoperability for developing integrated science products from multiple data sources. QA4EO has published documents on guiding principles (QA4EO, 2010) and a large number of other documents, for example on documentary procedures, reference standards, comparisons, and the expression of uncertainty. For a full list of the available documentation see http://qa4eo.org/documentation.html.

Despite these international efforts on coordinating and standardizing validation approaches for LST there has been no effort so far to create a comprehensive document that attempts to structure the various approaches to LST validation and to establish a hierarchical system of validation quality.

There have been previous documents describing a measurement protocol for validation of satellite-derived land surface products, for example the measurement protocol for in situ validation of AATSR, which has been provided as part of the AATSR validation plan (Parkes et al., 2005). However, this protocol was heavily focused on sea surface temperature (SST) and did not offer much detail in terms of a validation protocol dedicated to LST. To fill this gap is the main purpose of this document.

3.2. Current products

Operational LST products are currently available from a variety of instruments. Currently the three most commonly used LST products are delivered by the AATSR, MODIS, and SEVIRI instruments and are briefly described in the following section.

3.2.1. AATSR

The operational AATSR LST product (ATS_NR__2P) is based on the algorithm described in the original AATSR LST ATBD (Prata, 2002). The product provides gridded LSTs at a spatial resolution of 1 km for the entire globe. The target accuracy is 2.5 K for daytime retrievals and 1.0 K at night (Llewellyn-Jones et al., 2001). The product has been available since March 2004. Currently only data from the nadir view of the AATSR instrument is used for computing LST. More detail on the operational AATSR LST product can be found in the AATSR Product Handbook (European Space Agency, 2007). Efforts are currently underway to improve the algorithm by using auxiliary data with higher spatial resolution. This includes both a new dataset for global biome distribution and a new dataset for fractional vegetation cover.

3.2.2. MODIS

The operational MODIS LST product consists of Level-2 (MOD11_L2) and Level-3 (MOD11A1, MOD11A2, MOD11B1, MOD11C1, MOD11C2, MOD11C3) products. MOD11_L2 is available from the MODIS instruments on both the Terra and Aqua platforms. The product provides daytime and nighttime LST at a 1 km spatial resolution for the swath level. LST is derived using the generalized split-window LST algorithm (Wan and Dozier, 1996). The Level-3 products are gridded, projected, and generally producted by averaging the MOD11_L2 product in space and/or time. Details about the different products can be found in the MODIS LST Products Users' Guide (Wan, 2006).

3.2.3. SEVIRI

The operational LST product for SEVIRI is produced by the Land Surface Analysis Satellite Applications Facility (LAND-SAF). It is computed every 15 minutes at a spatial resolution of 3 km (at nadir) within the area covered by the MSG disk, i.e. primarily over Europe and Africa. The target accuracy is 2 K. The retrieval is based on the Generalized Split Window algorithm (Wan and Dozier, 1996). The SEVIRI LST product is described in detail in the corresponding Algorithm Theoretical Basis Document (Trigo et al., 2009).

3.3. State of the art of LST validation

Much of the international effort on validating data from spaceborne thermal infrared instruments has focused on SST. In comparison, the validation activities for LST are carried out by a fairly small but nonetheless very active research community. The following describes some of the more recent LST validation activities for AATSR, MODIS, and SEVIRI.

3.3.1. AATSR

Initial validation of the AATSR (Llewellyn-Jones et al., 2001) LST product was carried out over several Australian field sites and Lake Tahoe (Prata, 2003) and it was found that the accuracy of the algorithm is within the target specification. More recently, LST derived from AATSR was further validated over Valencia, Spain and Lake Tahoe, CA/NV, USA (Coll, Hook, et al., 2009), finding nearly zero average biases and standard deviations of 0.4 – 0.5 K for both daytime and nighttime data. They were further validated with a focus on long-term accuracy, finding that biases were negligible and that RMSE was ±0.5 K for full vegetation and water, and ±1.1 K for bare soil (Coll, Valor, et al., 2012). Both of these studies indicate the need for higher-resolution auxiliary to be used within the operational AATSR LST algorithm. The operational AATSR LST product was validated over Morocco where average biases of -1.00 K and -1.74 K were found for daytime and nighttime data, respectively (Noyes et al., 2007).

3.3.2. MODIS

Initial validation of MODIS (Justice et al., 1998) LST was carried out at multiple validation sites indicating that MODIS LSTs agree with in situ LSTs within ±1 K (Wan et al., 2002, 2004). Other studies using long-term nighttime ground measurements have found biases of 0.8 K for some sites and up to -3 K for other sites (Wang et al., 2008). Validation of the MODIS V5 level 2 LST product was further carried out in Valencia, Spain and Hainich, Germany, using both temperature- and radiance-based validation techniques (Coll, Wan, et al., 2009). The results indicate an average bias of -0.3 K and an RMSE of ±0.7 K for the temperature-based validation.

3.3.3. MSG-SEVIRI

The SEVIRI (Aminou, 2002) LST product provided by the facility on land surface analysis (LAND-SAF) has been validated against the MODIS LST product and against in situ observations made at the Evora station in Portugal, finding nighttime differences for the latter of 1-2 K, and higher discrepancies for daytime comparisons (Trigo et al., 2008). More recently, the uncertainty of LST from SEVIRI was assessed over the Gobabeb validation site in Namibia, where RMS difference between 1 K and 2 K were found (Freitas et al., 2010). The operational LST product derived from the SEVIRI instrument has further been validated over two sites in Eastern Spain, and accuracies of ±1.5 K were found in this study (Niclòs et al., 2011)

4. LST Validation Categories and common elements

The following section briefly describes an organizational structure of the various approaches to LST validation and gives some recommendations on elements that are common to all methodologies described here.

4.1. Validation Categories and accuracy classes

There is a wide variety of approaches to LST validation. The different types of approaches to LST validation are structured here within four different categories. Briefly, the four categories are described as follows

Category A: Comparison of satellite LST with in situ measurements
This is the traditional and most straightforward approach to validating LST. It involves a direct comparison of satellite-derived LST with collocated and simultaneously acquired LST from ground-based radiometers.

Category B: Radiance-based validation
This technique uses top-of-atmosphere (TOA) brightness temperatures (BTs) in conjunction with a radiative transfer model to simulate ground LST using data of surface emissivity and a atmospheric profiles of air temperature and water vapour content.

Category C: Inter-comparison with similar LST products
A wide variety of airborne and spaceborne instruments collects thermal infrared data and many provide operational LST products. An inter-comparison of LST products from different satellite instruments can be very valuable for determining LST.

Category D: Time series analysis
Analysing time series of satellite data over a temporally stable target site allows for the identification of potential calibration drift or other issues of the instrument that manifest themselves over time. Furthermore, problems associated with cloud contamination for example may be identified from artefacts evident in the time series. Care must be taken in distinguishing between instrument-related issues such as calibration drift and real geophysical changes of the target site or the atmosphere.

It should be noted that, while exhibiting certain similarities, these categories do not directly translate to the hierarchical validation classes as suggested previously for SST validation (Parkes et al., 2005) and by CEOS. While a Category A validation will generally be more complex and resource-demanding than a Category D validation, the four categories given here can be quite complementary, and a comprehensive LST validation will ideally entail certain elements from all four of them.

