remote sensing

MDPI

Article

Stability Assessment of the (A)ATSR Sea Surface Temperature Climate Dataset from the European Space Agency Climate Change Initiative

David I. Berry [1,*], Gary K. Corlett [2,3], Owen Embury [4,5] and Christopher J. Merchant [4,5]

[1] National Oceanography Centre, University of Southampton Waterfront Campus, European Way, Southampton SO14 3ZH, UK
[2] Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK; gkc1@le.ac.uk
[3] National Centre for Earth Observation, University of Leicester, University Road, Leicester LE1 7RH, UK
[4] Department of Meteorology, University of Reading, Reading RG6 6AL, UK; o.embury@reading.ac.uk (O.E.); c.j.merchant@reading.ac.uk (C.J.M.)
[5] National Centre for Earth Observation, University of Reading, Reading RG6 6AL, UK
* Correspondence: dyb@noc.ac.uk; Tel.: +44-(0)23-8059-7740

Received: 24 November 2017; Accepted: 15 January 2018; Published: 18 January 2018

Abstract: Sea surface temperature is a key component of the climate record, with multiple independent records giving confidence in observed changes. As part of the European Space Agencies (ESA) Climate Change Initiative (CCI) the satellite archives have been reprocessed with the aim of creating a new dataset that is independent of the in situ observations, and stable with no artificial drift (<0.1 K decade^{-1} globally) or step changes. We present a method to assess the satellite sea surface temperature (SST) record for step changes using the Penalized Maximal t Test (PMT) applied to aggregate time series. We demonstrated the application of the method using data from version EXP1.8 of the ESA SST CCI dataset averaged on a 7 km grid and in situ observations from moored buoys, drifting buoys and Argo floats. The CCI dataset was shown to be stable after ~1994, with minimal divergence (~0.01 K decade^{-1}) between the CCI data and in situ observations. Two steps were identified due to the failure of a gyroscope on the ERS-2 satellite, and subsequent correction mechanisms applied. These had minimal impact on the stability due to having equal magnitudes but opposite signs. The statistical power and false alarm rate of the method were assessed.

Keywords: Along Track Scanning Radiometer (ATSR); sea surface temperature; stability; homogeneity; drifting buoys; Argo; Global Tropical Moored buoy Array (GTMBA); Penalized Maximal t Test

1. Introduction

Observations of the sea surface temperature (SST) forms one of the key components of the climate record (e.g., [1,2]), with in situ observations extending back over 150 years (e.g., [3–7]) and satellite-based estimates over 20 years (e.g., [8]). Whilst some confidence is gained from the agreement between the different datasets few of them are truly independent from each other. For example, all of the in situ based datasets (e.g., [3–7]) are derived from the International Comprehensive Ocean–Atmosphere Data Set (ICOADS; e.g., [9,10]). Blended datasets (e.g., [11]) tend to contain in situ observations from ICOADS, and one or more satellite based sources. Satellite-only estimates (e.g., [12]) have tended to be calibrated and validated using the same in situ sources as present in ICOADS.

In recognition of the importance of independent estimates the (A)ATSR Reprocessing for Climate (ARC) project [13] aimed to produce a satellite sea surface temperature record based on measurements from the (Advanced) Along Track Scanning Radiometer ((A)ATSR) series of sensors. Within ARC, independence from the in situ record was achieved through the use of radiative transfer modeling to determine the retrieval coefficients used to calculate the skin sea surface temperature from the observed brightness temperatures [13,14]. These were then converted to bulk temperatures and a standard time of 1030 am/pm local time using the models of Fairall et al. [15] and Kantha and Clayson [16].

More recently, in recognition of the importance of sustained observations, high quality independent climate data records and the contribution that Earth Observation (EO) data can make, the European Space Agency (ESA) launched the Climate Change Initiative (CCI) [17]. The primary aim of the CCI is to realize the full potential of EO data from archives held by both the ESA and other satellite agencies. As part of this initiative, the ESA SST CCI project [18] was set up with the goal of reprocessing the satellite based radiometric SST record from the Advance Very High Resolution Radiometer (AVHRR) and ATSR series of sensors to produce an independent satellite-based SST dataset with a target stability of 10 mK year^{-1} (0.1 K decade^{-1}) [18] and building on the previous effort of the ARC project.

Stability of observation is a key aspect of quality for any record intended for use in climate applications. Measurement of climatic change involves tracking of small differences over years and decades, with the signal being often similar in size to the error characteristics of the raw observations. By careful treatment of the observations the impact of errors can be minimized, allowing any long-term changes to be quantified. However, even with careful treatment, residual artifacts may exist, either in the mean values or the error characteristics. These artifacts may be step changes or long-term drift in the mean error or in the higher moments (such as variance). Detection of these residual artifacts, or inhomogeneity, relies on comparison with independent sources, which may themselves be inhomogeneous. Interpreting observed instability between sources then becomes an expert judgement taking account of the timescales, timing, and nature of the changes, and the availability of supplemental metadata.

Various algorithms exist for assessing time series for change points and homogeneity (e.g., see [19]) but these tend to have been developed using data from land stations in fixed locations. Within this paper we develop a method to assess the homogeneity of the ESA SST CCI dataset, using the Penalized Maximal t Test (PMT) [20] with satellite in situ differences aggregated over many different observations and platforms, and an ensemble approach to quantify the uncertainty in the timing and size of any change points. We then apply the method using three different in situ reference sources, using observations from the: Global Tropical Moored Buoy Array (GTMBA) (e.g., [21]); the global drifting buoy array [22]; and the array of Argo profiling floats [23]. Additionally we fit an auto-regressive trend model to identify any long-term drift in the satellite data relative to in situ data. Section 2 of this paper describes the different data sources used. Section 3 gives a brief summary of change-point analysis and describes how we have applied the PMT to our data. Results are given in Section 4, and a summary and conclusion are given in Section 5.

2. Data

Data from Experimental Version 1.8 of the ESA SST CCI project [24] have been used to develop and test the method presented in this paper. These data have been extracted from a multi-sensor match-up database (MMD) [25] containing collocated swath or level 2 pre-processed (L2P) data from the (A)ATSR series of satellites and in situ observations from a variety of platforms. The in situ observations and satellite retrievals are briefly described in this section.

2.1. In Situ Data

In situ data from the MMD was originally been extracted from the Met Office Hadley Centre Integrated Ocean Database (HadIOD) [26]. These were, in turn, extracted from Release 2.5 of the International Comprehensive Ocean and Atmosphere Data Set (ICOADS2.5; [9]) and the Hadley Centre EN4 dataset [27]. ICOADS2.5 and EN4 contain data from common sources, such as the

World Ocean Database (WOD) [28] Where an observation exists in both ICOADS2.5 and EN4, the observation from EN4 was retained in preference [26]. Whilst the match-up database contains observations from many different platforms, only those match-ups containing drifting buoy, GTMBA and near surface (<5 m) Argo data were used in this study. Temperature observations from ships, sub-surface measurements and extra-tropical moorings were been used due to either quality (ships [29], extra-tropical moorings [30]) or representativeness issues (sub-surface).

The drifting buoy temperature observations from HadIOD were extracted from ICOADS2.5 [9]. The data within ICOADS2.5 came from a number of overlapping sources, with observations duplicated between the different sources. These duplicates were removed by ICOADS2.5 processing [9]. Prior to ingestion into HadIOD, the observations also underwent quality control following Rayner et al. [4], with implausible values and gross errors in location (time and space) and temperature flagged. Additionally, the observations had a platform level quality check applied prior to ingestion into HadIOD, with the quality of the observations made by individual buoys tracked over time through comparison with a satellite-based analysis [31]. The GTMBA data underwent the same processing as the drifting buoy data and came from either ICOADS2.5 or EN4. As noted above, where a duplicate was found, the EN4 data were retained in preference.

The surface Argo data within HadIOD were extracted from EN4 [27]. As with ICOADS2.5, EN4 contains data from a number of sources, including from the Argo Global Data Assembly Centre (GDAC) [23], the World Ocean Database (WOD) [28], and the Global Temperature Salinity Profile Programme. As part of the EN4 processing, duplicate temperature profiles were identified and removed, with the Argo GDAC data kept in preference to the other sources. Prior to ingestion into EN4 the profiles undergo additional quality control checks, including *inter alia*: gray list checks, parameter range checks, profile checks (spikes, steps, etc.), bathymetry, and depth checks. Full details on the duplicate elimination and quality control can be found in Good et al. [27] and Ingleby and Huddleston [32].

The primary characteristics of the drifting buoy, GTMBA and near surface Argo data are summarized in Table 1. The uncertainty values listed are those as reported by Atkinson et al. [26]. Also listed are validation statistics for the initial version of the ESA SST CCI ATSR global analysis as reported by Merchant et al. [18]. The drifting buoy observations are the most numerous and closest to the surface but have higher uncertainties compared to the other sources. Alternative estimates of the uncertainty in the drifter data ranges from ~0.2 K [33] to ~0.5 K [34]. The manufacturer has little impact on the quality of drifting buoy data [30]. The GTMBA data have smaller uncertainties compared to the drifting buoy data but are limited to the tropics and at a slightly greater depth. The near surface Argo measurements have the smallest uncertainty due to measurement errors but are made at the greatest depth compared to the other sources, and with far fewer observations available. Overall, based on the validation results listed in Table 1, the expected error variance is broadly equal across the three different in situ platforms.

Table 1. Uncertainty characteristics and depth of the in situ observations following Atkinson et al. [26]. Also listed are the robust standard deviation of the differences between the ESA SST CCI product and listed platforms, the median difference and number of match-ups from phase 1 of the ESA SST CCI project [18].

	Atkinson et al. [26]			Merchant et al. [18]		
Platform	**Uncertainty due to Noise/Random Errors (K)**	**Uncertainty due to Correlated Errors (K)**	**Nominal Depth**	**RSD (K)**	**Median Difference (K)**	**Number of Match-Ups**
GTMBA	0.020	0.000	1 m	0.28	+0.09	25,492
Drifting buoys	0.260	0.290	0.2 m	0.22	+0.05	2,392,462
Argo	0.002	0.000	3–5 m	0.26	+0.04	8867

There have been few technological changes to the GTMBA, drifting buoy and Argo observations that will have significantly impacted the quality of the data over the study period. However, the density and distribution of the observing systems has changed significantly. There was a marked increase in the number of drifting buoy observations in the late 1990s and early 2000s. The number of available match-ups are shown in Figure 1. There has also been a spatial evolution in the location of the data. For example, prior to ~2000, drifting buoy observations of the SST tended to be poleward of the tropics, with very few observations within 10° of the equator. Figure 2 shows the percentage of drifting buoy match-ups available from the MMD in a given 5° latitude band calculated annually. The lack of tropical observations prior to 2000 was clearly seen.

