Stability of cloud detection methods for Land Surface Temperature (LST) Climate Data Records (CDRs)

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Stability of cloud detection methods for Land Surface Temperature (LST) Climate Data Records (CDRs)

Claire E. Bulgin [a,b,*], Ross I. Maidment [a,b], Darren Ghent [c,d], Christopher J. Merchant [a,b]

[a] *Department of Meteorology, University of Reading, Reading RG6 6ET, UK*
[b] *National Centre for Earth Observation, University of Reading, Reading RG6 6AL, UK*
[c] *University of Leicester, Space Park Leicester, Leicester LE4 5SP, UK*
[d] *National Centre for Earth Observation, Space Park Leicester, Leicester LE4 5SP, UK*

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Stability
Land surface temperature
Climate data record
Cloud detection

ABSTRACT

The stability of a climate data record (CDR) is essential for evaluating long-term trends in surface temperature using remote sensing products. In the case of a satellite-derived CDR of land surface temperature (LST), this includes the stability of processing steps prior to the estimation of the target climate variable. Instability in the masking of cloud-affected observations can result in non-geophysical trends in a LST CDR. This paper provides an assessment of cloud detection performance stability over a 25-year LST CDR generated using data from the second Along-Track Scanning Radiometer (ATSR-2), the Advanced Along-Track Scanning Radiometer (AATSR), the Moderate Resolution Imaging Spectroradiometer (MODIS) and the Sea and Land Surface Temperature Radiometer (SLSTR). We evaluate three cloud detection methodologies, one fully Bayesian, one naïve probabilistic and the operational threshold-based cloud mask provided with each sensor, at four in-situ ceilometer sites. Of the 12 algorithm-site combinations assessed, only two (17 %) were stable across the full timeseries with respect to both cloud contamination and missed clear-sky observations. Five (42 %) were stable with respect to missed clear-sky observations only. The associated impacts on LST trends in the CDR could be as large as (+/−) 0.73 K per decade (0.43 K per decade above the target stability), which means that attention needs to be paid to this aspect of stability in order to understand uncertainty in long-term observed trends. Given that cloud detection stability has not to our knowledge been previously assessed for any target climate variable, this conclusion may apply more broadly to other satellite-derived CDRs.

1. Introduction

Climate data records (CDRs) of Earth surface temperature are becoming increasingly important for assessing global temperature trends over recent decades (Bento et al., 2017; Riffler et al., 2015; Foster and Rahmstorf, 2011; Merchant et al., 2019; Bulgin et al., 2020). CDRs generated using remote sensing data typically combine observations from several sensors to make multidecadal records (Duguay-Tetzlaff et al., 2015; Lieberherr and Wunderle, 2018; Merchant et al., 2019), improving the signal-to-noise ratio for detecting climate induced changes (Foster and Rahmstorf, 2011). The temporal stability of a surface temperature CDR is critical for valid detection and attribution of a climate signal in temperature trends (Good et al., 2022; Kogler et al., 2012). This research is therefore relevant for a wide range of climate applications of land surface temperature (LST) data, including urban LST (Ding et al., 2020; J. A. Peeling et al., 2024), surface energy balance (Ji et al., 2019), crop stress (Anderson et al., 2016), land cover change (Kayet et al., 2016) and LST angular effects (He et al., 2024; Na et al., 2024).

Surface temperature CDRs can be generated from satellite sensors measuring at infrared or microwave wavelengths (Li et al., 2013). For climate studies, the LST user community requests global data at 0.05 degrees (Aldred et al., 2023), necessitating the use of infrared sensors providing data at a higher spatial resolution than their microwave counterparts (Li et al., 2013). Threshold level stability requirements (the minimum level for which the data are useable) for LST are 0.3 K per decade, with breakthrough and goal targets of 0.2 K and 0.1 K per decade respectively (Global Climate Observing System, 2022). CDR stability is often assessed by comparison of the retrieved geophysical variable (in this case surface temperature) to another source of data, for

[*] Corresponding author at: Department of Meteorology, University of Reading, Reading RG6 6AL, UK.
E-mail address: c.e.bulgin@reading.ac.uk (C.E. Bulgin).

https://doi.org/10.1016/j.rse.2024.114440
Received 14 June 2024; Received in revised form 28 August 2024; Accepted 21 September 2024
Available online 2 October 2024

example in-situ measurements (Good et al., 2017, 2022; Merchant et al., 2019; Berry et al., 2018). Interpretation of such comparisons needs to account for the level of instability in the reference data as well as the CDR.

Retrieving LST from infrared sensors requires a pre-processing step to detect and remove cloud contaminated observations (Bulgin et al., 2022; Frey et al., 2008; Simpson et al., 2001; Závody et al., 2000). No cloud detection methodology is perfect (Bulgin et al., 2018; Simpson et al., 2000), and therefore all surface-temperature CDRs suffer from a degree of cloud contamination causing corresponding errors in retrieved surface temperature.

Temporal stability in the performance of a cloud detection algorithm as applied to a surface temperature CDR, is key to ensuring that non-geophysical temperature trends do not arise as a direct result of this pre-processing step. Where CDRs are constructed using data from multiple sensors, cloud detection methodologies may differ as data producers often focus on applying a sensor-specific "best" algorithm rather than a CDR-consistent algorithm (Kogler et al., 2012). This can result in systematic changes to the fraction of cloud-contaminated data over time or non-physical "jumps" in cloud detection performance between sensors, attributable to differences in algorithm or channel selection. Even in the application of a consistent algorithm across multiple sensors, changes in channel calibration can cause differences in cloud detection performance.

Although cloud contamination can lead to significant biases in the retrieved surface temperature, quantifying the uncertainty that arises from miss-classification in the cloud-clearing process remains a challenge. Establishing a cloud detection 'truth' against which to evaluate automated cloud-clearing performance is challenging. Field of view differences between satellite and ground-based instruments can limit comparisons and generating cloud masks from semi-automated systems or expert inspection is time consuming even for small amounts of data (Bulgin et al., 2022).

For LST retrievals made on the satellite image grid, a given 'clear-sky' pixel meets one of two criteria: 1) it has been mistakenly classified as clear-sky, in which case the retrieved surface temperature is uncertain due to the presence of cloud in the satellite field of view or 2) it has been correctly classified as clear-sky and there is no associated uncertainty due to cloud contamination. For higher-level products where LST retrievals are re-gridded or averaged, the propagation of this uncertainty from the cloud contaminated observations needs to be accounted for. As such, many uncertainty budgets are unable to quantify this uncertainty component despite noting its importance (Bulgin et al., 2016a; Ghent et al., 2019).