Each of the four validation categories is further subdivided into several classes based on the complexity of the methodology and the expected accuracy of the validation. For each category a set of quality criteria is established. A Class 1 validation must fulfil all these criteria. For each subsequent class, one of the listed criteria can be relaxed. Note that not

every category makes use of the full number of classes. Table 1 shows an overview of the categories and their classes.

Sections 5 through 8 provide details on each category and discuss the multiple accuracy classes. Each section first gives basic guidance on the recommended general methodology, discusses key criteria used in structuring the accuracy classes, briefly describes each accuracy class, and gives examples when available. In many ways, the classifications given in the these sections are somewhat idealized in that some of the classes are not represented by currently existing approaches. However, it is conceivable that some of the currently non-populated classes will be filled in future as more work is done on the topic and LST validation techniques become even more mature and powerful.

Table 4.1: Schematic overview of LST validation categories and complexity levels suggested in this document. Note that categories have a varying number of accuracy classes. Accuracy classes are only intended provide guidance within each validation category and should not be compared across validation categories. For details see Sections 5 through 8.

	Category			
	A In situ	**B** Radiance-based	**C** Inter-comparison	**D** Time series
Highest accuracy	A1	B1	C1	D1
	A2			
Accuracy Class	A3	B2	C2	D2
	A4			D3
	A5	B3	C3	
Lowest accuracy	A6			D4

4.2. Methodology common to all categories

Several techniques are common to all categories for LST validation. This section briefly describes some of the methodology that is shared by the various validation techniques. This includes cloud and snow masking, statistical techniques, and recommendations on visualization and reporting.

In general it is recommended to use LST products that maintains as much of the characteristics of the original dataset as possible, i.e. it is preferable to work with Level-2

products rather than Level-3 (unless investigating the accuracy of a specific Level-3 product is the main purpose of the validation).

4.2.1. Image characterisation

Validation techniques usually require knowledge about the contents of a scene and the atmospheric conditions at the time of overpass. In particular, any pixels in the scene that are affected by any kind of cloud cover need to be rigorously flagged and eliminated before any further analysis can take place. Both permanent and non-permanent snow cover should not however be removed – in the casethe snow surface temperature is the LST to be validated. In the case of AATSR for example the correct identification of a snow pixel will be part of the validation process

The approaches for all four categories require that the data has been subject to rigorous cloud-masking procedures. Validation of satellite LST requires particular care with respect to masking pixels affected by any type of cloud since accidental inclusion of even just slightly cloudy satellite data can negatively affect the representativeness of the validation results. Using robust statistics such as the median and the median absolute deviation can to some extent reduce the impact of including slightly cloudy satellite data; however it is recommended to take the utmost care in avoiding any kind of cloud contamination within the satellite dataset selected for validation purposes.

Automatic cloud screening techniques generally are able to provide a good first guess of cloud-affected pixels, but since they often are unreliable with respect to cirrus and other optically thin cloud layers, which can still cause a considerable bias in the LST estimates, they are not always suitable for the purposes of critical validation work. In many cases, the operational cloud mask products delivered with data from various satellite instruments (Závody et al., 2000; Derrien and Le Gléau, 2005; Frey et al., 2008), which are generally carried out using a series of threshold tests, are known to have certain issues (in particularly over land) and are thus not reliable enough to be used directly for LST validation work. Therefore, alternative cloud screening methods are often required. These can consist of semi-automated cloud screening processes possibly based on a subset of the cloud tests performed within the operational cloud screening procedures, additional cloud tests not considered in the operational cloud mask product, and most importantly a manual visual analysis of the satellite data to be used within the validation.

While not always possible, particularly for extremely large datasets, a visual examination of the satellite imagery to be used is highly recommended for any validation work in order to determine truly cloud-free conditions. Other instrumentation such as ceilometers can be used to identify clouds in the case of in situ validation. If not available then thresholds applied to the sky brightness temperatures offer an alternative possibility. In addition, plotting time series of presumably “cloud-free” LST data can often highlight any remaining cloud-affected data points that have not been flagged by the automatic cloud screening process.

In addition to characterisation of LST data with respect to cloud contamination, it can be very valuable to screen the image for atmospheric conditions that might have an adverse affect on validation accuracy, such as high aerosol loads.

4.2.2. Emissivity

Detailed knowledge of spectral surface emissivity is required for several LST validation techniques. Such information can be obtained from in situ measurements or from spectral libraries (Baldridge et al., 2009). While it might be sufficient to obtain only one emissivity value for highly homogeneous study sites, heterogeneous sites require information about the spectral emissivities of all endmembers occurring in the scene. In addition, some areas might exhibit phenology-dependent emissivity values or have other reasons for temporally changing emissivities, such as changing soil moisture. In such cases it is essential to obtain multiple emissivity values representing the entire range of variation.

4.2.3. Statistical techniques

Validation statistics are often computed from datasets that contain one or several outliers in the data. Classic statistical measures such as the mean, standard deviation, etc. are often severely affected by such outliers and others deviations from assumptions and can therefore result in misleading interpretations of the data at hand. Robust statistics on the other hand are to some extent resistant with respect to deviations from model assumptions (e.g. normality). In such cases, it is therefore advised to also use robust statistics such as the median in lieu of the mean and the median absolute deviation as a measure of statistical dispersion in lieu of the variance or standard deviation. It is recommended to use such robust statistics as additional reported values complementing standard summary statistics, and not necessarily as a replacement.

Statistical significance

It is recommended that the results from statistical tests are given together with the respective significance value, where appropriate. While it is recommended to always indicate the α-value used, doing so is particularly critical when values other than the conventional standard value of $\alpha = 0.05$ are used.

Uncertainty

It is recommended to provide estimates of uncertainty for any measurement. Extensive information on measuring, evaluating, and reporting uncertainty is available in the literature (Cox and Harris, 2006; Joint Committee for Guides in Metrology, 2008). It is further important to report the spatial and temporal structure of uncertainties.

4.2.4. Reporting of validation results

When reporting the results of a LST validation in quantitative form, it is recommended to give any bias or deviation calculated as

$$Bias = LST\_sat - LST\_ref$$

where LST_sat represents the satellite-derived LST and LST_ref represents the observation of a given reference LST, which is assumed to be closer to the true value. This convention ensures that a positive bias is indicative of an overestimate of the satellite LST, whereas a negative bias reflects a satellite LST that is too low with respect to the reference data set.

When validating LST it can further often be useful to report separately on the results obtained during night and during day. Since LST biases are usually higher during daytime than at night

due to insolation effects, giving a day/night average value will thus be less helpful as it is not representative of either night or day LST performance.

A common file format should be used in data delivery and reporting as much as possible. It is recommended to use the netCDF format developed by the University Corporation for Atmospheric Research (UCAR) (Rew and Davis, 1990; Rew et al., 1997) in conjunction with metadata that are compliant with the Climate and Forecasts (CF) Metadata Convention (Eaton et al., 2011). NetCDF is an open standard providing a self-describing and platform-independent binary data format that is particularly suited for the creation, access, and sharing of array-oriented scientific data. The format allows for efficient subsetting of the data.

4.2.5. Visualization

A good method for visualizing the deviations between two measured parameters, such as LST from satellite versus another source of LST, is a scatter plot of all the matchups. If a reference LST measurement is available, it is recommended to plot the reference LST on the x-axis of the plot and the satellite-derived LST on the y-axis.