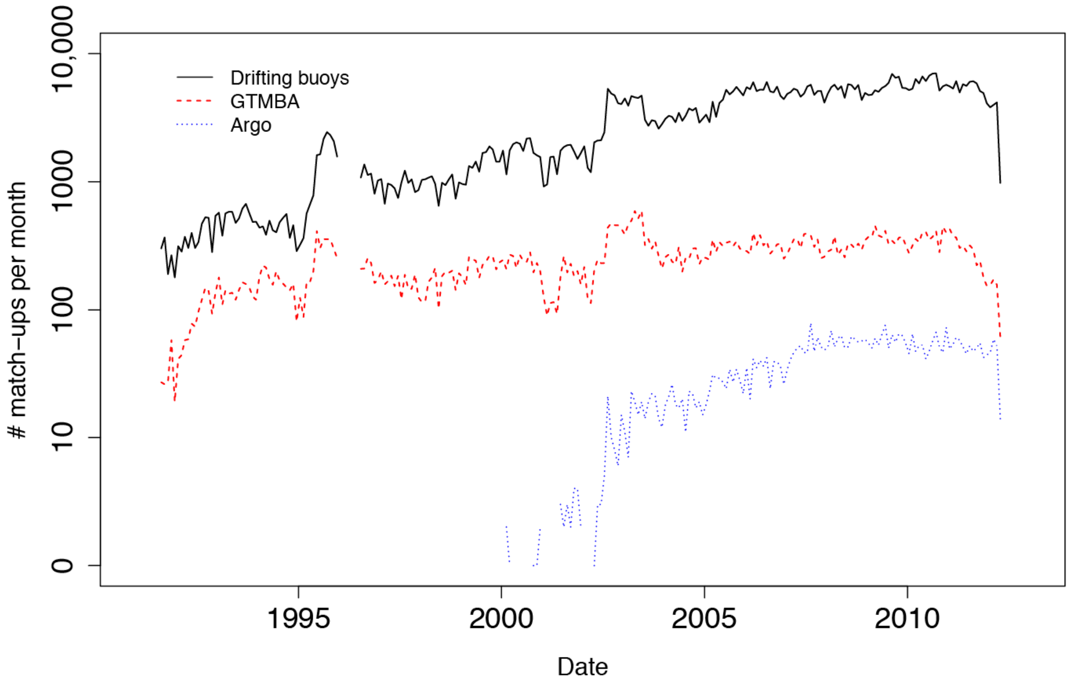

Figure 1. Number of match-ups per month between drifting buoy (black/solid line), GTMBA (red/dashed line) and Argo (blue/dotted line) observations and the ESA SST CCI data meeting the match-up criteria.

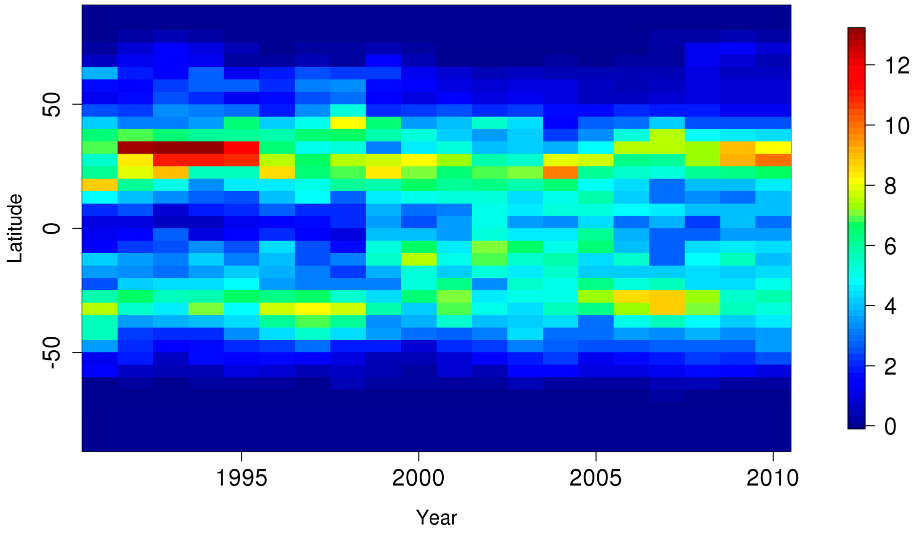

Figure 2. Percentage of drifting buoy match-ups available per 5° latitude band calculated annually.

2.2. Satellite Data

The match-up database from the ESA SST CCI project used within this study contained level 2 pre-processed (L2P) data from experimental version 1.8 of the project for the ATSR series of sensors. At the time of analysis, AVHRR data from v1.8 were unavailable. These will be available in subsequent versions. The algorithms used to generate the L2P data are summarized by Merchant et al. [18] and fully described in Merchant [24]. A general description of the (A)ATSR characteristics is given by Merchant et al. [35]. A brief summary is given in this section.

The L2P data contain estimates of the skin temperature (e.g., [36]) from a single swath with a resolution of 1 km, a swath width of 500 km, a 3-day repeat cycle and a local equator crossing time of either 1030 am/pm (ATSR and ATSR2) or 1000 am/pm (AATSR). For the ATSR sensors, the skin temperature estimates are based on the linear combination of 2 or 3 brightness temperature channels [24] following the ARC project [13,14]. The number of channels depends on whether the retrievals are for the daytime or nighttime and the satellite (See Table 2). The coefficients used in the linear combination of channels have been derived using radiative transfer modeling and atmospheric profiles from the ERA 40 reanalysis model. Cloud screening, trace gas atmospheric profiles and the impact of aerosols are accounted for as part of the retrieval process. Full details are given in Merchant [24]. Consistency between ATSR sensors was achieved using the overlap between the sensors and by referencing the brightness temperatures of all channels and sensors to be consistent with the 3.7 and 11 µm channels of the AATSR (e.g., [35]).

Table 2. Primary channels, satellite views and local equator crossing time for the different ATSR sensors and daytime/night time retrievals.

	ATSR1		**ATSR2**		**AATSR**	
	Day	**Night**	**Day**	**Night**	**Day**	**Night**
View	Dual	Dual	Dual	Dual	Dual	Dual
Channels (µm)	11, 12	11, 12	11, 12	3.7, 11, 12	11, 12	3.7, 11, 12
LECT	1030	2230	1030	2230	1000	2200
Period	August 1991–December 1995		August 1995–June 2003		July 2002–April 2012	

In addition to the skin temperature, the L2P files contain estimates of the sub-skin temperature and water temperature at different depths in the near surface ocean (0.2 m, 1 m and 5 m) adjusted to 1030 am/pm local time. The sub-skin temperatures were estimated using the Fairall et al. [15] model to adjust the skin temperatures to sub-skin temperatures. These were then adjusted to bulk temperatures using the model of Kantha and Clayson [16]. Both models were forced using the output from the ERA Interim Reanalysis model [37] and were used to adjust the temperatures to be representative of 1030 am/pm local time [24].

A number of events may impact on the quality of the (A)ATSR based retrievals. These are summarized in Figure 2 of Merchant et al. [35]. Shortly before the launch of the first ATSR sensor on board the ERS-1 satellite Mount Pinatubo erupted in June 1991, injecting a large amount of aerosol into the stratosphere. Whilst the (A)ATSR sensors have been designed to be robust to aerosol through the use of dual view sensors (e.g., [35]) there may be some residual effect present in the SST retrievals in the years shortly after 1991. In May 1992 the 3.7 micron channel failed on board the ATSR1 sensor. This can be seen in Table 2, with only two channel retrievals available for the ATSR1 sensor. Whilst data are available for 1996 from ATSR1, these are of lower quality and were excluded. Data from ATSR2 is available from August 1995 through to June 2003, with the exception of a six-month gap in the first half of 1996. During 2001 a gyroscope failed on the ERS-2 satellite, with data quality impacted between January 2001 and June/July 2001 when a zero gyro mode was implemented to improve the quality. Data from the AATSR sensor are available between July 2002 and April 2012. There are no known events thought to impact on the quality of the data in this period.

3. Method

The stability of the ESA SST CCI data was assessed using two different methods. The first tested for change points in the satellite data using a test based on the PMT algorithm [20]. The second tested for a residual trend in the time series of differences between the satellite data and in situ reference series. This section describes those tests. All tests were implemented using the R programming language [38].

3.1. Penalized Maximal t Test and Application to ESA SST CCI Data

3.1.1. Homogeneity Testing and the Penalized Maximal T Test

The majority of homogeneity tests for climate data are based on testing the relative homogeneity of a time series from one station with another nearby station that has a similar climatic signal. A discussion of the background and benchmarking of different homogeneity tests is discussed in Venema et al. [19]. Early tests, such as the Standard Normal Homogeneity Test (SNHT) [39], were based on classical statistical tests, comparing a target series to a reference series believed to be homogenous. More recently, tests have been developed to detect change points using inhomogeneous reference series and pairwise comparisons (e.g., [40]). In both techniques, using a reference series and those based on pairwise comparisons, test time series are created by either differencing or calculating the ratio between the target and reference time series or between pairs of time series across a local network of stations. A test statistic is then calculated for all possible change points and the most likely change point identified. If this exceeds some critical value it is then flagged as a change point. These tests are often applied recursively to find multiple change points.

Over the oceans, with the exception of the GTMBA [21], we have few high quality fixed stations spanning multiple years to use. The extra-tropical moorings tend to be in coastal regions where a small change in location can have a significant impact on the SST time series. Additionally, the quality of these observations is low compared to the GTMBA (e.g., see [26,30,41]). This makes application of tests based on multiple pairwise comparisons difficult with marine data. Instead we have opted to use the PMT [20] algorithm and aggregated time series of satellite and in situ data (see Section 3.1.2).