In this paper, we evaluate the stability in cloud detection performance across a 25-year LST CDR generated within the European Space Agency (ESA) Climate Change Initiative (CCI) programme (Hollmann et al., 2013). Three different cloud-clearing algorithms are compared: a) a fully Bayesian clear/not-clear classifier (Bulgin et al., 2022; Merchant et al., 2005), b) a naïve probabilistic approach (Bulgin et al., 2014) and c) the threshold-based operational cloud detection algorithms for each sensor (Ackerman et al., 1998; Birks, 2007; European Space Agency, 2023; Závody et al., 2000). Comparisons are made against in-situ ceilometer data at all locations with long-term data records: Ny Alesund, North Slope of Alaska, Oliktok Point and Southern Great Plains. We first asses the stability of the cloud detection algorithms, considering the frequency with which clear-sky observations are actually cloud-contaminated and the number of clear-sky observations erroneously flagged as cloud. We then assess the impact of the cloud detection performance on the LST data record in comparison with clear-sky data and quantify the timeseries biases associated with mis-classification of clear and cloudy pixels, before assessing the characteristics of the observations that are incorrectly flagged.

The rest of this paper is structured as follows: Section 2 describes the overall workflow of the study. Section 3 describes both the contents (satellite data, in-situ data, cloud detection algorithms and LST retrieval) and the construction of the match-up dataset used for all the analysis in this paper. Section 4 gives details of the metrics and methods used for data analysis and Section 5 contains the results. Section 6 includes a detailed discussion of the findings and the paper concludes in Section 7.

2. Methodology and workflow

Fig. 1 shows the workflow for the study. On the left are the main components of the workflow and on the right the inputs at each stage of the workflow, which relate to different sections of this manuscript. The first stage is the generation of a match-up database, which requires both satellite and ceilometer data inputs as described in Section 2. The next stage is the calculation of the key metrics for assessing cloud detecting stability; the fraction of clear-sky pixels that are actually cloud contaminated (CC) and the fraction of clear-sky pixels 'missed' by the cloud detection algorithm (MC). Full details of this stage are found in Section 3. The third stage is to assess the cloud detection stability including the consideration of uncertainties and external factors. These results are shown in Section 4. The final stage is to evaluate the impact of cloud detection instability on the LST stability (Section 5).

3. Match-up dataset

A match-up dataset was developed for the analysis in this paper, synthesising both satellite observations and in-situ data. The subsections below describe each of the inputs in detail, after which we describe the matching process.

3.1. Satellite data

The LST CDR is comprised of data from four satellite instruments: ATSR-2 aboard the second European Remote sensing Satellite (ERS-2), AATSR aboard the Environmental Satellite (EnviSat), MODIS aboard the National Oceanic and Atmospheric Administration Terra platform and SLSTR-A aboard the Sentinel 3A satellite (Donlon et al., 2012; IDEAS+AATSR QC Team, 2016; Masuoka et al., 1998). Each satellite in the CDR is polar-orbiting with a local-time equator overpass of 10:30 (22:30) for ATSR-2 and MODIS, and 10:00 (22:00) for AATSR and SLSTR. The four instruments all have channels at infrared and visible wavelengths, designed to facilitate both surface temperature retrieval and cloud detection.

Table 1 summarises the characteristics of the four instruments. ATSR-2 and AATSR data have a resolution of 1 km at nadir. For MODIS, only the 1 km resolution channels are listed in Table 1 (the instrument has a further 8 channels at 250–500 m resolution, not relevant to this study). SLSTR data at reflectance wavelengths have a resolution of 500 m, mapped independently of the infrared data to the satellite image grid. The higher resolution channels are aligned with the infrared using a simple 2 × 2 averaging (Coppo et al., 2010). MODIS and SLSTR have a wider swath width than ATSR-2 and AATSR instruments. MODIS and SLSTR observations are therefore limited to the satellite zenith angle range of 0–22 degrees (effectively reducing the usable swath width to match the earlier instruments), to ensure consistency between all instruments in the CDR. The data used in this study are from v1.00 of the LST CCI algorithm (Perry et al., 2020). No temporal correction has been made to the nominal overpass time of the satellite in the data used for this study to account for the time difference across the satellite swath, but time differences are limited by the restricted satellite zenith angle range.

3.2. In-situ ceilometer data

The satellite data are matched to the locations of four ceilometers, which provide a measure of the height of the lowest cloud base. These ceilometers are located in Ny Alesund in Svalbard (Maturilli and Herber,

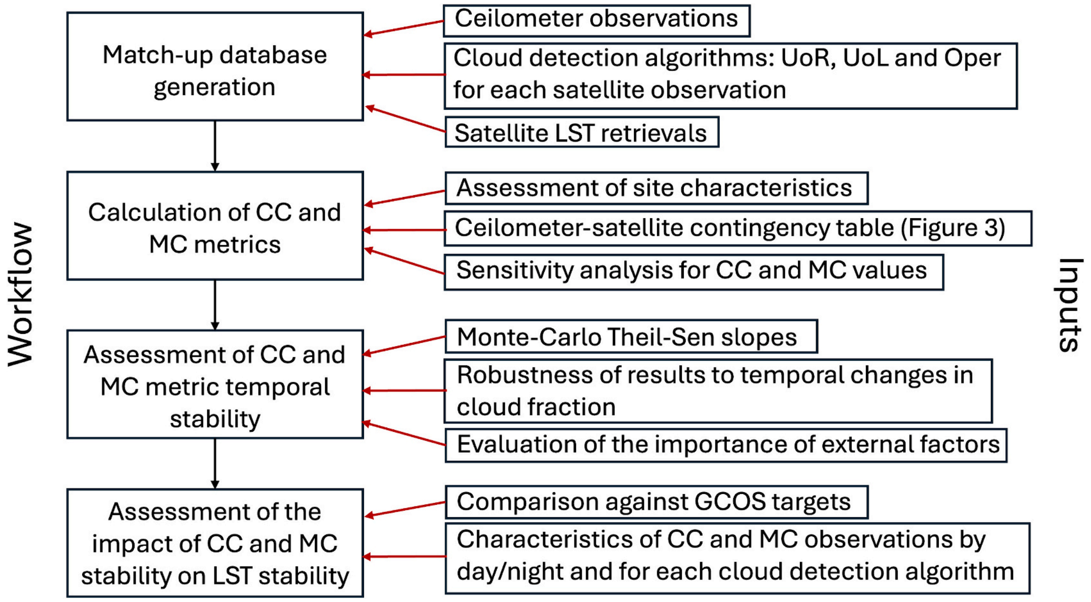
Workflow
Match-up database generation
Calculation of CC and MC metrics
Assessment of CC and MC metric temporal stability
Assessment of the impact of CC and MC stability on LST stability
Inputs
Ceilometer observations
Cloud detection algorithms: UoR, UoL and Oper for each satellite observation
Satellite LST retrievals
Assessment of site characteristics
Ceilometer-satellite contingency table (Figure 3)
Sensitivity analysis for CC and MC values
Monte-Carlo Theil-Sen slopes
Robustness of results to temporal changes in cloud fraction
Evaluation of the importance of external factors
Comparison against GCOS targets
Characteristics of CC and MC observations by day/night and for each cloud detection algorithm

Fig. 1. Workflow for the assessment of cloud detection stability and its impact on LST stability.