When a reference dataset is available and when matchups are obtained continuously over a significant time period, it can also be very useful to plot the error between satellite LST and reference LST over time. This technique allows for a straightforward identification of major issues, such as instrument malfunction, calibration drift, etc.

Differences in LST derived from two data sources can further be shown by plotting histograms of the deviations. Such visualization can clearly indicate a potential bias between the two datasets and further illustrate the spread of the errors around the mean bias (and the associated symmetry).

More information on best practices in data visualization can be found in the literature (Gilby and Walton, 2006).

5. Category A: Comparison with in situ measurements

A comparison of satellite-derived LST against in situ observations of LST is the traditional and most straightforward way of LST validation. Accurate in situ observations of LST at a suitable site and taken under the right conditions arguably offer the highest-quality validation that can currently be achieved. Unfortunately, due to the high cost and complexity of operating a dedicated, highly reliable in situ LST validation site, the number of such sites is very low on a global scale.

5.1. General methodology for in situ validation

Several items need to be taken into account when collecting such validation data including the geographic properties of the site, the ideal measurement protocol, instrumentation, meteorological observations, etc.

5.1.1. Site Requirements

A Level 1 LST validation site is required to be either homogeneous over a large scale or heterogeneous with a well-documented spatial distribution of endmembers. The first type of LST validation sites should exhibit homogeneous surface properties on a scale from a few meters to many kilometres, depending on the spatial resolution of the spaceborne instrument that is to be validated. While data from high-resolution thermal IR sensors such as the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) can theoretically be validated using a site that is homogeneous over just a few hundred meters, most medium-resolution instruments such as AATSR, MODIS, AVHRR, or SLSTR require surface homogeneity on the order of several kilometres. Generally, it is recommended that any LST validation site is homogeneous over an area of at least 3 × 3 pixels.

For specifically testing LST algorithm performance over heterogeneous areas it is also helpful to use heterogeneous validation sites, such as the Evora validation site in Portugal. In this case brightness temperatures of all endmembers occurring at the sites must be measured independently by radiometers, emissivities of all endmembers must be known, and the spatial distribution of the endmembers must be adequately described - this needs to be described for each orbit since the geolocation of pixels overstriking the in situ site changes for each orbit of Low Erath Orbit satellites.

The LST-related homogeneity of the validation site can be assessed using high-resolution thermal infrared imagery such as ASTER. Homogeneous sites should exhibit a spatial LST variability (standard deviation) of no more than 0.5 K.

The location chosen for the in situ site should be representative of the immediate surroundings as well as of the land cover type that the LST site is supposed to describe.

5.1.2. Ideal measurement protocol

The original AATSR LST ATBD states the following about an ideal LST validation protocol (Prata, 2002): The ideal method for obtaining an independent validation data-set is to use two multi-channel radiometers with AATSR bandpasses from an aircraft and from the ground.

The measurements must be made as close as possible to the overpass time and ancillary data relating to atmospheric profiles and emissivity measurements must be collected. The target validation site must be uniform at scales of 100's m to several kilometres. The sky must be clear during the measurement period and the aerosol content of the atmosphere must be low. The measurement protocol would consist of the following:

The airborne and ground-based instruments must be calibrated to an accuracy of ±0.1 K and be traceable to a NIST blackbody.

Vertical profiles of temperature and moisture must be obtained within specified thresholds of the radiometer measurements and within a specified radius of the target area. The recommendation here is ±10 minutes and ±10 km respectively.

Directional sky radiance measurements – at ~53° from zenith (Kondratyev, 1969) - using a ground-based duplicate radiometer at the target should be made at the time of the overpass. Sufficient directional measurements should be made to permit an accurate determination of the flux density of spectral sky radiance.

Sun photometer measurements of at least the aerosol optical depth are required at the target area using the same temporal and spatial considerations as the profile data.

Directional spectral (8-14 μm) emissivity measurements must be made at the target with appropriate spatial sampling and within a few days of the overpass, provided no significant changes have occurred at the target surface during the period of measurements (e.g. no rainfall, fires etc.). The emissivity measurements should be made as close as possible to the local time of overpass.

For a vegetated target, measurements of the density and height of the vegetation might prove useful. The fractional vegetation cover must be estimated and emissivity measurements of the components must be made.

Daytime and nighttime validation data will be required.

An all-sky camera should be used at the target to provide an objective measure of cloudiness at the time of the overpass during the daytime. At night an upward viewing pyrgeometer should be used.

As these ideal observing conditions are not always possible, a validation against in situ data must often be complemented by other validation techniques described in the other categories below, such as model-based radiometric validation, inter-comparison with other LST datasets, or time series analysis.

5.1.3. Instrumentation

Radiometer

At homogeneous sites, a single, well-calibrated radiometer can be sufficient, although it can be helpful to distribute several radiometers throughout the study area in order to obtain estimates of spatial variability of LST. Heterogeneous sites will require at least one

radiometer for each of the site's endmembers. Ideally the radiometer should have an accuracy of ±0.1 K. The calibration of all radiometers must be traceable to a SI reference standard such as those provided by UK's National Physical Laboratory (NPL) or the National Institute of Standards and Technology (NIST). Furthermore, each radiometer should be independently calibrated before and after field campaigns, or at regular intervals throughout a continuous in situ collection period; the rationale here is to quantify potential drift in the instruments. The radiometers must be well documented and the documentation must be available to the community. A final recommendation is that radiometers be inter-compared with other radiometers.

Meteorological observations
Standard meteorological observations are essential for validation and should be ideally measured at the validation itself or, if that is not possible, at a nearby site situated at a distance no further than 50 km from the validation site. Observations should include air temperature, humidity, air pressure, wind speed, wind direction, and shortwave solar radiation. Precipitation data would be useful to assess potential impact on emissivity - this though is site specific and cannot be substituted by using data from a "nearby" site.

Radiosondes
Ideally, radiosondes measuring the atmospheric profiles of air temperature and water vapour will be launched concurrently to the satellite overpass at or in close proximity to the validation site. While atmospheric profiles of air temperature and water vapour content can also be retrieved from reanalysis datasets such as those provided by ECMWF (Uppala et al., 2005; Dee et al., 2011) or NCEP (Kalnay et al., 1996), the validation accuracy than can be obtained by using atmospheric profiles from dedicated radiosonde launches is significantly higher than what can be achieved by using reanalysis data, since spatial and temporal matching between reanalysis data and overpass measurements adds uncertainty.

5.1.4. Obtaining surface emissivity

Accurate information about the emissivity of the validation site must be available in order to convert measurements of brightness temperature into in situ LST. Whereas for spatially and temporally homogeneous validation sites, such as Lake Tahoe, it is sufficient to provide a single value for emissivity, heterogeneous sites require an estimation of multiple emissivity values for each of the site's endmembers. In addition, at some sites such as the one located in Evora, Portugal, the emissivity of certain scene endmembers changes throughout the year in response to the phenology of the vegetation. In such cases it is necessary to obtain representative seasonal emissivity estimates at least representing the major changes in emissivity. Values of surface emissivity can be obtained using the two-lid box method - a summary of which can also be found in Sobrino & Caselles (1993). If in situ emissivity measurements are not taken then emissivity from the ASTER spectral library (Baldridge et al., 2009) could insteasd be used for each endmember.