In the PMT algorithm [20], the test is applied to a test time series, X_t, with the null hypothesis given by:

$$H_0 : \{X_t\} \sim N\left(\mu, \sigma^2\right) \tag{1}$$

and the alternative hypothesis given by:

$$H_a : \begin{cases} \{X_t\} \sim N(\mu_1, \sigma^2), & t = 1, \ldots, k \\ \{X_t\} \sim N(\mu_2, \sigma^2), & t = k+1, \ldots, N \end{cases} \tag{2}$$

where X_t is the time series being tested for change points; $\{X_t\} \sim N(\mu, \sigma)$ indicates that the population $\{X_t\}$ follow a normal distribution with a mean μ and standard deviation σ. The test time series, X_t, is usually a time series of differences between the base time series, and a reference time series although ratios can be used. For the null hypothesis, the data from the test series (X_t) are assumed to come from a normal distribution with a mean difference (or ratio) of μ and standard deviation of σ. For the alternative hypothesis, the mean of the time series before the change point at time k is given by μ_1 and by μ_2 after time k.

Within the PMT algorithm, the null hypothesis was tested by calculating the test statistic for the two-sample t test with unknown but equal variance at every possible change point. The most likely position of a change point is then identified by the location where this test statistic, multiplied by a penalization factor, is a maximum. The penalization factor accounts for the increased false alarm rate observed when the sample sizes are unequal [20]. The test statistic for the two sample t test is given by:

$$T(k) = \frac{1}{\hat{\sigma}_k} \left[\frac{k(N-k)}{N} \right]^{0.5} \left(\overline{X_1} - \overline{X_2} \right) \tag{3}$$

where:

$$\overline{X_1} = \frac{1}{k}\sum_{1\leq t\leq k}(X_t) \tag{4}$$

$$\overline{X_2} = \frac{1}{N-k}\sum_{(k+1)\leq t\leq N}(X_t) \tag{5}$$

$$\hat{\sigma_k} = \frac{1}{N-2}\left[\sum_{1\leq t\leq k}\left(X_t - \overline{X_1}\right)^2 + \sum_{(k+1)\leq t\leq N}\left(X_t - \overline{X_2}\right)^2\right] \tag{6}$$

and N is the number of time steps and k the position of the potential break point. The test statistic for the PMT is then given by [20]:

$$PT_{max} = \max_{1\leq k\leq (N-1)}[P(k)T(k)] \tag{7}$$

where $P(k)$ is the penalization factor. When PT_{max} exceeds a critical value, the null hypothesis is rejected and a break point at position k identified. Both the penalization factor and critical values for PT_{max} have been determined empirically, and full details are given in Wang et al. [20]. The PMT algorithm has been updated [42] to pre-whiten the test time series prior to the calculation of step size, in order to account for any auto-correlation in the data. Within this study we used the original PMT algorithm [20] implemented in the RHTest software package [43]. This should not impact our conclusions as we were only interested in the detection of steps and not the subsequent adjustment.

3.1.2. Application of the PMT to ESA SST CCI Data

Three different change point analyses were using the PMT algorithm and in situ data. The analyses were identical other than the source platforms for the in situ observations. The platforms used were:

1. Observations from the GTMBA;
2. Observations from drifting buoys;
3. Near surface observations from the Argo profiling floats.

An additional assessment has been performed using synthetic data to test the sensitivity of the method to detect a step of 0.05 K at different points in time.

Pre-Processing and Selection of Match-Ups

The following criteria have been used to select match-ups:

1. Satellite quality level equal to 4 (acceptable) or 5 (highest) (see [24] for description of quality levels).
2. HadIOD quality control [31] passed for in situ observations.
3. Separation distance between in situ observation and satellite retrieval <100 km.
4. Maximum time separation between in situ observation and satellite retrieval $\leq$ 1 h.
5. Satellite in situ difference <5 standard deviations of all match-ups for a given platform.

Within the MMD match-ups are defined as the L2P pixel from the various satellite sensors containing the location of the reference in situ observation and where the overpass time is within ±2 h of the reference observation [25]. The data extracted from the MMD contained one L2P match-up per reference observation for each (A)ATSR sensor plus the corresponding L2P data from the surrounding pixels on a 7 × 7 grid, with the grid centered on the match-up pixel. Pixels with quality level <4 are set to missing (criteria 1). As described in Section 2, the resolution of the pixels was ~1 km, and the L2P data was adjusted to 0.2 m depth and 1030 am/pm local time. In addition to the L2P data, the data extracted from the MMD included the in situ reference observation for each match-up plus observations nearby in time (±36 h) from the platform making the reference observation. Those failing the HadIOD QC were discarded (criteria 2). For Argo data, only the observation corresponding to the match-up was available due to the ~10 day repeat cycle for Argo floats.

Due to cloud screening and quality control, a significant proportion of the pixels in a 7 × 7 scene, including the central pixel, contained missing values. To mitigate the impact of missing data and increase the number of match-ups, the L2P data was averaged over the 7 × 7 scene to give an areal average, analogous to Level 3 data. To account for the time difference between the in situ observations and the time of the satellite data, adjusted to 1030 am/pm local time, the in situ observations, and location, either side of the satellite data and within 1 h (criteria 4) was interpolated to 1030 am/pm local time. The average location of the pixels containing valid data from the 7 × 7 scene and interpolated buoy locations were used to test criteria 3. This test was required due to a small number of erroneous drifting buoy locations in the HadIOD dataset. For Argo, a single observation was available and only those match-ups within 1 h of 1030 am/pm local time were used. It should be noted that the nearest pixel to the reference observation could have been selected in the case of missing data for the central pixel. However, during testing the choice of satellite data to use, nearest pixel or mean of 7 × 7 scene, made little difference to the results.

As a final step, following extraction and selection of the match-ups, the mean and standard deviation of the satellite in situ 0.2 m SST across all match-ups were calculated for the GTMBA, drifting buoy data and profiling floats separately. Any difference exceeding 5 standard deviations of the differences for the respective platform was then excluded from further processing (criteria 5).

Ensemble and Aggregation

As noted above, we have used aggregated time series in the application of the PMT. From Equation (3), it can be seen that by minimizing the standard deviation of the test series ($\hat{\sigma_k}$) we increased the sensitivity of the test. This term will include contributions from instrumental noise and errors due to both the satellite data and in situ observations. Climatic variations were minimized through use of collocated data. By averaging multiple match-ups per month the contribution of instrumental noise and errors to the monthly mean values was minimized. Based on propagation of error we would expect the standard deviation of the monthly mean values to be reduced. For independent normally distributed errors we expect the reduction in $\hat{\sigma_k}$ to be proportional to $\frac{1}{\sqrt{n}}$ where n is the number of match-ups per month.

As an example, we sub-sampled the selected match-ups for the drifting buoys and AATSR sensor for sample sizes between 1 and 100 match-ups per month and calculated the monthly mean difference. For each sample size 20 realizations were generated and the standard deviation of the monthly mean differences calculated. These were then averaged across all realizations of a given sample size and plotted as a function of sample size in Figure 3 (dots). The theoretical relationship was also shown, with a strong agreement between the observed and expected reduction in standard deviation with increasing sample size clearly seen. The rate of reduction was slightly less than expected for independent data, indicating that the samples were not strictly independent. This was to be expected as a given buoy may contribute to multiple match-ups in a given month. However, this impact was small.

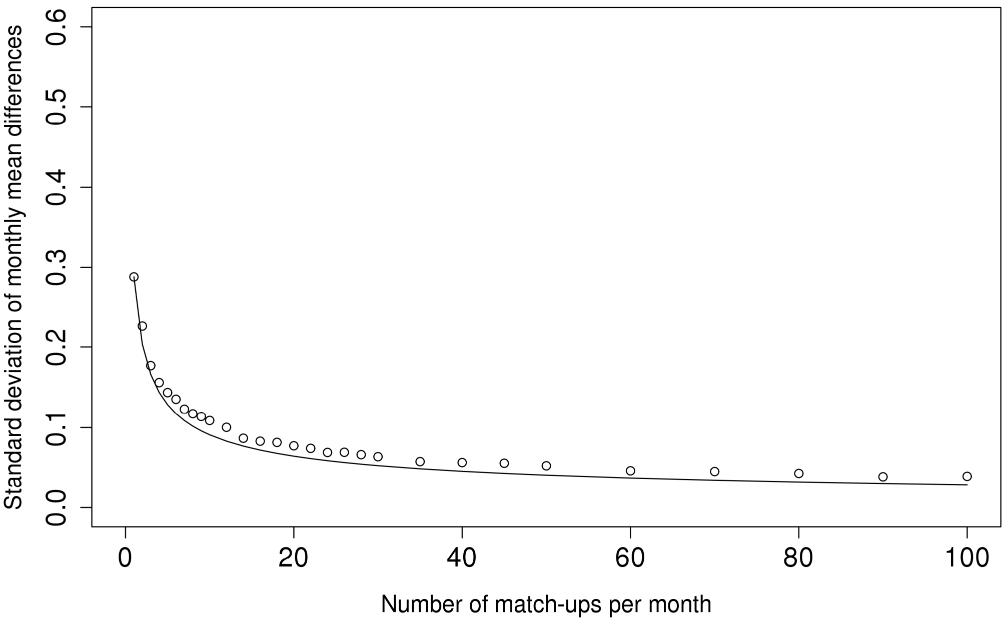

Figure 3. Standard deviation of monthly mean AATSR drifting buoy SST observations as a function of number of match-ups per month (dots). The theoretical relationship, $\frac{\sigma}{\sqrt{n}}$, is also shown (solid line).

To minimize the impact of errors, and increase the sensitivity of the PMT, aggregate time series were generated by sub-sampling the available match-ups and averaging to give monthly mean values. For each analysis, an ensemble of 1000 realizations was generated, with a fixed number of match-ups used to minimize the variance of the monthly standard deviation in time. The number of match-ups used depended on the data availability (see Section 4). The use of an ensemble allowed the impact of the sub-sampling on the analysis and the uncertainty in the date and size of any detected change points to be quantified.

3.1.3. False Alarm Rate

Within our application of the PMT we used a nominal false alarm rate (FAR) of 1% (see Section 4.1). Based on a FAR of 1%, for every 100 applications of the PMT to independent data, we expected one application to incorrectly identify a step change. Due to our use of ensembles that contained a proportion of the same match-ups between different ensemble members, our applications of the PMT were not independent. As a result, we expected a FAR that differs from the nominal FAR specified as part of the PMT. To quantify the impact of our method on the nominal FAR we calculated the observed FAR for each analysis presented by replacing the match-ups with synthetic data and reapplying the tests. Within the synthetic dataset, the in situ observations were replaced by zeros and the satellite data with normally distributed noise with a mean of zero and standard deviation corresponding to the observed standard deviation of the satellite in situ differences calculated each month and for each sensor/platform configuration.