Table 1
Instrument characteristics for the sensors used in the satellite CDR.

Sensor	Satellite	Equator Overpass Time	Channels (μm)	CDR time period
ATSR-2	ERS-2	10:30 and 22:30	0.55, 0.67, 0.87, 1.61, 3.7, 10.85, 12.0	1995–2002
AATSR	EnviSat	10:00 and 22:00	0.55, 0.67, 0.87, 1.61, 3.7, 10.85, 12.0	2002–2012
MODIS	Terra	10:30 and 22:30	0.42, 0.44, 0.49, 0.53, 0.56, 0.65, 0.68, 0.75, 0.87, 0.91, 0.936, 0.94, 1.38, 3.75, 3.96, 4.05, 4.47, 4.52, 6.72, 7.33, 8.55, 9.73, 11.03, 12.02, 13.34, 13.64, 13.94, 14.24	2012–2016
SLSTR	Sentinel 3A	10:00 and 22:00	0.56, 0.66, 0.87, 1.38, 1.6, 2.25, 3.74, 10.85, 12.0	2016–2020

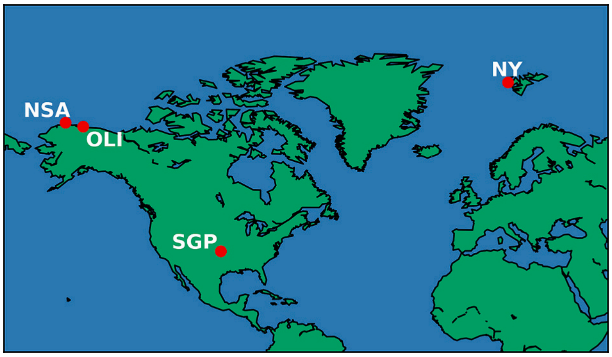

Fig. 2. Site locations of the four ceilometers: Ny Alesund (NY), North Slope of Alaska (NSA), Oliktok Point (OLI) and Southern Great Plains (SGP).

2017), the North Slope of Alaska and Oliktok Point (both on the Arctic coastline) (Morris et al., 1996) and the Southern Great Plains in Oklahoma (Morris et al., 1996). The primary characteristics of the ceilometer locations and measurements are given in Table 2 and Fig. 2. These four sites were chosen as they meet the following two criteria: a) dataset length sufficient to overlap with two or more of the satellite sensors used in the CDR, and b): attenuation height of at least 13 km, sufficient to detect cirrus cloud (many ceilometers have attenuation heights of ~8 km so do not meet this criterion). Throughout the remainder of this paper, the sites will be referred to by their short names as given in Table 2: NY, NSA, OLI and SGP.

Table 2
Characteristics of the four ceilometer sites used to determine cloud base height.

Location	Facility	Latitude	Longitude	Attenuation Height	Dataset Length	Measurement Frequency	Site Characteristics
Ny Alesund (NY)	National Environment Research Council (NERC)	78.9 N	11.9 E	13th July 1998 to 24th Aug 2011 13 km, 15 km thereafter	1998–2016	1 min	Within the Arctic Circle. Snow cover is dominant year round, temperatures exceed freezing in summer months.
North Slope of Alaska (NSA)	Atmospheric Radiation Measurement (ARM)	71.3 N	156.6 W	15–20 km	2000–2019	30 s	Within the Arctic Circle. Snow cover in winter months. Tundra thaws in summer.
Oliktok Point (OLI)	ARM	70.5 N	149.9 W	20 km	2014–2019	30 s	Within the Arctic Circle. Snow cover in winter months. Tundra thaws in summer.
Southern Great Plains (SGP)	ARM	36.6 N	97.5 W	25 km	1997–2018	1 min	Located in an area of cattle pasture. Continental climate (warm summers and cool winters).

3.3. Satellite to in-situ data matching

The satellite data are matched to the in-situ observations with a maximum spatial separation of 1 km (commensurate with the resolution of the satellite data) and a maximum time difference of 30 s. The time difference is minimised to prevent cloud movement between satellite and in-situ data pairs giving genuinely different cloud detection results.

The satellite matchups that are extracted comprise 31 across-track pixels by 7 along-track pixels, centred on the location of the ceilometer instrument. This larger across-track data extraction is required, to take account of the viewing geometry differences between the satellite and ceilometer. The ceilometer instrument looks directly upwards in the vertical, whilst the satellite looks downward, with a satellite viewing zenith angle restricted to between 0 and 22 degrees. The satellite pixel matched to the ceilometer observation is the one having the geometry nearest to that shown in Fig. 3, i.e., that whose line-of-sight views cloud above the cloud-base detected by the ceilometer. This requires accounting for both cloud base height and satellite zenith angle. The uncertainty on the cloud base height estimation from the ceilometer data is +/− 5 m (Morris, 2016) and is negligible in this choice of pixel.

In the absence of cloud (clear-sky conditions as seen by the ceilometer) the reference height for collocating the satellite observation is set to 6 km. This mid-tropospheric height, where many clouds are located is chosen to maximise the overlapping volume sensed by the satellite and ceilometer viewing geometries in the clear-sky case. Also recorded in these cases is the length of time before and after the match-up time, during which the ceilometer records clear-sky conditions, referred to as the clear-sky history.

3.4. Cloud detection algorithms

Cloud detection stability is assessed using three different algorithms: 1) Bayesian cloud detection, 2) probabilistic cloud detection and 3) operational cloud detection algorithms. The reason for choosing these algorithms is as follows. The Bayesian cloud detection scheme has been designed to work for surface temperature retrievals and has a long history of application to sea surface temperature climate data records (Embury et al., 2024; Merchant et al., 2019). The probabilistic cloud detection algorithm is routinely applied to LST products from polar orbiting sensors within the ESA LST CCI project (Ghent et al., 2019; Ghent et al., 2017), including those used together to form CDRs. Finally, the operational cloud detection algorithms are the ones supplied directly with the satellite data and would be available to users applying their own LST retrieval schemes (Ackerman et al., 1998; Birks, 2007; European Space Agency, 2023; Závody et al., 2000).