5.1.5. Determination of in situ LST

The true kinetic temperature is calculated from the brightness temperatures measured by the radiometer, the sky radiance reflected by the surface into the path of the radiometer, and the emissivity of the surface. This can be accomplished using the expression

$$B_c(T_c) = \varepsilon_c \cdot B_c(T_{sfc}) + (1 - \varepsilon_c) \cdot B_c(T_{sky})$$

where $B_c(T_c)$ is the measured radiance given by the Planck function for the effective brightness temperature T_c in the radiometer channel c, $B_c(T_{sfc})$ is the emitted surface radiance given by the Planck function for the surface temperature T_{sfc} in channel c, and $B_c(T_{sky})$ is the down-welling atmospheric radiance given by the Planck function for the effective brightness temperature T_{sky} of the atmosphere, and ε_c is the emissivity of the Earth's surface in channel c. The atmospheric radiance has a small but non-negligible impact on LST and therefore must be corrected for.

5.1.6. Spatial Sampling

The in situ measurement of LST needs to be measured such that it is representative for the entire area of a pixel of the satellite instrument, projected to the ground. If the site is very homogeneous it can be sufficient to measure LST using only one well-calibrated radiometer. If multiple radiometers are available it can be valuable to distribute them throughout the study area in order to obtain information about spatial variability in LST. Heterogeneous sites require the use of multiple radiometers, at least one for each endmember. This includes sampling of shadow areas beneath vegetation. For upscaling to the satellite pixel a geometric projection model can estimate the percentage of the pixel in shadow. As mentioned earlier the percentages of each endmember need to be identified for each overstrike pixel. The in situ LST to be compared with the satellite-derived LST is the weighted sum of the LSTs from the individual endmembers.

5.1.7. Temporal Sampling

Ideal in situ measurements of LST need to be continuous in time with a temporal sampling rate ranging from 1 second to 2 minutes. Observations at longer time intervals can still be helpful but generally result in either a significantly lower number of satellite vs. in situ matchups, or alternatively in a lower validation accuracy when the maximum allowed time difference between the two observations is increased. For longer matchups the in situ measurements should be temporally interpolated for comparison with the satellite-derived observation. Long gaps in the time series should be avoided as much as possible.

5.1.8. Uncertainty budget

An uncertainty budget needs to be provided for the spatial, temporal and geophysical uncertainty at the validation site. This includes instrument error, the natural variability of the ground LSTs, and the emissivity correction error.

5.1.9. Reporting

Any operator of an in situ validation site should report the original observations of the instruments as well as any derived measurements, such as LST. The data should be provided in a common format. Comprehensive information about the in situ validation site, including a detailed documentation of the instrumentation should be made available on a website. The information should be provided in similar form as for the CEOS Reference Sites (see http://calval.cr.usgs.gov/sites_catalog_ceos_sites.php), where a standardized questionnaire is

completed by the site maintainer and provided in PDF format at a public location on a website.

Besides information about the instrumentation (specific devices, calibration, interpretation), the site description should include information about land cover, phenology, elevation, climate, soil type, and other invariant parameters. If the type of measurements made at the site differs from the generally adopted definitions, this needs to be clearly stated. Information about the specific implementation of the site and information about ancillary data should also be provided.

In terms of other metadata it is also very important that the owner of an in situ validation site reports results from a radiometer calibration and an inter-comparison with other radiometers. The period for which measurements have been taken at the site needs to be given and significant gaps in the data need to be highlighted. It is particularly necessary to report if gaps in the time series have been filled using statistical interpolation techniques. For validation purposes it is recommended to always deliver time series without statistically filled gaps. The site descriptions needs to further include information about the calibration coefficients of the instrument and if they have changed at any point.

Ideally the in situ data from all stations should be provided to the research community at a common, well-documented, and easily accessible online location (such as an FTP server).

When visualizing the results of an LST validation against in situ data it is advisable to begin with simple scatter plots showing the LST for all available satellite vs. in situ matchups, where in situ LST is plotted on the x-axis and satellite LST on the y-axis.

Subsequently it is highly recommended to provide plots indicating the temporal behaviour of the differences between satellite LST and in situ LST. This can be accomplished through plotting the deviation (defined as satellite LST minus in situ LST) over time. Such plots can provide important information about possible seasonal variations in satellite LST quality or indicate other temporal patterns in the data.

When presenting plots for several in situ sites, the temperature range plotted on each axis should be kept constant, if possible. Otherwise the corresponding figure captions should clearly indicate the change in data range for the axes.

5.2. Accuracy Criteria in Category A

Here we define six classes of in situ data, describing a wide range of possibilities with respect to the data quality and expected validation accuracy as well as complexity and expense of data acquisition. In addition to proper instrumentation and calibration, a key element in defining the different levels was the availability of multi-year time series of radiometer data, since long and consistent time series are becoming increasingly valuable, not only for allowing a significantly increased number of satellite vs. in situ matchups for better statistical analysis, but also for better identification of instrumental issues such as drift.
The criteria selected for the quality classes in this category are:

Proper instrumentation and demonstrated calibration

Length of the data collection period (minimum 3 years)
No or only minimal gaps in the time series (maximum 5 days/month)
High-frequency temporal sampling (minimum 2 minutes)
Comprehensive documentation of the validation site
An existing uncertainty budget
(Seasonal) in situ emissivity provided
Measures actual BT of surface (not a proxy)

5.3. Class 1 in situ data (A1)

A class 1 in situ validation site completely fulfils all the criteria listed above. At the time of writing no such sites exist, primarily due to the lack of a comprehensive and documeneted uncertainty budget at the currently existing Class 2 sites.

5.4. Class 2 in situ data (A2)

Class 2 in situ data are established long-term in situ validation sites, completely equipped with one or more highly accurate and calibrated radiometers, which provide continuous observations with no or only small data gaps. The sites are either completely homogeneous over an area of ideally 3 x 3 pixels or alternatively heterogeneous with well-documented and measured heterogeneity. In addition to radiometers, these sites measure the standard meteorological parameters. All instrumentation and procedures must be well documented and the site description must be available to the LST research community. The sites in this class only lack one of the criteria listed above, in most cases this is a comprehensive uncertainty budget.

Examples of established sites that collect Class 2 in situ data are

Gobabeb LST validation site, Namibia (Olesen and Göttsche, 2009; Göttsche et al., 2011)

Evora LST validation site, Portugal (Olesen and Göttsche, 2009; Göttsche et al., 2011)

Lake Tahoe LST validation site, California/Nevada, USA (Hook et al., 2003, 2007)

5.5. Class 3 in situ data (A3)

In situ validation sites that fail to meet two of the criteria listed above are considered Class 3 sites. In addition to the lack of an uncertainty budget, such sites typically have long (multi-year) time series but are not continuous in the sense that measurements are only taken during certain times of the year.

An example of an established sites collecting Class 3 in situ data is the Valencia LST validation site, Spain (Coll et al., 2005). At this site, LST measurements have been carried out with calibrated radiometers for several years, however the sampling was limited to the summer months of each year.