3.2. Stability Assessment

The long-term stability of the satellite SST estimates relative to the in situ data was quantified by fitting a linear trend model to the satellite in situ differences. The (A)ATSR SST retrievals are known to contain small seasonally varying biases relative to the in situ data. Within a time series of the differences between the (A)ATSR and in situ temperatures this will appear as auto correlated errors. To take these into account when assessing the stability we used a lag 1 autoregressive (AR1) model, with the model given by (e.g., [44]):

$$X_i = \beta_0 + \beta_1 t_i + \epsilon_i \tag{8}$$

$$\epsilon_i = \rho\epsilon_{i-1} + e_i \tag{9}$$

where X_i is the mean difference for time step i; β_0 the intercept for the trend model; β_1 the slope of the trend model; t_i the time variable for time step i; ϵ_i the auto regressive error at time step i; ρ the autocorrelation parameter at lag 1; and e_i the independent error or noise term for time step i. If the autocorrelation is not taken into account, the degrees of freedom used to calculate the confidence intervals will be overestimated and the uncertainty range will be too narrow.

4. Results

4.1. Step Change Analyses

Table 3 summarizes the configuration of the three different analyses and time period covered. The sample size lists the number of match-ups selected per ensemble member. The number selected was determined to balance the length of the time series available with the sensitivity of the analysis. The sample sizes for GTMBA and Argo were smaller than for the drifting buoy analysis due to the more limited availability of the data.

Table 3. Summary of the analyses performed using the PMT. The FAR column indicates the specified false alarm rate used in the tests.

Reference Platform (Period)	Sample Size	Ensemble Size	Reference SST	FAR
GTMBA (1991 onwards)	25	1000	Linear interpolation to 1030 AM/PM local time	1%
Drifting buoy (~1996 onwards)	100	1000	Linear interpolation to 1030 AM/PM local time	1%
Argo (~2004 onwards)	16	1000	Nearest observation to 1030 AM/PM local time (max separation 1 h)	1%

4.1.1. GTMBA

The first analysis repeated the homogeneity assessment made as part of the ARC project [35]. Within ARC, the PMT was applied to a single aggregate time series for a subset of the moorings from the GTMBA in the Pacific. Based on this analysis, the (A)ATSR record was found to be stable in the tropics after the effects of the Pinatubo eruption had dissipated (by around 1994/1995). Similar results were found within this analysis, but by using all available GTMBA moorings from the match-up database. Table 4 summarizes the results of the analysis, listing the number of change points detected per ensemble as a fraction of the ensemble size. A large number of ensemble members (~20%) had at least one change point detected compared to a nominal FAR of 1% and actual FAR of <1% based on the synthetic data. The majority of the ensemble members with a change point have a single change point detected (17%). A small fraction (1.6%) had two change points detected.

Table 4. Number of change points detected vs. percentage (%) of the 1000 ensemble members for the GTMBA analysis. Also listed is the number of change points detected when the observations are replaced with noise.

Number of Change Points	% of Ensemble Members with Detected Change Points	Observed False Alarm Rate (FAR) (%)
0	80.9	99.4
1	17.1	0.2
2	1.6	0.4
3	0.4	0.0

Figure 4 shows the results of the change point analysis as time series of the mean change point models averaged over those ensemble members with the same number of change points (panel a, black lines). Also shown are the monthly mean satellite in situ differences for the individual ensemble members (red lines). Figure 4b shows a histogram of the change point dates. The majority of change points detected occurred prior to 2000 and Figure 5 shows a scatter plot of the step size versus date for those ensemble members with a single change point occurring prior to the start of 2000. Three different clusters were seen, with a step size of ~0.08 K around the end of 1992 for the earlier cluster. Subsequent clusters decreased in magnitude of step size over time. Overall, the evidence suggests a warming in the satellite data during this period, with a median step of 0.054 K (95% confidence interval 0.022 to 0.089 K) occurring around February 1994 (October 1992 to December 1995). The confidence intervals were 2.5% and 97.5% quantiles estimated using the R function "quantile" and based on Hyndman and Fan [45]. The cluster of points in December 1995 coincided with the end of the ATSR1 data used, with the PMT algorithm and RHTest preferentially using the metadata to set the date of the change point.

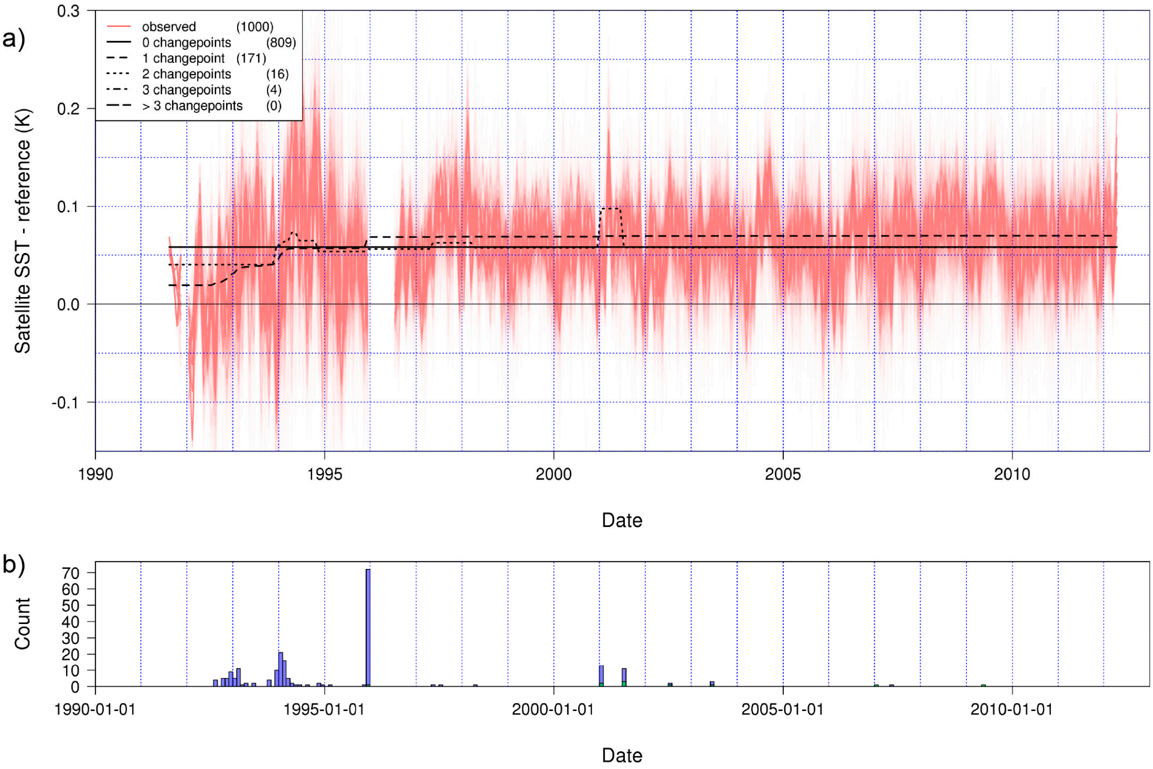

Figure 4. Results of the ensemble application of PMT using the GTMBA. (**a**) The monthly mean differences (satellite in situ) for the individual ensemble members are shown in red. Also shown are the means of the model fitted by PMT grouped by number of change points detected. (**b**) Histogram of the data of the change points detected across all ensemble members.

The evidence for the change points after 2000 shown in Figure 4 is weaker, with only 16 out of 1000 ensemble members showing the change points during January and July 2001. These change points are coincident with the gyroscope failure on the ERS-2 satellite and implementation of the zero-gyro mode in July 2001 to correct for the effects of this on the SST (e.g., Merchant et al., 2012). The size of these changes are +0.08 K and −0.07 K respectively. Without the supporting metadata neither of these change points would be significant based on the GTMBA data. However, similar change points have been detected in the drifting buoy analysis, but with opposite signs, suggesting that the impact of the gyro failure varies between the tropics and higher latitudes.

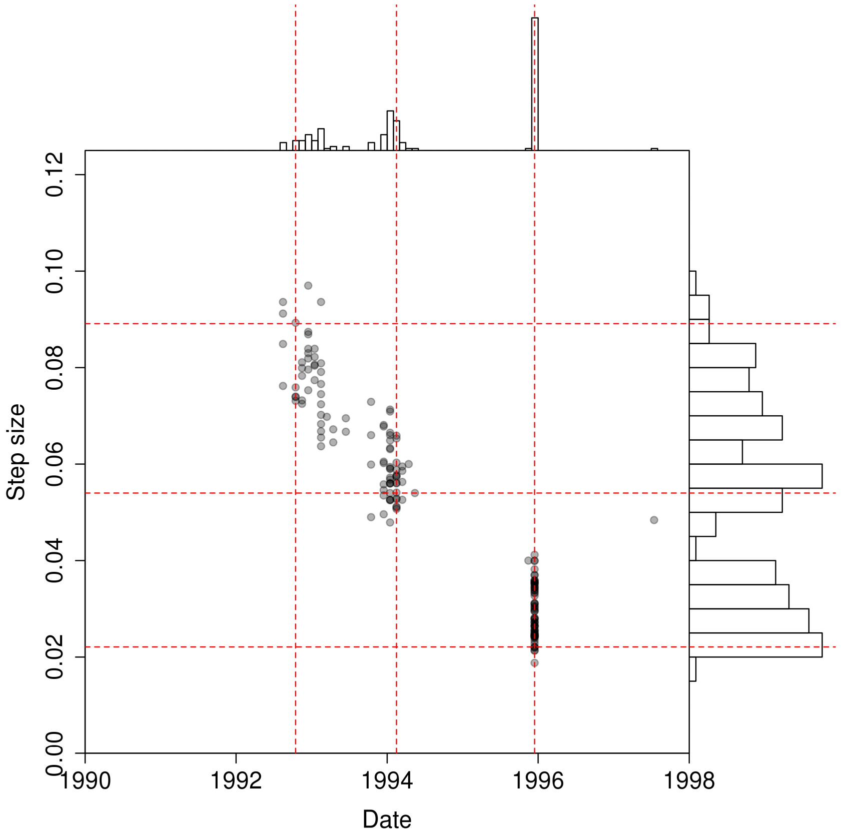

Figure 5. Scatter plot and histograms of the change point date versus step size for the GTMBA ensemble members where 1 change point was detected. Also shown are the median and 95% confidence intervals for the date and step size (red dashed lines).