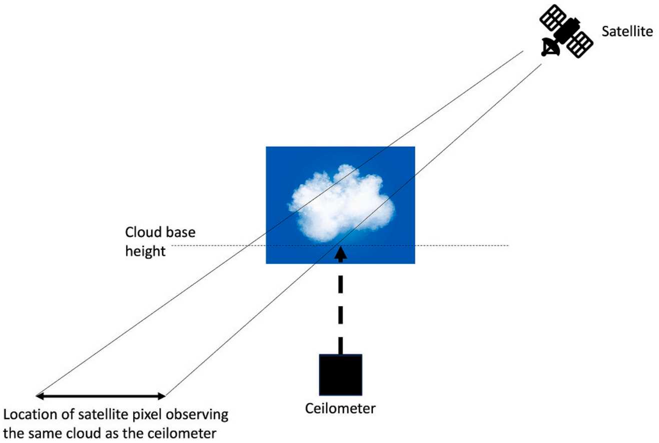
Satellite
Cloud base height
Ceilometer
Location of satellite pixel observing the same cloud as the ceilometer

Fig. 3. Schematic illustrating the possible across-track offset in the ground projection of the satellite pixel observing the same cloud feature as the ceilometer. Where a cloud base height is detected by the ceilometer, the best matched satellite pixel for cloud detection comparison purposes is the one which intersects with the atmospheric column immediately above this cloud base height.

3.4.1. Bayesian cloud detection (UoR)

A detailed description of the Bayesian cloud detection algorithm as applied over land is provided elsewhere (Bulgin et al., 2022; Merchant et al., 2005) so we provide only a brief overview here. Bayes' theorem can be applied to the problem of cloud detection to calculate the probability that a given observation is cloud-free. The probability of a clear-sky pixel $P(c|\boldsymbol{y}^o, \boldsymbol{x}^b)$, is conditional on prior information of the background state ($\boldsymbol{x}^b$) and the observation vector ($\boldsymbol{y}^o$) as shown by Eq. (1).

$$P(c|\boldsymbol{y}^o, \boldsymbol{x}^b) = \left[1 + \frac{P(\bar{c})P(\boldsymbol{y}^o|\boldsymbol{x}^b, \bar{c})}{P(c)P(\boldsymbol{y}^o|\boldsymbol{x}^b, c)}\right]^{-1} \tag{1}$$

The background state ($\boldsymbol{x}^b$) is a reduced state vector including skin temperature, total column water vapour and aerosol optical depth, constrained by hourly ERA5 numerical weather prediction (NWP) reanalysis data and aerosol optical depth from the Copernicus Atmosphere Monitoring Service (CAMS) (Bulgin et al., 2022; Hersbach et al., 2020; Inness et al., 2019). $P(\bar{c})$ and $P(c)$ are the prior probabilities of cloud and cloud-free conditions respectively as specified by the ERA5 cloud fraction, with $P(\bar{c})$ constrained between 0.5 and 0.95 (Bulgin et al., 2022). The observation vector ($\boldsymbol{y}^o$) is comprised of the satellite observations at infrared and, during the day, reflectance wavelengths. The 11 and 12 µm channels are always used, with the addition of the 3.7 µm channel at night and the 0.6, 0.8 and 1.6 µm channels during the day (Table 3). $P(\boldsymbol{y}^o|\boldsymbol{x}^b, c)$ is the clear-sky probability of the observations given background conditions and is simulated using a fast radiative transfer model (RTTOV 12.3, (Hocking et al., 2019)). $P(\boldsymbol{y}^o|\boldsymbol{x}^b, \bar{c})$ is the cloudy sky equivalent, pre-calculated and stored in the form of a look-up table due to the significant computation expense of simulating all possible cloudy sky conditions (Bulgin et al., 2022). A threshold of 0.5 is placed on the resultant clear-sky probability to generate a binary cloud mask, above which, pixels are considered clear. This algorithm was developed at the University of Reading and is referred to using the shorthand 'UoR' throughout this manuscript.

3.4.2. Probabilistic cloud detection (UoL)

The naïve probabilistic cloud detection algorithm compares a given satellite observation against a pre-calculated probability density function (PDF) of clear-sky conditions for the pixel location. These PDFs are generated from clear-sky simulations using a fast radiative transfer model (RTTOV11.2, (Hocking et al., 2015)) run at the European Centre for Medium Wave Forecasting (ECMWF) profile locations in the ERA-Interim reanalysis (Bulgin et al., 2014). Temporal interpolation is employed between the 6-hourly timesteps of ERA-Interim and spatial interpolation is bi-linear between ERA-Interim profile locations (Bulgin et al., 2014). The PDF is constructed by taking the mean simulated brightness temperature and the standard deviation of the observational climatology from the corresponding 5 × 5 degree grid cell, dependent on month, biome and day/night partitioning (Bulgin et al., 2014). The observational climatology is constructed using data from AATSR for 27

Table 3
Summary of wavelengths used for each cloud detection algorithm over land.

Algorithm	Channel wavelengths / µm	
	Day	Night
Bayesian	0.6, 0.8, 1.6, 11, 12	3.7, 11, 12
Probabilistic	11, 12	3.7, 11, 12
SADIST	0.5, 0.6, 0.8, 11, 12	3.7, 11, 12
Basic Cloud Mask	0.5, 0.6, 0.8, 1.4, 11, 12	3.7, 11, 12
MODIS Cloud Mask	0.6, 0.8, 1.4, 3.7, 3.9, 11, 12, 13.9	3.7, 3.9, 6.7, 7.3, 11, 12, 13.9

biomes, derived from the GlobCover land cover classification (Ghent et al., 2017). Pixels are determined to be cloudy if they fall outside of the 95 % limit of either of the two tests employed. At night the observations are tested against the simulated 12 μm temperature and 11–3.7 μm differences. During the day the observations are tested against the simulated 12 μm temperature and 11–12 μm differences (Table 3) (Bulgin et al., 2014). This algorithm was developed at the University of Leicester and is referred to using the shorthand 'UoL' throughout this manuscript.

3.4.3. Operational cloud detection (Oper)

The operational cloud masks are all based on a series of threshold tests. ATSR-2 and AATSR data are provided with the Synthesis of ATSR Data Into Sea-Surface Temperature (SADIST) cloud mask, which has been adapted for use over land (Birks, 20077; Závody et al., 2000). The tests employed are for gross cloud, thin cirrus, medium/high level cloud, fog/stratus and during the daytime only, a test based on the normalised difference vegetation index (NDVI) and a snow test. The channels employed in these tests are shown in Table 3. SLSTR follows on from ATSR-2 and AATSR in the ESA satellite instrument series (although there was a 4-year gap between the failure of AATSR in 2012 and the launch of SLSTR in 2016). The SLSTR cloud mask is therefore an evolution of the SADIST cloud mask, referred to as the 'basic cloud mask' in SLSTR products (European Space Agency, 2023). It employs the same series of tests as those used by SADIST, with the addition of an 11 μm spatial coherence test and a 1.375 μm threshold test (European Space Agency, 2023). In this study we use the summary cloud mask.