5.6. Class 4 in situ data (A4)

In situ validation sites that fail to meet three of the criteria listed are considered Class 4 sites. This includes stations that have potential as Class 2 (or even Class 1) stations, but that currently only have relatively short time series which may have substantial gaps due to instrument failure or cloud conditions. Furthermore this class contains station networks established for reasons other than LST validation as well as most validation campaigns that provide high-quality and reasonably well-documented data but only for a very short time period.
Examples of Class 2 in situ sites/datasets are:

Dahra, Senegal (Olesen and Göttsche, 2009; Göttsche et al., 2011)
RMZ/Heimat, Namibia (Olesen and Göttsche, 2009; Göttsche et al., 2011)
Sites of the Atmospheric Radiation Measurement (ARM) Climate Research Facility (Stokes and Schwartz, 1994)
Most short-term LST validation campaigns

It is likely that sites such as Dahra and RMZ/Heimat will move to one of the higher classes when they have been operational for several years and provide longer time series with fewer data gaps.

5.7. Class 5 in situ data (A5)

Class 5 in situ sites violate four of the listed criteria. An example of Class 5 in situ data includes measurements made by the United States Climate Reference Network (USCRN) (NOAA/NESDIS, 2007). This network provides continuous surface temperature data at 114 stations located within the continental United States and is planned to be operated for many decades in order to provide consistent datasets for climate research. However, while the stations provide data on infrared ground surface temperature, the calibration level of the instruments and thus the data quality is currently unknown. In addition, the data are currently only recorded at hourly intervals. Further research needs to be undertaken to determine the feasibility of using USCRN data for LST validation. If the outcome of such an analysis is positive, it is likely that USCRN data can be classified in one of the higher classes in future.

5.8. Class 6 in situ data (A6)

Class 6 in situ observations violate at least five of the listed criteria. This lowest level of complexity and validation quality is generally provided by in situ measurements of LST proxies, such as air temperature. It has been shown that the air temperature and ground surface temperature are almost the same during certain hours of the night and that air temperature can thus be used as a surrogate dataset to validate LST as a first approximation if the satellite overpass occurs temporally close to those times (Prata, 2003). While this technique obviously only allows for rough comparisons, air temperature is ubiquitously observed at a large number of meteorological stations worldwide, so this class potentially includes a multitude of stations that can be used for providing general guidance on LST quality at a global scale. Such an approach has for example been used in Greenland (Hall et al., 2008).

6. Category B: Radiance-based validation

Radiance-based validation offers an alternative to validation with in situ LST measurements as it does not require measurements of LST on the ground (Wan and Li, 2008). This technique simulates top-of-atmosphere brightness temperatures with a radiative transfer model using data of surface skin temperature, surface emissivity and nearly concurrent atmospheric profiles of air temperature , water vapour and if available aerosols. Perturbations are applied to the input skin temperature until the simulated BTs match the satellite retrieved BTs – the LST retrieval error is then the difference between satellite-retrieved LST and inverted skin temperature. This method has been applied for validation of the MODIS LST product (Wan and Li, 2008), validation of AATSR LST in Spain (Coll, Valor, et al., 2012) and for validating at-sensor radiance from ASTER and MODIS at Lake Tahoe (Hook et al., 2007). While radiance-based validation of LST cannot completely replace accurate in situ measurements of LST, it can provide a viable alternative for long-term, semi-operational LST product evaluation at the global scale (Coll, Valor, et al., 2012).

6.1. General methodology for radiance-based validation

While there are various ways to perform a radiance-based validation of LST and methodological details vary tremendously between the different approaches, there are a number of general steps are commonly followed.

6.1.1. Obtaining atmospheric profile

A radiance-based validation requires knowledge about the vertical structure of the atmosphere, in particular about the vertical profiles of air temperature and water vapour. This information can be obtained from either radiosonde launches in the vicinity of the validation site or from the output of atmospheric models such as the reanalyses provided by ECMWF (Uppala et al., 2005; Dee et al., 2011) or NCEP (Kalnay et al., 1996). A detailed radiance-based validation should generally use radiosonde profiles obtained at or near the study site but model-based atmospheric profiles can be valuable for some studies. It is possible to test the accuracy of the atmospheric profiles through the difference Δ(T11–T12) between the satellite-retrieved BT differences (T11–T12) and the simulated BT differences (T11–T12) (Wan and Li, 2008; Coll, Caselles, et al., 2012). Δ(T11–T12) should be close to zero when the atmospheric temperature and water vapour profiles used in simulations represent the real atmospheric conditions and effect of the surface emissivities for the satellite observations.

6.1.2. Obtaining surface emissivities

As for temperature-based validation with in situ data, a radiance-based validation requires knowledge of the surface emissivity at the validation site. For homogeneous sites it can be sufficient to just have one emissivity value, whereas heterogeneous sites require knowledge of spectral emissivities for each individual scene endmember. It should be noted that some sites, such as Evora in Portugal, exhibit emissivities that change throughout the year in conjunction with the phenology of the vegetation. In such cases it can be necessary to obtain multiple emissivity values throughout the year to reflect the seasonal change.

6.1.3. Applying radiative transfer code

In this step, the top-of-atmosphere BTs are simulated using a radiative transfer model, such as MODTRAN (Berk et al., 1999) or RTTOV (Saunders et al., 1999), in conjunction with surface skin temperatures, spectral surface emissivities and near-concurrent atmospheric profiles of temperature and water vapour. Radiative transfer codes include line-by-line models and band transmission models. Line-by-line models calculate the contribution of each spectral line for all molecules in the atmospheric layer and provide the highest accuracy albeit generally associated with high computational expense. Radiative transfer code performing band transmission calculations is computationally less demanding but its results area also more approximate.

6.1.4. Reporting

When reporting the results from a radiance-based validation it is vital to provide exact information about the radiative transfer model that was used as well as how the atmospheric profile data was obtained. Furthermore it is critical to provide data on emissivities used in the validation procedure.

6.2. Accuracy Criteria in Category B

The accuracy classes for this category are established based upon the expected accuracy of both the radiative transfer code and the atmospheric profiles. The two key criteria used for classifying the types of radiance-based validation of LST were

1. Atmospheric profile of temperature and water vapour content obtained from radiosonde launches at the validation site
2. Radiative transfer model based on line-by-line radiative transfer code rather than band transmission

Three classes have been defined, ranging from highest to lowest expected accuracy of the validation.

6.3. Class 1 radiance-based validation (B1)

A Class 1 radiance-based validation fulfils both of the criteria listed above, i.e. it uses line-by-line radiative transfer code in conjunction with atmospheric profiles of air temperature and water vapour observed directly at the validation site during the time of the satellite overpass. The combination of a line-by-line radiative transfer code together with a radiosonde-based atmospheric profile provides the highest possible level of accuracy for this type of LST validation.

6.4. Class 2 radiance-based validation (B2)

A Class 2 radiance-based validation is performed very similar to a Class 1 radiance-based validation; however one of the two classification criteria is not met. Therefore, a Class 2 radiance-based validation uses either a band-transmission radiative transfer model in

conjunction with radiosonde-based atmospheric profiles observed directly at the validation site during the time of the satellite overpass or a line-by-line radiative transfer model in conjunction with atmospheric profiles obtained from a less accurate source such as from areanalysis dataset.

6.5. Class 3 radiance-based validation (B3)

A Class 3 radiance-based validation relaxes both of the classification criteria necessary for a Class 1 validation: It uses a broad-band radiative transfer model together with atmospheric profiles obtained from an atmospheric model. As such, this validation class is expected to provide the lowest level of validation accuracy; however it is still a very valuable validation method when no ground LST measurements are available.