4.1.2. Drifting Buoys

Table 5 summarizes the results of the change point analysis using the drifting buoy observations. One or more change points were detected in a large proportion of the ensemble members, with only 35% having no change point detected. As with the GTMBA, this is in contrast to a nominal FAR of 1% and actual FAR of 4.4% using the synthetic data. The majority of ensemble members with a change point have multiple change points detected. Figure 6 shows the timing of the change points, with the mean change point model averaged over ensemble members with the same number of change points shown in panel (a) (black lines) and histogram of change point dates in the panel (b). Also shown are the monthly mean differences for the individual ensemble members (panel (a); red lines). The change points occurring in January 2001 and July 2001 were very clear, both in the meantime series for the change point models and histogram of change point dates. There was also a clear warming in the satellite–buoy differences towards the end of the 1990s, but with more uncertainty over the timing of the change. For those ensemble members with a single change, the change occurs during the 1990s, with the record after ~2001 being relatively stable.

As noted above, the change points during 2001 (January and July) were associated with the failure of a gyroscope on the ERS-2 satellite and the period when only nadir (single) view SSTs were available. The impact of the use of single view SST retrieval during this period can be clearly seen in the uncertainty estimates provided in the L2P data (Figure 7). Figure 8 shows a scatter plot of the detected step sizes and dates across the ensemble (top) split into four groups. Individual scatter plots and histograms for the individual groups are also shown. The median values and 95% confidence intervals for the different groups are listed in Table 6 together with the size of each group. The first group (black circles, Figure 8a,b) is clustered in the mid to late 1990s, with a median step of 0.053 K and a median date of February 1999. This group is discussed in more detail below. The second group (red squares, Figure 8a,c) was coincident with the gyroscope failure, with a step of −0.060 K occurring during January 2001. The third group (green diamonds, Figure 8a,d) coincided with the commencement of the zero gyro mode in July 2007 and had a median step size of +0.054 K. The final group (blue triangles, Figure 8a,e) was relatively small and spread out over the AATSR record from middle 2002 onwards, with a cluster of change points detected around 2002–2003 and a small number later in the record. The median size change point for this group was −0.029 K with a date of July 2002. This final group coincided with start of the AATSR data, with the majority of change points flagged for 2002 across the ensemble only significant if supported by metadata.

Table 5. As Table 4 but for the drifting buoy analysis.

Number of Change Points	% of Ensemble Members with Detected Change Points	Observed False Alarm Rate (FAR) (%)
0	35.0	96.6
1	20.5	2.3
2	21.1	1.1
3	18.7	0.0
>3	4.7	0.0

Table 6. Identified step changes, size and dates for the drifting buoy analysis. The 95% confidence intervals have been estimated based on the 2.5% and 97.5% quantiles of the step sizes and dates.

Break Number (Count)	Step Size (K) (95% Confidence Interval)	Position (95% Confidence Interval)
1 (520)	0.053 (0.029–0.083)	15-02-1999 (15-12-1995 to 15-11-1999)
2 (424)	−0.060 (−0.114–−0.021)	15-01-2001 (15-07-2000 to 15-01-2001)
3 (356)	0.054 (0.015–0.102)	15-07-2001 (15-06-2001 to 15-07-2001)
4 (56)	−0.029 (−0.052–−0.017)	15-07-2002 (26-06-2002 to 02-01-2007)

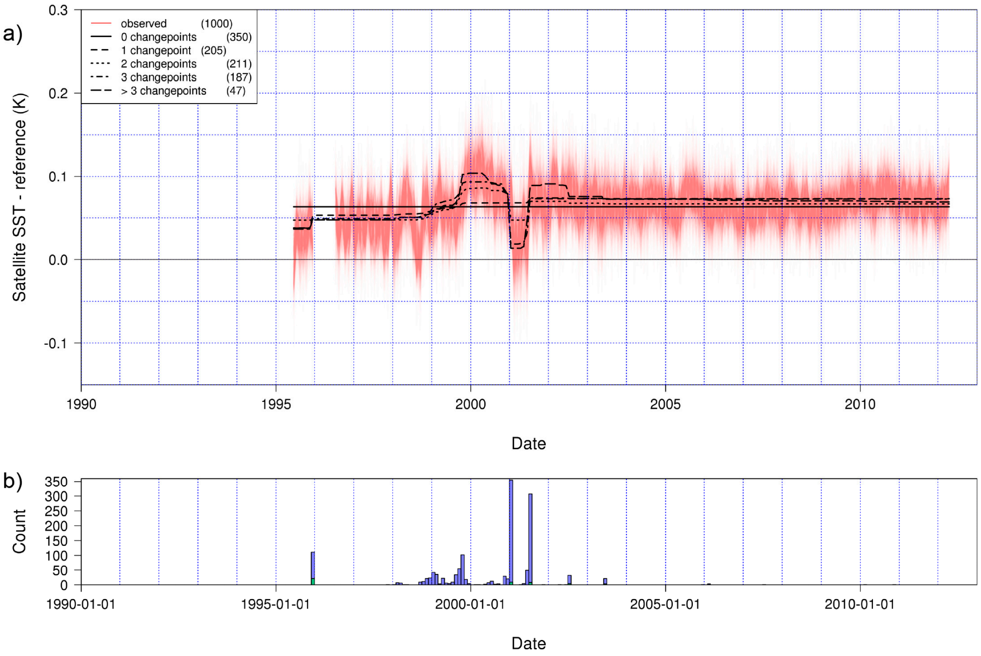

Figure 6. As Figure 3 but for drifting buoy observations. (**a**) The monthly mean differences (satellite in situ) for the individual ensemble members are shown in red. Also shown are the means of the model fitted by PMT grouped by number of change points detected. (**b**) Histogram of the data of the change points detected across all ensemble members.

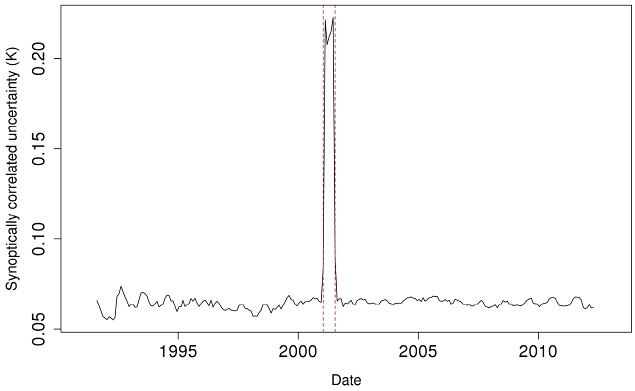

Figure 7. Monthly mean synoptically-correlated uncertainty estimate from the L2P data calculated across all available drifting buoy match-ups (black line). Also shown are the change points identified in 2001 (vertical red dashed lines).

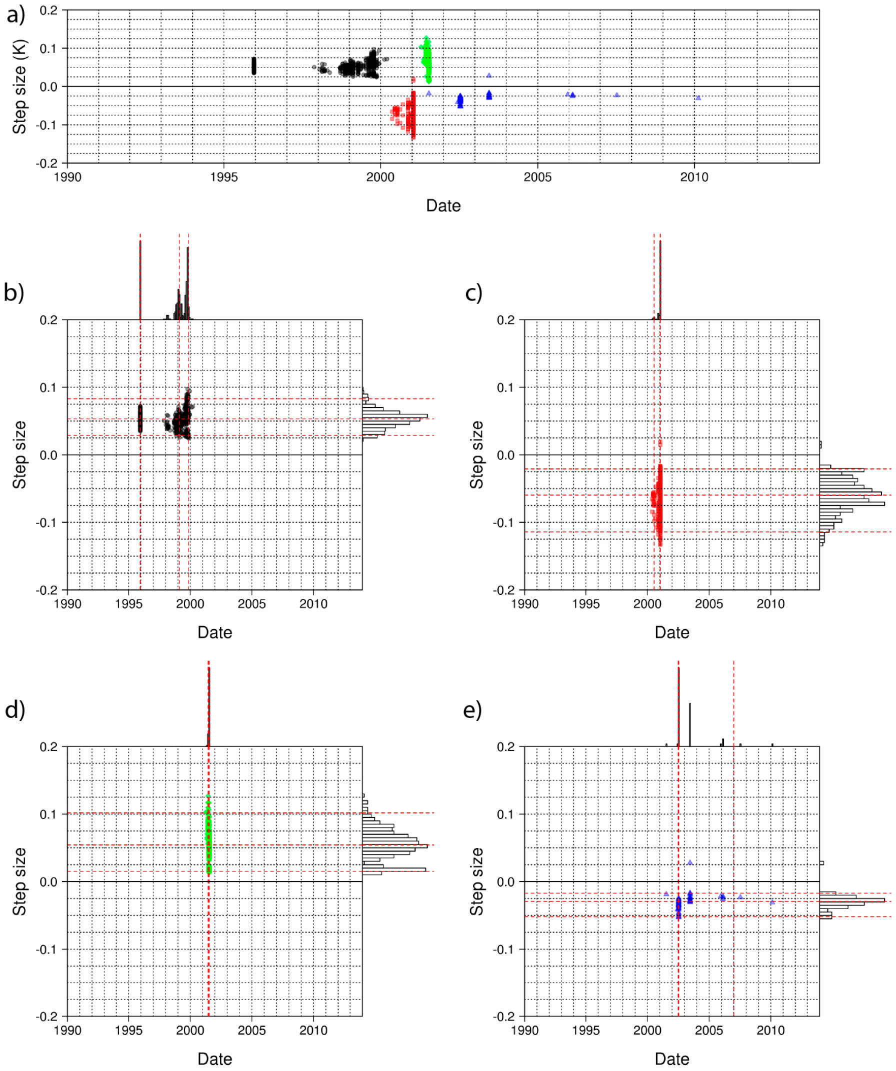

Figure 8. Scatter plot of the identified change point dates and step sizes for the drifting buoy analysis split into four groups. Panel (**a**) shows the steps identified over the full record—black represents break 1, red break 2, green break 3, and blue break 4. Panels (**b**–**e**) show scatter plots and histograms for the individual groups, the red dashed lines indicate the 2.5%, 50%, and 97.5% quantiles.

Whilst a large proportion of the ensemble had a change point detected pre-2000, this was a period where the drifting buoy network underwent a major evolution, with large changes to the spatial sampling by the drifting buoys (Figure 2). When these changes are coupled with regional differences previously reported for (A)ATSR in situ comparisons, with tropical differences tending to be positive and extra tropical differences cooler or negative (e.g., Figure 6 from [18]), we would expect to see some evidence in the time series of differences. As more observations and match-ups are made in the tropics we would expect to see a warming in the satellite–buoy differences, and this is what is seen in Figure 6 and the sign of the steps detected in the 1990s. As a result, it is likely that the pre-2000 change points were a feature of the evolution of the drifting buoy network and the changing spatial sampling by drifting buoys, rather than being attributed to the satellite data.