The MODIS operational cloud mask provides a gradated confidence, classifying each pixel as either cloudy, probably cloud, probably clear or confidently clear (Ackerman et al., 2010). For the purpose of this study, we implement a binary mask, which considers all 'cloudy' and 'probably cloud' pixels to be cloud and the remainder clear-sky. MODIS tests include identification of thin cirrus, low-level water clouds, high and mid-level clouds, and surface/low clouds. During the day additional reflectance tests are used and at night a comparison is made between the observations and expected clear-sky surface temperatures (Ackerman et al., 2010). The channels used in the cloud-masking algorithm are summarised in Table 3. These series of algorithms are referred to by the shorthand 'Oper' throughout this manuscript.

3.5. LST retrieval

A LST retrieval was made for all observations within the match-up database (irrespective of the cloud masking outcome of any given algorithm), facilitating comparison of LST values as would be retrieved when applying different cloud-clearing algorithms. The Leicester ATSR and SLSTR Processor for Land Surface Temperature (LASPLAST) (Ghent et al., 2017) algorithm is applied to all sensors within the CDR. Full details of this split-window retrieval algorithm are provided in (Ghent et al., 2017).

4. Metrics and definitions

We define in this section the series of metrics and definitions that we use throughout this paper. The results section refers to these definitions as appropriate to save repetition throughout the manuscript.

4.1. Match-up data comparisons

Match-up data at four in-situ locations are evaluated in this paper. Ny Alesund (Svalbard, Europe), North Slope of Alaska (Alaska, North America) and Oliktok Point (Alaska, North America) are all high-latitude sites in the northern hemisphere, whilst the Southern Great Plains (Oklahoma, North America) is mid-latitude site. The temporal frequency of satellite to in-situ matches varies by site. At higher latitudes, match-up frequency reaches 90-min time intervals where consecutive satellite overpasses view the same location. In the mid-latitudes overpass frequency occurs a maximum of twice a day with intervals up to 2–3 days in-between, depending on the swath width and return time of the satellite. We therefore analyse the data at quarter-year intervals: January–March (JFM), April–June (AMJ), July–September (JAS) and October–December (OND) and ensure that any statistics calculated are based on a minimum of 20 matchups in each season.

4.2. Cloud performance metrics

For each ceilometer-satellite match-up we know the state of the ceilometer path (clear/cloud), the length of clear-sky history before and after the ceilometer-satellite match (given by the ceilometer) and the satellite path's classification (clear/cloud) according to each of the cloud detection algorithms (Fig. 4). The true state of the satellite path is unknown and may differ from the ceilometer path due to the difference in viewing geometry and spatial footprint of the observation (see Section 3.3).

The ceilometer-satellite matches can be split into three groups: a) clear-sky ceilometer data with at least 90 s of clear-sky observations before and after the time of the match (*S*), b) clear-sky ceilometer data with fewer than 90 s of clear-sky observations before and/or after the time of the match (*T*) and c) cloudy ceilometer data (*U*). Choosing a value of +/− 90 s for the clear-sky history ensures that at a minimum the ceilometer observation prior to and following the matched data were both clear. For ceilometers with an observation frequency greater than every minute, this window would encompass more observations.

For each of these classes of observations (*S*, *T* and *U*), the fraction of observations where the satellite path is truly clear or the satellite path is truly cloud will sum to one, but the fractions themselves are unknown (represented by letters *a-f*, Fig. 4). In each case (*a-f*), the cloud detection algorithms will then classify the satellite path as either clear or cloudy (each pair of fractions again sums to one, represented by the letters *G-R*, Fig. 4).

Using the variables in this contingency table (Fig. 4) we can formulate two metrics of interest: 1) the fraction of clear-sky LST classifications that is likely to be contaminated by cloud (CC) and 2) the fraction of clear-sky pixels that are misclassified as cloud (MC).

$$CC = \frac{SbI + TdM + UfQ}{S(aG + bI) + T(cK + dM) + U(eO + fQ)} \tag{2}$$

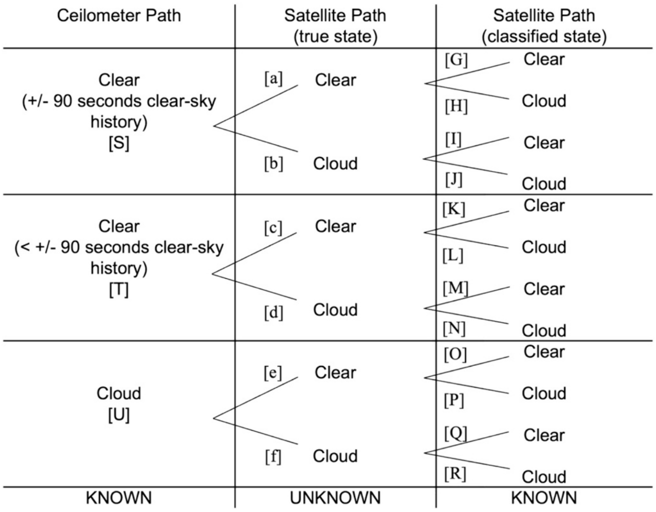

Fig. 4. Schematic illustrating the relationship between the state of the ceilometer path, the true state of the satellite path and the classified state of the satellite path.

$$MC = \frac{SaH + TcL + UeP}{S(aH + bJ) + T(cL + dN) + U(eP + fR)} \tag{3}$$

We assume that the underlying probabilities of truly clear-sky pixels being classified as clear or misclassified as cloudy by a given cloud detection algorithm are consistent and independent of the reported state from the ceilometer. These assumptions imply the following: $G = K = O$; $H = L = P$; $I = M = Q$ and $J = N = R$. For cloudy observations, we also match the satellite path directly to the height of the cloud base observed for the ceilometer, which gives a high degree of certainty that a truly cloudy satellite path was viewed, so we can assume that $e = 0$ and therefore $f = 1$.

Considering first the calculation of CC (Eq. (2)), we can define the following quantities where V is the number of all clear-sky ceilometer observations with a long clear-sky history, classified as clear-sky by the given cloud detection algorithm. W is the equivalent metric for the clear-sky ceilometer observations with a shorter clear-sky history. X is the number of cloudy ceilometer observations where the satellite path is classified as clear-sky.

$$V = S(aG + bI) \tag{4}$$

$$W = T(cK + dM) \tag{5}$$

$$X = UQ \tag{6}$$

Substituting these quantities into Eq. (2) gives:

$$CC = \frac{(Sb + Td)\,{}^{X}\!/_{U} + X}{V + W + X} \tag{7}$$

Here, all variables are known from the match-up datasets except for b and d, which cannot be directly observed. However, we can assess the sensitivity of CC and MC to plausible values. By choosing upper (worst case), lower (best case) and most-likely values for these variables, we can calculate a sensitivity of the CC metric to varying amounts of cloud contamination.