7. Category C: Inter-comparison with other datasets

Intercomparison of an LST product with another source of spatially distributed LST data, i.e. LST information derived from another satellite instrument or from an airborne instrument, is a very powerful tool for validation. While it generally cannot provide the same (pseudo)-absolute validation as a comparison against in situ observations, an intercomparison is often a good supplement and can give important quality information with respect to spatial patterns in LST deviations.

The 'intercomparison' between products allows evaluation of their relative consistency. This is very important for users when exposed to several products. It is also mandatory when combining several products into the 'best available product'. However this step on its own is not sufficient for comprehensive validation since the various products could be 'consistent' but nonetheless biased with respect to reference data.

7.1. General methodology for LST inter-comparison

There are various ways to perform a validation of LST through an inter-comparison with other data sources of LST. Therefore the methodological details vary tremendously between the different approaches and the data used. However certain general steps are commonly followed in most approaches and they are described in the following.

7.1.1. Data quality control and selection

Prior to any LST inter-comparison all data should be quality checked and pixels not meeting the highest quality control levels should be eliminated (unless the aim of the validation is to specifically evaluate lower quality data).

Any LST inter-comparison requires scene characterisation and rigorous cloud screening (see section 4.2.1 for details). A particular concern in the context of inter-comparisons between multiple satellite-based LST datasets is the view angle of the instrument. When matchups are being selected for a particular location, different instrument view angles can alter the results of the validation. Depending on the products used and the final goal of the validation, view angle dependency should be quantified in the reporting of the intercomparison findings.

7.1.2. Spatial regridding

In order to inter-compare data acquired at significantly different spatial resolutions it is recommended to resample both data sets to a common grid. A variety of approaches exist for this purpose, however one of the most straightforward techniques is to average the values of all the pixels whose centre coordinate falls within each common grid cell.

7.1.3. Temporal matching

In order to match two LST datasets temporally, time interpolation should be carried out. For example, if one of the datasets has a substantially different sampling rate than the other, as often occurs when comparing data from Low Earth Orbit instruments with data acquired in a

geostationary orbit. The value from the geostationary instrument used in the matchup should be an interpolatation between successive measurements to correspond to the overpass time of the respective Low Earth Orbit satellite.

7.1.4. Reporting

Depending on the validation strategy and the chosen methodology, the results can be reported in a wide variety of ways. One of the most straightforward ways to report the results from a multi-sensor intercomparison of LST is to show a map of LST differences between the two datasets. Such a map can not only highlight overall biases between the two datasets but most importantly it also helps identify spatial patterns in the deviations, including artificial artefacts caused by algorithm issues such as poor-quality auxiliary data. If the validation data set is collected over a sufficiently long time period, it is further useful to plot the temporal behaviour of the deviations between the two LST datasets, in order to show, among others, possible seasonal signals in the data. When reporting the results statistically, it is recommended to differentiate between day- and night-time inter-comparisons.

7.2. Accuracy Criteria in Category C

The key criteria selected for a Category C validation are:

1. Matchup times must be within a specified threshold – a recommendation being ±10 minutes.
2. The reference dataset must be operationally available

7.3. Class 1 inter-comparison (C1)

A Class 1 inter-comparison is considered the most accurate type of inter-comparison. It fulfils the criteria listed above. Such an inter-comparison can for example be accomplished by using data from two instruments onboard of the same satellite platform or alternatively from similar instruments flying successively along a similar orbit (such as the A-Train group of satellites).

7.4. Class 2 inter-comparison (C2)

A Class 2 inter-comparison relaxes one of the criteria established above. As such, this accuracy class encompasses LST inter-comparisons of datasets that deviate in their respective matchup times by more than the recommended threshold.

A special case of Class 2 inter-comparisons are airborne LST observations. Aircraft equipped with radiometers are a powerful tool for validating both TOA radiances and LST derived from spaceborne instruments, as they provide much higher spatial detail than satellite data. They further allow detailed observations of the atmospheric conditions. The primary advantage of the airborne platform is that it allows coverage over a much larger area than could normally be achieved with field measurements and provides significantly more spatial detail than satellite observations can deliver. However, it is a requirement that the instrument is well calibrated. Airborne TIR instruments are generally flown in dedicated validation

campaigns and as such are specifically set up to match the satellite instrument to be validated in terms of overpass time and viewing angle. However, due to the episodic nature of such campaigns and the resulting lack of operational availability, data acquired in airborne LST campaigns does not fulfil one of the criteria listed above and thus is classified as a Class 2 inter-comparison.

7.5. Class 3 inter-comparison (C3)

The third accuracy class within the inter-comparison category encompasses the inter-comparison of satellite-derived LST data with spatially distributed LST proxy data, such as reanalysis data as provided by the ECMWF (Uppala et al., 2005; Dee et al., 2011) or NCEP (Kalnay et al., 1996).

8. Category D: Time series analysis

Time series analysis is an efficient technique for screening a remote sensing dataset for potential problems that occur over time, i.e. calibration drift, offsets between different instruments of a series, or identifying unrealistic outliers. As such, time series analysis can be helpful as a screening tool for LST validation. The quality classes for this category are organized with increasing area over which the time series is computed. It is argued that time series provided at small spatial scales will be able to provide more specific information with respect to temporal validation than global time series. While the latter can be interesting in their own right, their data tends to include a variety of instrumental issues and geophysical signals which are difficult to separate, and they are therefore less useful for validation purposes.

A precision assessment reflects the repeatability of the products. This step is very important when analyzing long time series or comparing different regions. It could be derived from the comparison to reference in situ data if enough data are available and if they are associated with small uncertainties. Because of the lack of reference data and the sometime significant uncertainties associated with them, precision is assessed by quantifying the variability of the products over surfaces that are known to be homogeneous and stable. If such surfaces are difficult to find because of the heterogenous and dynamic nature of vegetation, it may be quantified by evaluating the 'smoothness' of the temporal profile or geostatistical metrics such as variograms that may show some significant variability (nugget) for pixels separated by very short distances. The evaluation should also consider the frequency of valid data with additional details on data QA (such as the application of gap filling techniques). Histograms of the data are also expected to show whether the distributions seem reasonable.

8.1. General methodology for time series analysis

Time series analysis of LST data can be carried out in a variety of ways and the methodology depends highly on the specific type of data used and the primary goals of the validation. Therefore only very general methodological recommendations on using time series analysis for purposes of LST validation can be given here.

8.1.1. Data acquisition

The value of time series for LST validation as well as for studying geophysical phenomena generally increases with length, completeness, and increasing temporal resolution. On the other hand, long and complete time series consisting of daily or even orbit data tend to consume a large amount of storage space (in particular at continental and global scales) and their analysis can be very resource-intensive. When acquiring data from different sources it is recommended to store all data in a common data format in order to simplify automatic extraction of LST values from the satellite data.

8.1.2. Data extraction

In order to generate a time series, data for a specific region must be extracted from each image. This process generally involves automatically processing the archive of LST data,

averaging over a number of pixels in each image (unless the time series is supposed to reflect only one pixel), and storing a time series of the mean values.

8.1.3. Time series analysis

Once the data has been extracted from the archive of images and meaningful averages have been computed for a given area, the actual time series analysis can begin. This generally involves visual analysis of temporal plots as well as computing statistical measures, such as moving averages. In addition, trend analysis and statistical modelling techniques can be applied at this stage. It can be particularly helpful to study long-term LST trends over known stable targets in order to estimate possible calibration drift or other sensor issues.