4.1.3. Argo

Figure 9 summarizes the results of the change point analysis using Argo data, with the monthly mean satellite in situ differences plotted for each ensemble member (red lines). In contrast to the drifting buoy network there were few match-ups available, e.g., Figure 1, and only 16 match-ups per month were used in this analysis. The impact of this reduced number of match-ups was evident in the variability of the monthly mean differences shown in Figure 9, and gaps in the record prior to the start of 2005. The lack of independence of the ensemble members was clearly seen, with less variability between the different ensemble members. This was also evident in the single change point detected compared to the ~10 that we would expect for an ensemble size of 1000 and 1% FAR. This suggests that there are no significant change points (relative to Argo) for the period 2005 onwards. The actual false alarm rate using synthetic data was slightly higher, with change points detected in 14 (1.4%) out of the 1000 ensemble members.

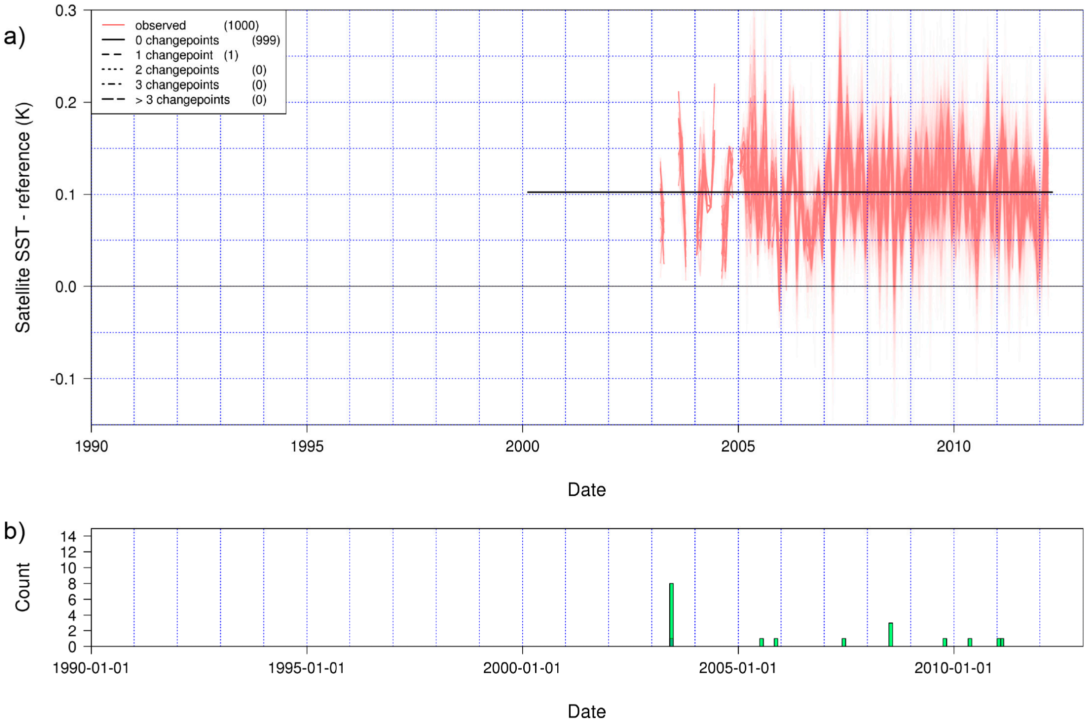

Figure 9. As for Figure 3 but for Argo observations. (**a**) The monthly mean differences (satellite in situ) for the individual ensemble members are shown in red. Also shown are the means of the model fitted by PMT grouped by number of change points detected. (**b**) Shows a histogram of the false alarms.

4.1.4. Sensitivity Tests

In addition to specifying our confidence in any detected change points and that they were not the result of chance (i.e., type I errors or FAR), it was also useful to specify our confidence of detecting a change point given our data and the probability of falsely accepting our null hypothesis. This concept is known as the statistical power of a test or the type II error rate, with the statistical power given by $1 - \beta$ where β is the type II error rate. The statistical power of the PMT has been quantified and published by Wang et al. [20]. Table 7 lists the ensemble median standard deviation of the monthly mean differences for the different analyses and number of months with valid data. Also listed are the mean hit rates of the PMT from Table 5 of Wang et al. [20], interpolated to match those in Table 7 and converted to statistical power (%).

Table 7. Median standard deviation of the monthly mean differences calculated across the different ensembles, ratio of the target step size (0.05) to standard deviation, number of months with valid data and estimated power interpolated from Table 5 of Wang et al. [20].

Analysis	Median Inter-Month Standard Deviation (σ)	$\frac{0.05}{\sigma}$	Number of Months	Power (%)
GTMBA	0.062	0.806	249	42.6%
Drifting buoys	0.039	1.282	203	69.6%
Argo	0.073	0.685	110	26.1%

From Table 7 it can be seen that the analysis based on the drifting buoy data had the greatest estimated statistical power, followed by the GTMBA and then the Argo analyses. To check the actual power of the analyses described we re-estimated the statistical power using the synthetic datasets described above. For each potential change point we inserted a step of 0.05 K into the satellite data in the synthetic datasets and repeated the change point analysis, sub-sampling, and applied the PMT 1000 times. We then calculated the power for each potential change point as the hit rate observed at that change point divided by 1000. The results expressed as % are shown in Figure 10. For the GTMBA array, the results indicated that there was a ~40–60% chance of detecting a step change using a single ensemble member over the majority of the period with data, but decreasing at either end of the time series. The impact of the sample size and reduction of noise with increasing sample size was seen, with a greater power of the drifting buoy analysis evident after ~2000. Based on this analysis, there was an ~80% chance of detecting a step of 0.05 K in a single ensemble member using the drifting buoys and sample size of 100. As with the GTMBA analysis, the power decreased towards either end of the time series. In contrast to the GTMBA and drifting buoy analysis, the Argo analysis showed very little power in detecting a step, due to the small number of match-ups available and larger inter-month standard deviation.

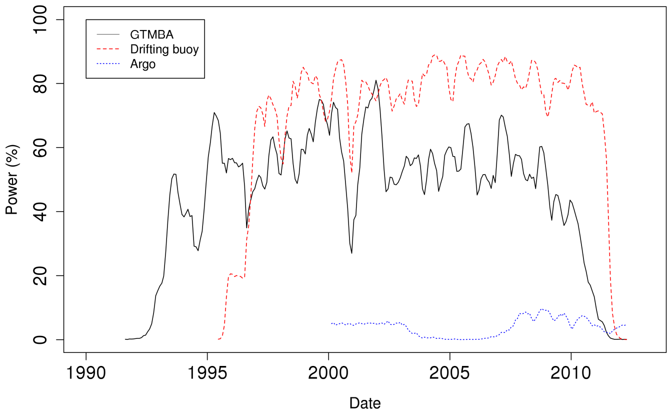

Figure 10. Estimated statistical power (%) for detection of 0.05 K step in the satellite data given the observing system characteristics and analysis configuration listed in Table 3.

These results were broadly comparable with those listed in Table 7. The results of the sensitivity tests were also broadly in line with the results for the different analysis presented above. Assuming that the detected change points were real, ~75% of the ensemble members for the drifting buoy analysis detected a change point, compared to a power of ~80%. For the GTMBA, change points were detected in ~20% of ensemble members, but towards the start of the time series. This was compared to a power of 40–50% in the middle of the time series, decreasing to 20% or less prior to 1995. Whilst the Argo statistical power was less than expected, the ensemble members used was not independent. This would lead to reduced power and an increased false alarm rate, as seen in this section and the previous section.

4.2. Stability Assessment

Figure 11 shows the monthly mean differences (black lines) for the different ensemble members from the GTMBA (panel (a)) and drifting buoy analysis (panel (c)) without adjustment for identified change points. Also shown are the results of fitting the AR1 trend model for each ensemble member (red lines). Figure 11 also shows the histogram of the fitted trend component for the GTMBA (panel (b)) and drifting buoys (panel (d)). No trend analysis has been performed using the Argo data due to the short time series available.

For both the GTMBA and the drifting buoy analysis the trends in the differenced time series were small. The median trend fitted to the GTMBA time series was 0.012 ± 0.015 K decade^{-1}, with the uncertainty equivalent to the median 95% confidence interval from the fitted AR1 models. The median trend component of the AR1 model fitted to the drifting buoy time series was 0.010 ± 0.014 K decade^{-1}. Table 8 lists the 2.5%, 50%, and 97.5% quantiles for both the trend estimate and confidence intervals across the ensemble. In both analyses, even with the detected change points, the stability of the satellite SST estimates relative to the in situ data was within the target 0.1 K decade^{-1}.

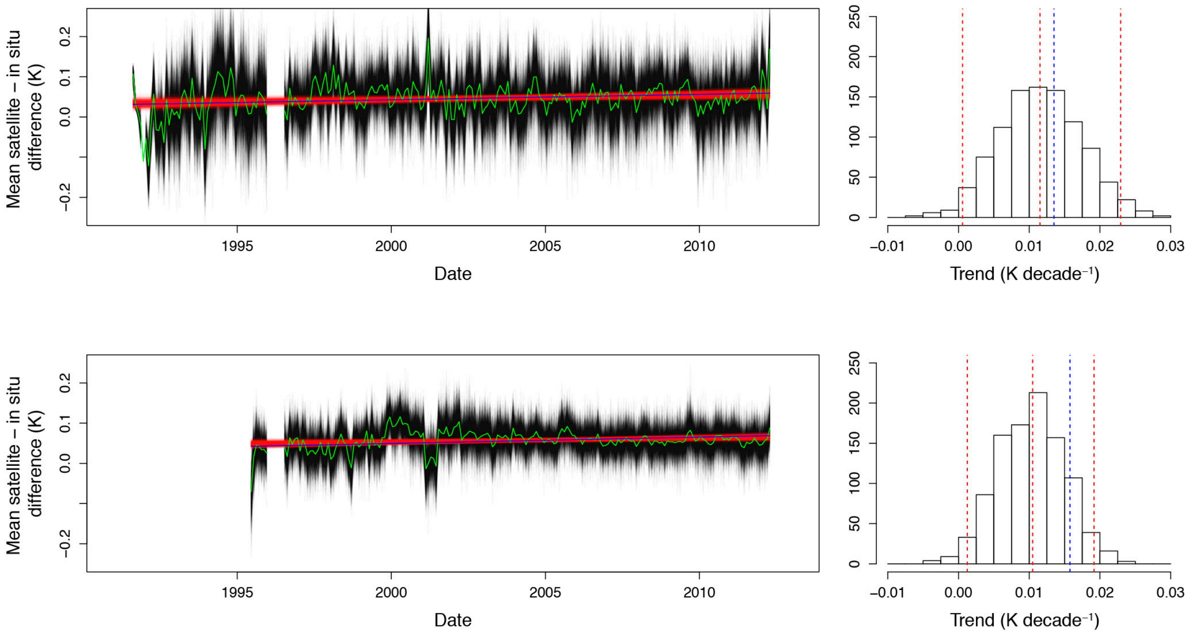

Figure 11. Panels (**a**,**c**): time series of the individual ensemble mean differences as a function of time (black lines) using observations from the (**a**) GTMBA and (**c**) drifting buoys, fitted AR1 trend models (red lines), mean differences averaged across all available match-ups (green lines) and AR1 model fitted to the mean of all available match-ups. Panels (**b**,**d**): histograms of the estimated trends across the ensembles for the (**b**) GTMBA and (**d**) drifting buoy data. The red dashed lines indicate the 2.5%, 50% and 97.5% quantiles. The blue dashed line indicates the trend fitted using all available match-ups.