The best-case scenario is that the ceilometer result always accurately reflects the true state of the satellite path, in which case $b = d = 0$ %. The reasons to expect this not to be the case are 1) the larger atmospheric column observed by the satellite (1 km pixels at nadir) in comparison with the ceilometer, 2) cloud movement between the times of the ceilometer and satellite observations and 3) inconsistencies in the atmospheric path observed by the ceilometer and the satellite due to differences in viewing geometry.

The likelihood of inconsistency in the atmospheric path increases with satellite viewing zenith angle and can be represented by the distance of the matched satellite pixel as projected onto the ground, from the satellite pixel containing the ceilometer (pixel shift). This ranges between 0 and 2 for the viewing geometries considered in this paper. The values chosen for b and d are therefore dependent on this difference.

Values for b and d are shown in Table 4. The fraction of clear-sky ceilometer cases where the true state of the satellite path is cloudy (b) is 0 % in the best case for all pixel shifts. As b is a fraction of S, where we have a clear-sky history of +/− 90 s from the time of the match, we assume that in all cases where there is no shift, the likelihood of cloud contamination in the satellite path is 0 %. As the pixel shift increases and the ceilometer-satellite paths diverge, the likelihood of cloud contamination increases. We use values of 2 % and 5 % (most likely and worst case) for a 1-pixel shift and 5 % and 10 % respectively for a 2-pixel shift, to test the sensitivity of CC to b.

d is the equivalent fraction to b, but applied to the set of matches in T, where the clear-sky history for each match is shorter than +/− 90 s. In this case, the chances of cloud contamination are greatly increased. In the best case, we again assume 0 % for each pixel shift. We then test the sensitivity by defining d as 5 % for the no-shift most-likely case, 20 % for the 1-pixel shift and 50 % for the 2-pixel shift 50 %. In the worst-case scenario, we use values of 10, 50 and 80 % respectively.

The percentage of pixels with different ground-pixel shifts is location specific (Table 5). We use this information along with the range of values given in Table 4 for the sensitivity analysis to calculate location specific values for b and d for each case (best, most likely and worst). For example, to calculate d in the most likely case for Ny Alesund we use: $d = (19{*}0.05) + (41{*}0.2) + (40{*}0.5) = 29.2\%$. The location specific split in the percentage of clear-sky observations that fall into groups S and T is accounted for by independently multiplying the total number of observations in each group by b or d.

Returning to Eq. (3), we can define three further metrics from the data we have available. α is the number of clear-sky observations in group S that are classified as cloud by the given cloud detection algorithm. β is the equivalent metric for group T. γ is the number of cloudy ceilometer observations also classified as cloud by the given algorithm.

$$\alpha = S(aH + bJ) \tag{8}$$

$$\beta = T(cL + dN) \tag{9}$$

$$\gamma = UR \tag{10}$$

Substituting these values into Eq. (3) gives Eq. (11). a and c can be calculated as $(1 - b)$ and $(1 - d)$ respectively.

$$MC = \frac{H(Sa + Tc)}{\alpha + \beta + \gamma} \tag{11}$$

Finally, we solve for H using previously defined parameters:

$$S = SaG + SaH + SbI + SbJ \tag{12}$$

Substituting in Eqs. (4 and 10), then dividing by S gives:

$$1 = aH + b\,{}^{\gamma}\!/_{U} + {}^{V}\!/_{S} \tag{13}$$

$$H = \frac{1 - b\,{}^{\gamma}\!/_{U} - {}^{V}\!/_{S}}{a} \tag{14}$$

H is therefore calculated on a location and algorithm-specific basis.

4.3. Monte-Carlo Theil-Sen slopes

To calculate timeseries trends, we follow an updated version of the methodology used by (Good et al., 2022), when calculating LST temporal stability with reference to two-metre air temperature data. The Theil-Sen regression calculates pairwise slopes between each data point

Table 4
Sensitivity analysis for calculating b and d dependent on ground-shift in matched satellite pixel.

b	No shift	1-pixel shift	2-pixel shift
Best case	0 %	0 %	0 %
Most likely (realistic case)	0 %	2 %	5 %
Worst case	0 %	5 %	10 %

d	No shift	1-pixel shift	2-pixel shift
Best case	0 %	0 %	0 %
Most likely (realistic case)	5 %	20 %	50 %
Worst case	10 %	50 %	80 %

Table 5
Percentage of matches with no, one and two ground-pixel shifts in the matching of the ceilometer and satellite observations.

Location	No Shift	1-Pixel Shift	2-Pixel Shift
NY	19 %	41 %	40 %
NSA	18 %	46 %	36 %
OLI	9 %	53 %	38 %
SGP	21 %	50 %	29 %

in the timeseries and takes the median of these slopes as the slope estimator, making it less sensitive to outliers (Sen, 1968). One limitation of this approach as used by (Good et al., 2022) is that it doesn't account for the uncertainty in the input data. We therefore adopt a Monte-Carlo Theil-Sen approach, whereby we calculate the Theil-Sen median over 10,000 iterations (Metropolis and Ulam, 1949; Sen, 1968). We seed the Monte-Carlo iteration with the original datapoints and sample randomly within the upper and lower bounds (best and worst case scenarios) for each observation thereafter. From the resultant distribution of 10,000 median slopes, we take the mean and two sigma to describe the slope and variability. If the two sigma bounds on the slope encompass a slope of zero, we consider the timeseries to be stable.

5. Results

5.1. Match-up data characteristics

Fig. 5 (panels a-d) shows the quarterly mean cloud fraction, calculated independently for each sensor using all matches in each timeseries. A sensor is included in the analysis where there are three or more years of match-ups available. The data record at OLI is shorter and only MODIS-T and SLSTR-A have more than three years of ceilometer matches. Error bars indicate the standard error on the calculation of the mean. Note that the scales on the y-axis of these figures are different for each location. NY has the highest cloud fraction (ranging between 0.68 and 0.81 for AATSR and MODIS), peaking in JAS, with a minimum in JFM. NSA and OLI are both located in northern Alaska but have quite different annual cycles in cloud fraction. For NSA, the cloud fraction is lowest in AMJ (0.22–0.35) and highest in OND (0.45–0.49) for AATSR, MODIS and SLSTR (but less variable for ATSR-2). At OLI, the cloud fraction is consistent year-round for MODIS (0.36–0.41). The SLSTR cloud fraction is more variable, but the data also have a higher uncertainty. In SGP, cloud fraction is at a minimum in JAS and a maximum in JFM, ranging between 0.38 and 0.66.