8.1.4. Reporting

The reporting of time-series analysis results for LST validation purposes is primarily accomplished through providing temporal plots of the data and description of their interpretation. If trend analysis is being performed on the time series, it can further be useful to provide its result in quantitative form. When reporting trends, it is further recommended to always report uncertainty estimates for the trend value.

8.2. Accuracy Criteria in Category D

In contrast to the other three validation categories, no specific key criteria were defined for category D. While criteria such as the length of time series, the number and size of gaps in the time series, and the temporal sampling could be used to define classes, such requirements are highly dependent on the goal of the time series analysis and vary widely.

Instead, accuracy classes for this category were established simply based on the geographical scale over which the time series are computed. This follows the idea that localized time series will be able to be interpreted more easily than global time series, as the latter usually combine a whole number of artificial and natural issues and thus make the identification of individual instrument problems or similar validation issues very challenging.

8.3. Class 1 time series analysis (D1)

A Class 1 time series analysis is performed at the single pixel level (or an average over a small number of neighbouring pixels). Plotting and studying reasonably long time series of such a small area can be very valuable for identifying algorithm and instrument issues, such as increased noise or calibration drift.

8.4. Class 2 time series analysis (D2)

A Class 2 time series analysis is performed at for a geographically invariant scene. For most instruments on low earth orbit platforms this means areas at a scale on the order of a few hundred kilometres.

8.5. Class 3 time series analysis (D3)

A Class 3 time series analysis is performed at the regional scale. This could be on the scale of a country, a continent, or a certain land cover type. While not as powerful as Class 1 and 2, this type of time series analysis can still provide valuable information about problems in the data. For biome based retrieval algorithms this may uncover biome-specific retrieval issues.

8.6. Class 4 time series analysis (D4)

A Class 4 time series analysis is computed at the global scale. While this type of analysis is not as useful for validation purposes as the other three classes, it can still provide interesting information. Furthermore, after establishing their accuracy, global time series are extremely relevant for global change research.

9. Conclusions

Validation of satellite-based LST products can be carried out in a variety of ways. This document establishes four categories that encompass the vast majority of LST validation work that is currently being carried out, namely validation against in situ data, radiance-based validation, intercomparison between different LST datasets, and finally time series analysis.
While a comparison against in situ data is generally the most accurate and reliable LST validation technique, it is also one of the most complex and resource-demanding. It requires extraordinary care both in acquiring accurate LST from well-calibrated radiometers over a suitable and representative location and in ensuring that the significant differences in spatial scale between the point-level in situ observations and the kilometre-scale satellite LST pixels can be overcome appropriately.

Radiance-based validation of LST has improved significantly in achievable accuracy and has gained popularity in recent years. It can be a valuable alternative to in situ-based validation in areas or land cover types where adequate coverage with in situ data is not available.

Inter-comparison of various satellite-based LST products is a helpful validation technique as it doesn't require the use of in situ data. An intercomparison of products however can only give information about the consistency between the different datasets, not an assessment of the true accuracy of individual products.

Finally, time series can be a useful validation technique for identifying a wider variety of instrument-related and other issues that cause significant artefacts in the data.

While each of these four validation categories can give important information about various aspects of LST product quality, a comprehensive validation of an LST product will ideally incorporate elements of all four categories.Furthermore, for LST validation to be meaningful. Measurements values, techniques applied and analysis findings should be fully documented and freely available to both the CalVal and user community.

10. Updates to QA4EO

No updates to QA4EO are considered necessary at this time.

11. References

Aminou, D.M.., 2002. MSG's SEVIRI Instrument. ESA bulletin 111, 15-17.

Baldridge, A.M., Hook, S.J., Grove, C.I., Rivera, G., 2009. The ASTER spectral library version 2.0. Remote Sensing of Environment 113, 711-715.

Berk, A., Anderson, G.P., Bernstein, L.S., Acharya, P.K., Dothe, H., Matthew, M.W., Adler-Golden, S.M., Chetwynd, J.H., Richtsmeier, S.C., Pukall, B., Allred, C.L., Jeong, L.S., Hoke, M.L., 1999. MODTRAN4 Radiative Transfer Modeling for Atmospheric Correction, in: SPIE Proceedings, Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research III. pp. 348-353.

Coll, C., Caselles, V., Galve, J., Valor, E., Niclos, R., Sanchez, J., Rivas, R., 2005. Ground measurements for the validation of land surface temperatures derived from AATSR and MODIS data. Remote Sensing of Environment 97, 288-300.

Coll, C., Caselles, V., Valor, E., Niclòs, R., 2012. Comparison between different sources of atmospheric profiles for land surface temperature retrieval from single channel thermal infrared data. Remote Sensing of Environment 117, 199-210.

Coll, C., Hook, S.J., Galve, J.M., 2009. Land Surface Temperature From the Advanced Along-Track Scanning Radiometer: Validation Over Inland Waters and Vegetated Surfaces. IEEE Transactions on Geoscience and Remote Sensing 47, 350-360.

Coll, C., Valor, E., Galve, J.M., Mira, M., Bisquert, M., García-Santos, V., Caselles, E., Caselles, V., 2012. Long-term accuracy assessment of land surface temperatures derived from the Advanced Along-Track Scanning Radiometer. Remote Sensing of Environment 116, 211-225.

Coll, C., Wan, Z., Galve, J.M., 2009. Temperature-based and radiance-based validations of the V5 MODIS land surface temperature product. Journal of Geophysical Research 114, 1-15.

Cox, M.G., Harris, P.M., 2006. Software Support for Metrology - Best Practice Guide No. 6: Uncertainty Evaluation (No. DEM-ES-011). National Physics Laboratory.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, a. J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137, 553-597.

Derrien, M., Le Gléau, H., 2005. MSG/SEVIRI cloud mask and type from SAFNWC. International Journal of Remote Sensing 26, 4707-4732.

Eaton, B., Gregory, J., Centre, H., Office, U.K.M., Drach, B., Taylor, K., Hankin, S., Caron, J., Signell, R., Bentley, P., Rappa, G., Hock, H., Pamment, A., Jockes, M., 2011. NetCDF Climate and Forecast (CF) Metadata Conventions.

European Space Agency, 2007. AATSR Product Handbook.

Freitas, S.C., Trigo, I.F., Bioucas-dias, J.M., Göttsche, F.-m, 2010. Quantifying the Uncertainty of Land Surface Temperature Retrievals From SEVIRI / Meteosat 48, 523-534.

Frey, R.A., Ackerman, S.A., Liu, Y., Strabala, K.I., Zhang, H., Key, J.R., Wang, X., 2008. Cloud Detection with MODIS. Part I: Improvements in the MODIS Cloud Mask for Collection 5. Journal of Atmospheric and Oceanic Technology 25, 1057-1072.

Gilby, J., Walton, J., 2006. Software Support for Metrology - Good Practice Guide No. 13: Data Visualisation (No. DEM-ES-009). National Physics Laboratory.

Göttsche, F., Olesen, F., Bork-Unkelbach, A., 2011. Validation of Operational Land Surface Temperature Products with Three Years of Continuous In-Situ Measurements, in: Proeceedings of the 2011 EUMETSAT Meteorological Satellite Conference. Oslo, Norway, 5-9 September 2011.

Hall, D., Box, J., Casey, K., Hook, S., Shuman, C., Steffen, K., 2008. Comparison of satellite-derived and in-situ observations of ice and snow surface temperatures over Greenland. Remote Sensing of Environment 112, 3739-3749.