Table 8. Trend and trend uncertainty estimates for the AR1 model applied to the GTMBA and drifting buoy ensembles.

Analysis	**Trend (K decade^{-1})**			**Trend Uncertainty (K decade^{-1})**		
	2.5%	**50%**	**97.5%**	**2.5%**	**50%**	**97.5%**
GTMBA	0.001	0.012	0.023	0.013	0.015	0.017
Drifting buoy	0.001	0.010	0.019	0.012	0.014	0.017

5. Conclusions

The ESA SST CCI project aims to generate an SST dataset that is independent of the in situ observations, free of inhomogeneity and with high temporal stability (drift <0.1 K decade^{-1}) in the global mean. Within this paper we have presented a method to test the homogeneity of the SST values from the CCI project in comparison to in situ observations, testing for change points in the SST retrievals. This work builds on that of others, extending the change-point analysis from stations at fixed locations to mobile station data through the use of aggregate time series and an ensemble approach. The use of aggregate time series increases the sensitivity of the tests applied by minimizing observational errors and noise through averaging. The use of ensembles allows us to quantify the sampling uncertainty due to our method and sub-sampling of the available match-up data to form our aggregate time series.

We have applied the method to three different sources of in situ data and SST retrievals based on the (A)ATSR series of sensors. The in situ sources were: (1) the GTMBA array, (2) the drifting buoy array (ATSR2 onwards), and (3) near surface Argo data (~2004 onwards). The results of the analysis using the GTMBA data were broadly comparable to that previously found in the ARC project [35], with evidence of a warming of the satellite SST retrievals compared to the GTMBA data between 1993 and 1995, and coincident with the decreasing levels of atmospheric aerosols following the Pinatubo eruption in 1991. In addition to recovery following Pinatubo, there is evidence of a pair of change points in the satellite data during 2001 in both the GTMBA and drifting buoy analysis. The first step detected occurs in January 2001 and is coincident with the failure of a gyroscope on the ERS-2 satellite. The second change point occurs in July 2001 and is coincident with beginning of the zero gyro mode SST retrievals in the ATSR 2 data. During this period (January–July 2001) only nadir view retrievals were available, with a substantial increase in the uncertainty in the satellite SST values. No change points were detected for the period including the Argo data.

The statistical power of the different analyses has been estimated, based on both the published power of the PMT [20] and an empirical assessment as part of the work presented in this paper. The drifting buoy analysis has the greatest statistical power, with a 70–80% chance of detecting a step of 0.05 K using a single ensemble member. For example, for an ensemble of 20 independent members containing a step of 0.05 K, we would expect 14 to 16 members to detect the step, and at most, one false positive. For a moderate ensemble size (~20–50) we can be virtually certain that a step of this magnitude would be detected. The power of the moored buoy analysis was in the range 40–50% for a step of 0.05 K. Again for a moderate ensemble size we can be virtually certain of detecting a change point in the data. The analysis using the Argo data showed much less statistical power and it is doubtful that we would be able to generate enough independent ensemble members to significantly increase the power. It should be noted that the success of the drifting buoy analysis and greater statistical power is primarily due to the large number of drifting buoys making SST observations. This can, in part, be credited to the use of design metrics, such as the equivalent buoy density and buoy need index of Zhang et al. [46], for the drifting buoy array.

In addition to the change point analysis, we have assessed the satellite in situ differences for drift by fitting an AR1 trend model to the monthly mean differences. We have applied the model to both the individual ensemble members and monthly mean differences using all available match-ups. Similar results are found using both methods, with small trends of order 0.01 K decade^{-1} present in the monthly mean differences. Uncertainties in the trend, both due to the model fit and sampling across the ensemble, are of a similar or slightly larger magnitude to the trend itself. The observed trends in the differences are significantly smaller than the target 0.1 K decade^{-1}.

Future assessments of the ESA SST CCI dataset(s) will be needed. For example, version 2 of the CCI dataset will contain SST retrievals from both the AVHRR and (A)ATSR sensors, and use updated algorithms. It is recommended that future assessments are made using the GTMBA match-ups in the tropics for the full period and drifting buoy data globally from ~2000 onwards. Whilst Argo has shown limited utility in the analyses presented, this is primarily due to the limited number of match-ups available for the (A)ATSR data. For AVHRR we would expect a greater number of match-ups due to the wider swath width, resulting in greater statistical power and an increased ability to detect small changes in the satellite data. This will be tested in future analyses.

Finally, the method developed in this paper is not restricted to the (A)ATSR sensors or to sea surface temperature. Provided that suitable sample sizes that are independent or near independent of each other can be generated from the available matchups and that the impact of changes to the in situ reference networks can be quantified the method can be applied. For example, the method could be applied to SST data from MODIS or from microwave sensors. Similarly, the data could be applied to wind speed or humidity retrievals where suitable reference data exists. Based on the results above for SST, and work by previous authors (e.g., [47]), a minimum sample size of ~100 matchups is recommended, but this will vary with the variable of interest and the quality of the reference data. For coastal regions, where sampling is more limited but there are more fixed stations, the use of the operational moored buoy network and methods such as the pairwise comparison test of Menne and Williams [40] may be more appropriate. However, this would require improvement to the operational buoy metadata record.

Acknowledgments: The work undertaken in this paper was funded by the European Space Agency (ESA ESRIN REF 3-13904/13/I-NB) Sea Surface Temperature Climate Change Initiative project. Analysis has been performed on the UK Science and Technology Facilities Councils (STFC) JASMIN supercomputing facility using version 3.4.0 of the R programming language. The RHTest v4 R package has been used to perform PMT tests and is available from http://etccdi.pacificclimate.org/software.shtml.

Author Contributions: David I. Berry wrote the paper, designed and developed the method to apply the homogeneity testing methods to mobile data, and performed the analysis. Gary K. Corlett provided the data from the match-up database used, provided guidance on how to use the match-ups, and contributed to the interpretation of the results. Owen Embury contributed to the interpretation of the results and provided guidance on events that may impact on the quality of the satellite data. Christopher J. Merchant conceived the assessment of the stability of the satellite SSTs using Argo and the other in situ platforms, building on the work of the ARC project, and contributed to the interpretation of the results.

Conflicts of Interest: The authors declare no conflict of interest.

Glossary

AATSR	Advanced Along Track Scanning Radiometer
AR1	Lag 1 Auto-Regressive Model
ARC	(A)ATSR Reprocessing For Climate Project
ATSR	Along Track Scanning Radiometer
AVHRR	Advanced Very High Resolution Radiometer
CCI	Climate Change Initiative
ECMWF	European Centre For Medium-Range Weather Forecasts
EN4	Hadley Centre EN4 Dataset
EO	Earth Observation
ERA	ECMWF Re-Analysis
ERS-1	European Remote Sensing Satellite 1
ERS-2	European Remote Sensing Satellite 2
ESA	European Space Agency
FAR	False Alarm Rate
GDAC	Global Data Assembly Centre

GTMBA	Global Tropical Moored Buoy Array
HadIOD	Hadley Centre Integrated Ocean Database
ICOADS	International Comprehensive Ocean-Atmosphere Data Set
L2P	Level 2 Pre-Processed
MMD	Multi-Sensor Match-Up Database
PMT	Penalized Maximal t Test
RSD	Robust Standard Deviation
SNHT	Standard Normal Homogeneity Test
SST	Sea Surface Temperature
WOD	World Ocean Database