The seasonal LST climatology is plotted in Fig. 5 (panels e-h) for all, day and nighttime observations. To calculate the climatology, we use all matches in group S, with a no ground pixel shift between the satellite and ceilometer (Section 4.2), where we are confident that the ceilometer clear-sky observations are also clear for the satellite path. Nighttime is defined as all solar zenith angles greater than 85 degrees. All sites have an LST minimum in JFM and a maximum in JAS as they are all located in the northern hemisphere. NY, NSA and OLI are all high latitude sites with similar LST ranges. Daytime minimums are 254.5, 248.5 and 248.7 K for each site respectively (occurring in JFM) and daytime maximums are 275.4, 273.3 and 275.1 K (in JAS). Nighttime minimums in JFM are 252.2, 244.6 and 247.5 K; 1.2–3.9 K cooler than the daytime minimums. Nighttime maximums in JAS are 267.4, 269.1 and 264 K. The largest temperature difference between day and night occurs in AMJ. In SGP, increased daytime solar heating increases the difference between daytime and nighttime LST. Nighttime LST has a minimum of 275.1 K in JFM and a maximum of 296.7 K. Daytime equivalents are 287.6 and 313.9 K.

Fig. 6 shows the CC and MC performance metrics (defined in section 4.2). The solid line shows the most likely case with the shading representing the range between the best and worst case for each metric. In NY, UoR has the lowest fraction of algorithm-defined clear-sky pixels contaminated by cloud, ranging between 0.36 and 0.53, with a minimum in AMJ. CC is higher for both the UoL and Oper algorithms (0.48–0.66). The possible range of CC values is small for all seasons and algorithms (00.02–0.04). In NSA the most likely CC value is similar for all algorithms in OND (0.43–0.46), diverging across the other seasons with a minimum for Oper in AMJ and a minimum for UoR and UoL in JAS. The range in values between the best and worst case is also a bit larger (0.03–0.06). OLI has similar values of CC to NSA, with lower fractions for Oper (0.24–0.31) and UoR (0.27–0.35), than UoL (0.28–0.43). In SGP, CC follows the same seasonal pattern for all algorithms with a minimum in JAS (0.16–0.21) and a maximum in JFM for Oper and UoL (0.3–0.39), and OND for UoR (0.29). The possible range in CC values in OLI and SGP is the lowest of all sites considered (0.01–0.03).

MC typically has an inverse relationship to CC. In NY, MC is lowest in JAS for all algorithms (0.08–0.15), peaking in JFM (0.24–0.31). The range of possible CC values varies between 0.02 and 0.05 for all algorithms. In NSA and OLI, MC is much larger than for NY and SGP. In OLI,

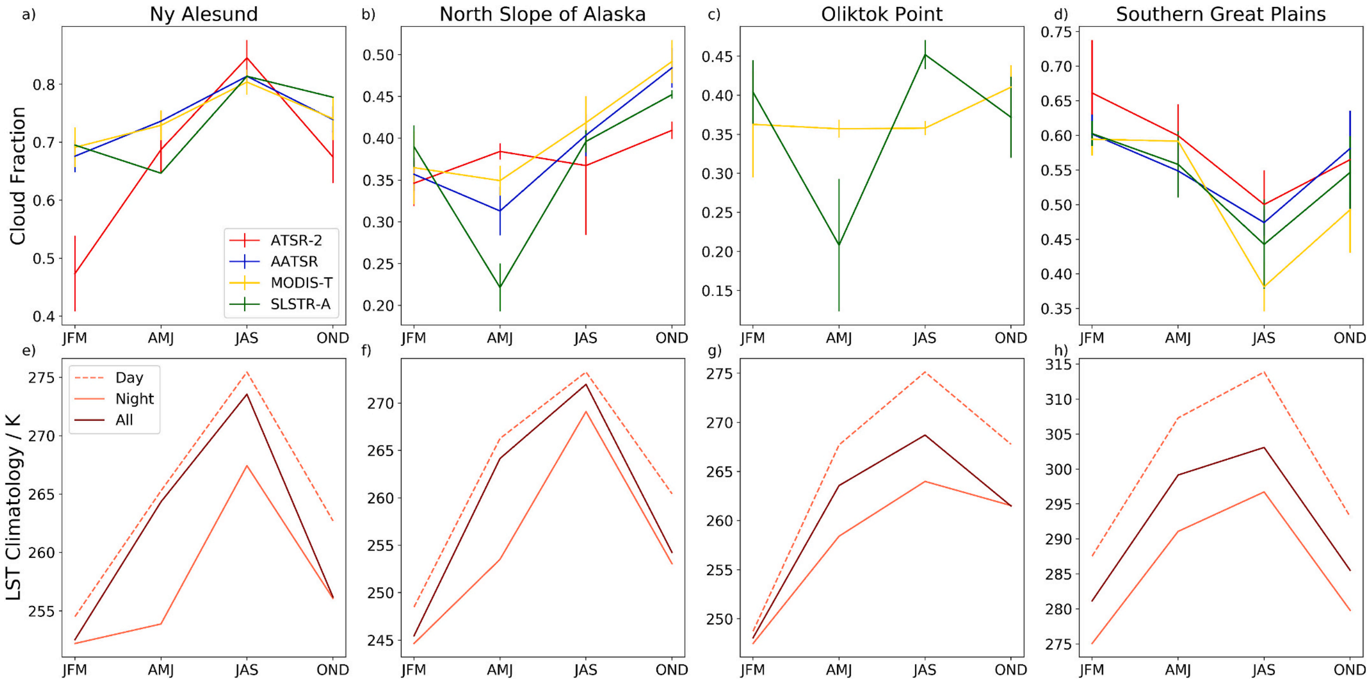

Fig. 5. Match-up characteristics for the four in-situ ceilometer sites (NY, NSA, OLI and SGP from left to right). Variables shown are the quarterly cloud fraction for each satellite sensor (top) and the climatological LST for each site (bottom).