Hook, S.J., Prata, F.J., Alley, R.E., Abtahi, A., Richards, R.C., Schladow, S.G., Pálmarsson, S.Ó., 2003. Retrieval of Lake Bulk and Skin Temperatures Using Along-Track Scanning Radiometer (ATSR-2) Data: A Case Study Using Lake Tahoe, California. Journal of Atmospheric and Oceanic Technology 20, 534.

Hook, S.J., Vaughan, R.G., Tonooka, H., Schladow, S.G., 2007. Absolute Radiometric In-Flight Validation of Mid Infrared and Thermal Infrared Data From ASTER and MODIS on the Terra Spacecraft Using the Lake Tahoe, CA/NV, USA, Automated Validation Site. IEEE Transactions on Geoscience and Remote Sensing 45, 1798-1807.

Joint Committee for Guides in Metrology, 2008. Evaluation of measurement data - Guide to the expression of uncertainty in measurement (No. JCGM 100:2008).

Justice, C.O., Vermote, E., Townshend, J.R.G., Defries, R., Roy, D.P., Hall, D.K., Salomonson, V.V., Privette, J.L., Riggs, G., Strahler, A., Lucht, W., Myneni, R.B., Knyazikhin, Y., Running, S.W., Nemani, R.R., Wan, Z., Huete, A.R., Leeuwen, W.V., Wolfe, R.E., Giglio, L., Muller, J.-P., Lewis, P., Barnsley, M.J., 1998. The Moderate Resolution Imaging Spectroradiometer (MODIS): Land Remote Sensing for Global Change Research. IEEE Transactions on geoscience and remote sensing 36, 1228-1249.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., 1996. The NCEP/NCAR 40-year reanalysis project. Bulletin of the American Meteorological Society 77, 437-471.

Kondratyev, K. Y., 1969. Radiation in the Atmosphere. New York Academic Press.

Llewellyn-Jones, D., Edwards, M., Mutlow, C., Birks, A., Barton, I., Tait, H., 2001. AATSR: global-change and surface-temperature measurements from ENVISAT. ESA bulletin 105, 10–21.

NOAA/NESDIS, 2007. United States Climate Reference Network (USCRN) Functional Requirements Document (No. NOAA-CRN/OSD-2003-0009R1UD0). U.S. Department of Commerce, National Oceanic and Atmospheric Administration (NOAA), National Environmental Satellite, Data, and Information Service (NESDIS).

Niclòs, R., Galve, J.M., Valiente, J. a., Estrela, M.J., Coll, C., 2011. Accuracy assessment of land surface temperature retrievals from MSG2-SEVIRI data. Remote Sensing of Environment 115, 2126-2140.

Noyes, E.J., Soria, G., Sobrino, J.A., Remedios, J.J., Llewellyn-Jones, D.T., Corlett, G.K., 2007. AATSR land surface temperature product algorithm verification over a WATERMED site. Advances in Space Research 39, 171-178.

Olesen, F.-S., Göttsche, F.-M., 2009. Validation of Land Surface Temperatures Obtained from METEOSAT-MVIRI and SEVIRI with in-situ measurements, in: 2009 EUMETSAT Meteorological Satellite Conference, Bath, UK, 21-25 September 2009. EUMETSAT.

Parkes, I.M., Steven, M.D., Mutlow, C.T., Donlon, C.J., Foot, J., Prata, F., Grant, I., Nightingale, T., Edwards, M.C., 2005. AATSR Validation Implementation Plan Part 1 - AATSR Validation Principles and Definitions (No. PO-PL-GAD-AT-005).

Prata, F., 2002. Land surface temperature measurement from space: AATSR algorithm theoretical basis document. CSIRO Atmospheric Research, Aspendale, Australia.

Prata, F., 2003. Land surface temperature measurement from space: Validation of the AATSR Land Surface Temperature Product. CSIRO Atmospheric Research, Aspendale, Australia.

QA4EO, 2010. A Quality Assurance Framework for Earth Observation: Principles.

Rew, R., Davis, G., 1990. NetCDF: An Interface for Scientific Data Access. IEEE Computer Graphics & Applications 10, 76-82.

Rew, R.K., Davis, G.P., Emmerson, S., Davies, H., 1997. NetCDF User's Guide for C, An Interface for Data Access, Version 3.

Saunders, R., Matricardi, M., Brunel, P., 1999. An improved fast radiative transfer model for assimilation of satellite radiance observations. Quarterly Journal of the Royal Meteorological Society 125, 1407-1425.

Sobrino, J. S. & Caselles, V., 1993, A field method for measuring the thermal infrared emissivity. ISPRS Journal of Photogrammetry and Remote Sensing 48, 24–31.

Stokes, G.M., Schwartz, S.E., 1994. The Atmospheric Radiation Measurement (ARM) Program - Programmatic Background and Design of the Cloud and Radiation Test Bed. Bulletin of the American Meteorological Society 75, 1201-1221.

Trigo, I.F., Freitas, S.C., Bioucas-Dias, J., Barroso, C., Monteiro, I.T., Viterbo, P., 2009. Algorithm Theoretical Basis Document for Land Surface Temperature (LST) (No. SAF/LAND/IM/ATBD_LST/1.0). Land Surface Analysis Satellite Application Facility (LAND-SAF).

Trigo, I.F., Monteiro, I.T., Olesen, F., Kabsch, E., 2008. An assessment of remotely sensed land surface temperature. Journal of Geophysical Research 113, 1-12.

Uppala, S.M., Kållberg, P.W., Simmons, A.J., Andrae, U., Da Costa Bechtold, V., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, a. C.M., Berg, L.V.D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., Isaksen, L., Janssen, P.A.E.M., Jenne, R., Mcnally, A.P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N.A., Saunders, R.W., Simon, P., Sterl, A., Trenberth, K.E., Untch, A., Vasiljevic, D., Viterbo, P., Woollen, J., 2005. The ERA-40 re-analysis. Quarterly Journal of the Royal Meteorological Society 131, 2961-3012.

Wan, Z., 2006. MODIS Land Surface Temperature Products Users ' Guide. ICESS, University of California, Santa Barbara.

Wan, Z., Dozier, J., 1996. A generalized split-window algorithm for retrieving land-surface temperature from space. IEEE Transactions on geoscience and remote sensing 34, 892–905.

Wan, Z., Li, Z.-L., 2008. Radiance-based validation of the V5 MODIS land-surface temperature product. International Journal of Remote Sensing 29, 5373-5395.

Wan, Z., Zhang, Y., Zhang, Q., Li, Z.-L., 2004. Quality assessment and validation of the MODIS global land surface temperature. International Journal of Remote Sensing 25, 261-274.

Wan, Z., Zhang, Y., Zhang, Q., Li, Z.-liang, 2002. Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. Remote Sensing of Environment 83, 163-180.

Wang, W., Liang, S., Meyers, T., 2008. Validating MODIS land surface temperature products using long-term nighttime ground measurements. Remote Sensing of Environment 112, 623-635.

Závody, A.M., Mutlow, C.T., Llewellyn-Jones, D.T., 2000. Cloud Clearing over the Ocean in the Processing of Data from the Along-Track Scanning Radiometer (ATSR). Journal of Atmospheric and Oceanic Technology 17, 595-615.