References

1. Stocker, T.F.; Qin, D.; Plattner, G.-K.; Tignor, M.; Allen, S.K.; Boschung, J.; Nauels, A.; Xia, Y.; Bex, V.; Midgley, P.M. (Eds.) *IPCC Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2013.
2. Blunden, J.; Arndt, D.S. State of the Climate in 2016. *Bull. Am. Meteorol. Soc.* **2017**, *98*, S93–S128. [CrossRef]
3. Huang, B.; Thorne, P.W.; Banzon, V.F.; Boyer, T.; Chepurin, G.; Lawrimore, J.H.; Menne, M.J.; Smith, T.M.; Vose, R.S.; Zhang, H.-M. Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *J. Clim.* **2017**, *30*, 8179–8205. [CrossRef]
4. Rayner, N.A.; Brohan, P.; Parker, D.E.; Folland, C.K.; Kennedy, J.J.; Vanicek, M.; Ansell, T.J.; Tett, S.F.B. Improved analyses of changes and uncertainties in sea surface temperature measured in situ sice the mid-nineteenth century: The HadSST2 dataset. *J. Clim.* **2006**, *19*, 446–469. [CrossRef]
5. Kennedy, J.J.; Rayner, N.A.; Smith, R.O.; Parker, D.E.; Saunby, M. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *J. Geophys. Res.* **2011**, *116*. [CrossRef]
6. Kennedy, J.J.; Rayner, N.A.; Smith, R.O.; Parker, D.E.; Saunby, M. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res.* **2011**, *116*. [CrossRef]
7. Hirahara, S.; Ishii, M.; Fukuda, Y. Centennial-Scale Sea Surface Temperature Analysis and Its Uncertainty. *J. Clim.* **2014**, *27*, 57–75. [CrossRef]
8. Casey, K.S.; Brandon, T.B.; Cornillon, P.; Evans, R. The Past, Present, and Future of the AVHRR Pathfinder SST Program. In *Oceanography from Space: Revisited*; Barale, V., Gower, J.F.R., Alberotanza, L., Eds.; Springer: Berlin, Germany, 2010; pp. 273–287.
9. Woodruff, S.D.; Worley, S.J.; Lubker, S.J.; Ji, Z.; Freeman, J.E.; Berry, D.I.; Brohan, P.; Kent, E.C.; Reynolds, R.W.; Smith, S.R.; et al. ICOADS Release 2.5: Extensions and enhancements to the surface marine meteorological archive. *Int. J. Climatol.* **2011**, *31*, 951–967. [CrossRef]
10. Freeman, E.; Woodruff, S.D.; Worley, S.J.; Lubker, S.J.; Kent, E.C.; Angel, W.E.; Berry, D.I.; Brohan, P.; Eastman, R.; Gates, L.; et al. ICOADS Release 3.0: A major update to the historical marine climate record. *Int. J. Climatol.* **2017**, *37*, 2211–2232. [CrossRef]
11. Smith, T.M.; Reynolds, R.W. Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997). *J. Clim.* **2003**, *16*, 1495–1510. [CrossRef]
12. Reynolds, R.W.; Smith, T.M.; Liu, C.; Chelton, D.B.; Casey, K.S.; Schlax, M.G. Daily high-resolution-blended analyses for sea surface temperature. *J. Clim.* **2007**, *20*, 5473–5496. [CrossRef]
13. Embury, O.; Merchant, C.J.; Filipiak, M.J. A reprocessing for climate of sea surface temperature from the along-track scanning radiometers: Basis in radiative transfer. *Remote Sens. Environ.* **2012**, *116*, 32–46. [CrossRef]
14. Embury, O.; Merchant, C.J. A reprocessing for climate of sea surface temperature from the along-track scanning radiometers: A new retrieval scheme. *Remote Sens. Environ.* **2012**, *116*, 47–61. [CrossRef]
15. Fairall, C.W.; Bradley, E.F.; Godfrey, J.S.; Wick, G.A.; Edson, J.B.; Young, G.S. Cool-skin and warm-layer effects on sea surface temperature. *J. Geophys. Res.* **1996**, *101*, 1295–1308. [CrossRef]

16. Kantha, L.H.; Clayson, C.A. An improved mixed-layer model for geophysical applications. *J. Geophys. Res. Ocean.* **1994**, *99*, 25235–25266. [CrossRef]
17. Hollmann, R.; Merchant, C.J.; Saunders, R.; Downy, C.; Buchwitz, M.; Cazenave, A.; Chuvieco, E.; Defourny, P.; de Leeuw, G.; Forsberg, R.; et al. The ESA climate change initiative Satellite Data Records for Essential Climate Variables. *Bull. Am. Meteorol. Soc.* **2013**, *94*, 1541–1552. [CrossRef]
18. Merchant, C.J.; Embury, O.; Roberts-Jones, J.; Fiedler, E.; Bulgin, C.E.; Corlett, G.K.; Good, S.; McLaren, A.; Rayner, N.; Morak-Bozzo, S.; et al. Sea surface temperature datasets for climate applications from Phase 1 of the European Space Agency Climate Change Initiative (SST CCI). *Geosci. Data J.* **2014**, *1*, 179–191. [CrossRef]
19. Venema, V.K.C.; Mestre, O.; Aguilar, E.; Auer, I.; Guijarro, J.A.; Domonkos, P.; Vertacnik, G.; Szentimrey, T.; Stepanek, P.; Zahradnicek, P.; et al. Benchmarking homogenization algorithms for monthly data. *Clim. PAST* **2012**, *8*, 89–115. [CrossRef]
20. Wang, X.L.; Wen, Q.H.; Wu, Y. Penalized maximal *t* test for detecting undocumented mean change in climate data series. *J. Appl. Meteorol. Climatol.* **2007**, *46*, 916–931. [CrossRef]
21. McPhaden, M.J.; Busalacchi, A.J.; Cheney, R.; Donguy, J.R.; Gage, K.S.; Halpern, D.; Ji, M.; Julian, P.; Meyers, G.; Mitchum, G.T.; et al. The tropical ocean global atmosphere observing system: A decade of progress. *J. Geophys. Res.* **1998**, *103*, 14169–14240. [CrossRef]
22. Lumpkin, R.; Centurioni, L.; Perez, R.C. Full access fulfilling observing system implementation requirements with the global drifter array. *J. Atmos. Ocean. Technol.* **2016**. [CrossRef]
23. Argo. Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). *Seanoe* **2000**. [CrossRef]
24. Merchant, C.J. Algorithm Theoretical Basis Document (Phase II EXP 1.8). European Space Agency Contract Report SST_CCI-ATBD-UOR-202. Available online: http://www.esa-sst-cci.org/PUG/pdf/SST_CCI-ATBD-UOR-202_Issue-1-signed.pdf (accessed on 16 January 2018).
25. Corlett, G.K. MMD Content Specification. European Space Agency Contract Report SST_CCI-TN-UOL-001. Available online: http://www.esa-sst-cci.org/sites/default/files/Documents/public/SST_cciMMDContentSpecificationIssue1(20120504).pdf (accessed on 16 January 2018).
26. Atkinson, C.P.; Rayner, N.A.; Kennedy, J.J.; Good, S.A. An integrated database of ocean temperature and salinity observations. *J. Geophys. Res.* **2014**, *119*, 7139–7163. [CrossRef]
27. Good, S.A.; Martin, M.J.; Rayner, N.A. EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res.* **2013**, *118*, 6704–6716. [CrossRef]
28. Boyer, T.P.; Antonov, J.I.; Baranova, O.K.; Coleman, C.; Garcia, H.E.; Grodsky, A.; Johnson, D.R.; Locarnini, R.A.; Mishonov, A.V.; Brien, T.D.O.; et al. *World Ocean Database 2013*; Levitus, S., Mishonov, A., Eds.; NOAA Printing Office: Silver Spring, MD, USA, 2013.
29. Kent, E.C.; Kennedy, J.J.; Smith, T.M.; Hirahara, S.; Huang, B.; Kaplan, A.; Parker, D.E.; Atkinson, C.P.; Berry, D.I.; Carella, G.; et al. A call for new approaches to quantifying biases in observations of sea surface temperature. *Bull. Am. Meteorol. Soc.* **2017**, *98*, 1601–1616. [CrossRef]
30. Castro, S.L.; Wick, G.A.; Emery, W.J. Evaluation of the relative performance of sea surface temperature measurements from different types of drifting and moored buoys using satellite-derived reference products. *J. Geophys. Res.* **2012**, *117*. [CrossRef]
31. Atkinson, C.P.; Rayner, N.A.; Roberts-Jones, J.; Smith, R.O. Assessing the quality of sea surface temperature observations from drifting buoys and ships on a platform-by-platform basis. *J. Geophys. Res.* **2013**, *118*, 3507–3529. [CrossRef]
32. Ingleby, B.; Huddleston, M. Quality control of ocean temperature and salinity profiles-Historical and real-time data. *J. Mar. Syst.* **2007**, *65*, 158–175. [CrossRef]
33. Lean, K.; Saunders, R.W. Validation of the ATSR Reprocessing for Climate (ARC) Dataset Using Data from Drifting Buoys and a Three-Way Error Analysis. *J. Clim.* **2013**, *26*, 4758–4772. [CrossRef]
34. Emery, W.J.; Baldwin, D.J.; Schlussel, P.; Reynolds, R.W. Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements. *J. Geophys. Res.* **2001**, *106*, 2387–2405. [CrossRef]
35. Merchant, C.J.; Embury, O.; Rayner, N.A.; Berry, D.I.; Corlett, G.K.; Lean, K.; Veal, K.L.; Kent, E.C.; Llewellyn-Jones, D.T.; Remedios, J.J.; et al. A 20 year independent record of sea surface temperature for climate from Along-Track Scanning Radiometers. *J. Geophys. Res.* **2012**, *117*. [CrossRef]
36. Donlon, C.J.; Martin, M.; Stark, J.; Roberts-Jones, J.; Fiedler, E.; Wimmer, W. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sens. Environ.* **2012**, *116*, 140–158. [CrossRef]

37. Dee, D.; Uppala, S.; Simmons, A.; Berrisford, P.; Poli, P.; Kobayashi, S.; Andrae, U.; Balmaseda, M.; Balsamo, G.; Bauer, P. The ERA—Interim reanalysis: Configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **2011**. [CrossRef]
38. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2017.
39. Alexandersson, H. A homogeneity test applied to precipitation data. *J. Climatol.* **1986**, *6*, 661–675. [CrossRef]
40. Menne, M.J.; Williams, C.N., Jr. Homogenization of Temperature Series via Pairwise Comparisons. *J. Clim.* **2009**, *22*, 1700–1717. [CrossRef]
41. Meindl, E.A.; Hamilton, G.D. Programs of the National-Data_buoy_center. *Bull. Am. Meteorol. Soc.* **1992**, *73*, 985–993. [CrossRef]
42. Wang, X.L. Accounting for autocorrelation in detecting mean shifts in climate data series using the penalized maximal t or F test. *J. Appl. Meteorol. Climatol.* **2008**. [CrossRef]
43. Wang, X.L.; Feng, Y. RHtestsV4 User Manual; Climate Research Division, Atmospheric Science and Technology Directorate, Science and Technology Branch, Environment Canada. Available online: http://etccdi.pacificclimate.org/software.shtml (accessed on 16 January 2018).
44. Neter, J.; Kutner, M.H.; Nachtsheim, C.J.; Wasserman, W. *Applied Linear Statistical Models*, 4th ed.; McGraw Hill: Boston, MA, USA, 1996.
45. Hyndman, R.J.; Fan, Y. Sample Quantiles in Statistical Packages. *Am. Stat.* **1996**. [CrossRef]
46. Zhang, H.M.; Reynolds, R.W.; Smith, T.M. Adequacy of the in situ observing system in the satellite era for climate SST. *J. Atmos. Ocean. Technol.* **2006**, *23*, 107–120. [CrossRef]
47. Kilpatrick, K.A.; Podestá, G.; Walsh, S.; Williams, E.; Halliwell, V.; Szczodrak, M.; Brown, O.B.; Minnett, P.J.; Evans, R. A decade of sea surface temperature from MODIS. *Remote Sens. Environ.* **2015**. [CrossRef]