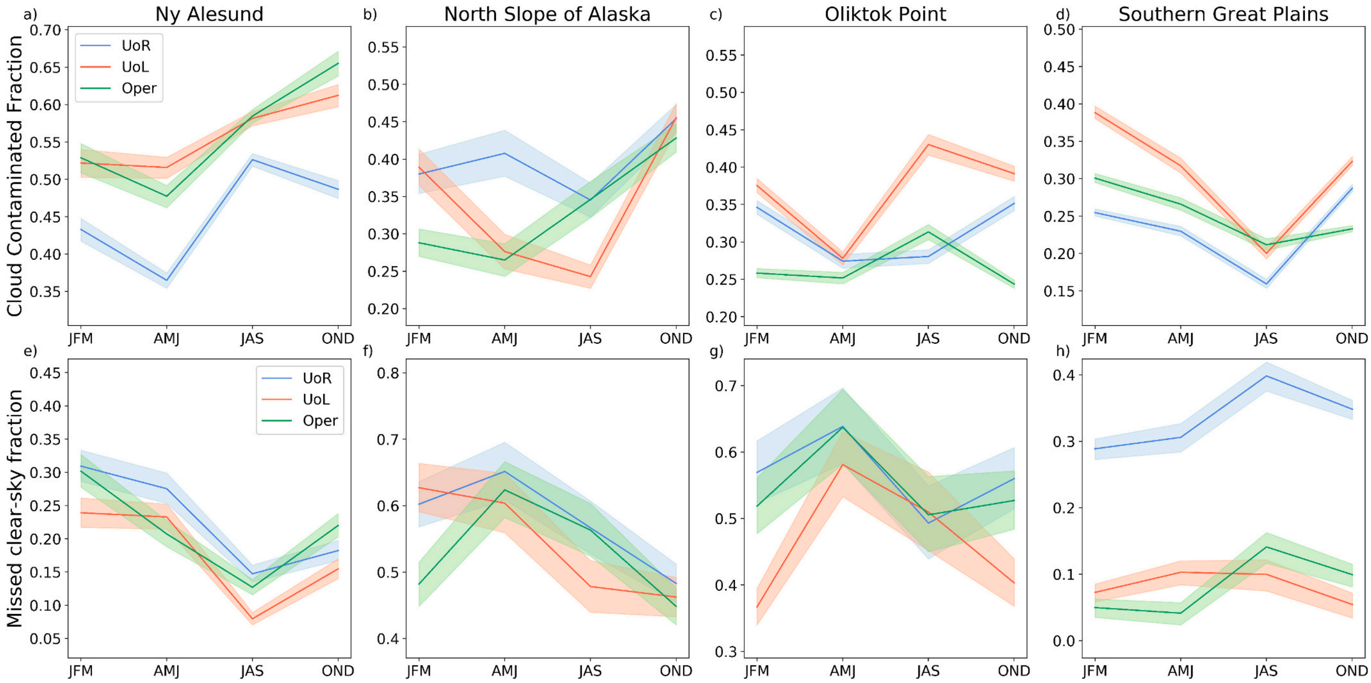

Fig. 6. Cloud detection performance for the ceilometer-satellite matches (NY, NSA, OLI and SGP from left to right). Variables shown are the fraction of clear-sky data contaminated by cloud (CC) for each cloud detection algorithm (top) and the fraction of clear-sky data erroneously screened as cloud (MC) (bottom) by each of the cloud detection algorithms. In all cases, the most likely case is represented by the solid lines, with shading representing the possible range of values (best case to worst case). For more information on how this range is defined, please refer to Section 4.2.

a larger fraction of the clear-sky observations has a short clear-sky history (group T, 29 %), reflected in the greater possible range of values for MC (0.05–0.12). In both NSA and OLI, MC peaks in AMJ (0.6–0.65 for NSA, 0.58–0.64 for OLI). The season with the minimum MC is algorithm and location specific. The lowest MC fractions for any site are found for the UoL and Oper algorithms in SGP, with a minimum in JFM (0.05–0.07) and a maximum in JAS (0.1–0.14). The UoR algorithm performs less well here, with MC ranging between 0.29 and 0.4.

5.2. Cloud detection metric timeseries

Fig. 7 shows timeseries plots of the annual CC and MC metrics (most likely values) for each ceilometer site. Data are only included where every season is represented in the annual average. This is important in the calculation of stability in the annual mean performance metrics as they show considerable seasonal variability (Fig. 6). Where full seasonal representation is unavailable due to gaps in the ceilometer or satellite data record (early 2000's and 2010 for some sites), these years are omitted.

Overall performance is algorithm, location and sensor specific. Consistent with the seasonal analysis in Fig. 6, CC is largest in NY, ranging between 0.34 and 0.63 for UoR, 0.44–0.68 for UoL and 0.35–0.71 for Oper. Some temporal variability is evident in CC for all algorithms, with lower values at the beginning of the AATSR part of the CDR and then again at the end of the AATSR record and first year of MODIS. In the next section we discuss whether this variability could be related to external factors.

In NSA, there is less coherent temporal variability in CC. CC is most variable for UoR, with peaks in 2005, 2006, and 2015. The overall CC rates are lower than for NY: 0.24–0.55 for UoR, 0.28–0.44 for UoL and 0.27–0.44 for Oper. In OLI, Oper has the lowest CC values (0.21–0.24) and there is a step-change between MODIS and SLSTR for the UoL algorithm. CC values for MODIS data are 0.28–0.39 and for SLSTR data, 0.4–0.46. In SGP, CC is the lowest of the four sites analysed, but there is significant interannual variability with a range of 0.34 for UoR, 0.2 for UoL and 0.27 for Oper.

For MC, there is a step-change for all algorithms for NY, where MC is high during the ATSR-2 data record and then lower thereafter. MC is highest in NSA and OLI. For NSA, MC ranges between 0.5 and 0.64 for UoR, 0.39–0.63 for UoL and 0.41–0.63 for Oper. The largest MC values are seen in 2003 and 2016–2019. In OLI there is evident upward trend in MC for UoL rising from 0.33 in 2014 to 0.46 in 2019. MC is lowest for SGP with a significant difference between UoL and Oper (0.0–0.26 and 0.0–0.15 respectively) and UoR (0.25–0.41). MC rises for UoL during the SLSTR data record.

5.3. Stability in cloud-detection performance metrics

The first step in assessing whether the cloud clearing algorithm employed prior to the LST retrieval has the potential to generate non-geophysical trends in the LST data is to assess the temporal stability of the performance metrics, CC and MC. We do this using the Monte-Carlo Theil-Sen methods described in Section 4.3. The results are provided in Table **6**. We assess the stability of the performance metrics for each sensor individually, where we have at least three years of matches, and then for the full timeseries. The slope and confidence interval values provided in Table 6 are for the Monte-Carlo Theil-Sen fit to the whole timeseries.

Considering first CC, only two timeseries meet the stability criteria: UoR at NSA and SGP. In these cases where CC is considered stable for the entire timeseries of matches, each sensor taken on an individual basis would be considered unstable with the exception of MODIS for SGP. With the exception of ATSR-2 at NSA for UoL and Oper and MODIS for UoR at SGP, all sensor-specific results for other site-algorithm comparisons also fail to meet the stability criterion (shown by the red colours in Table 6).

For MC, timeseries stability is achieved for UoR at NSA, OLI and SGP, and for Oper at OLI and SGP. A greater number of individual sensor records are also stable across all cloud detection algorithms. The only data record where all sensors and the timeseries are stable is for UoR at OLI. In some cases, all contributing sensors can be stable e.g. Oper at NY, but the timeseries as a whole is unstable.

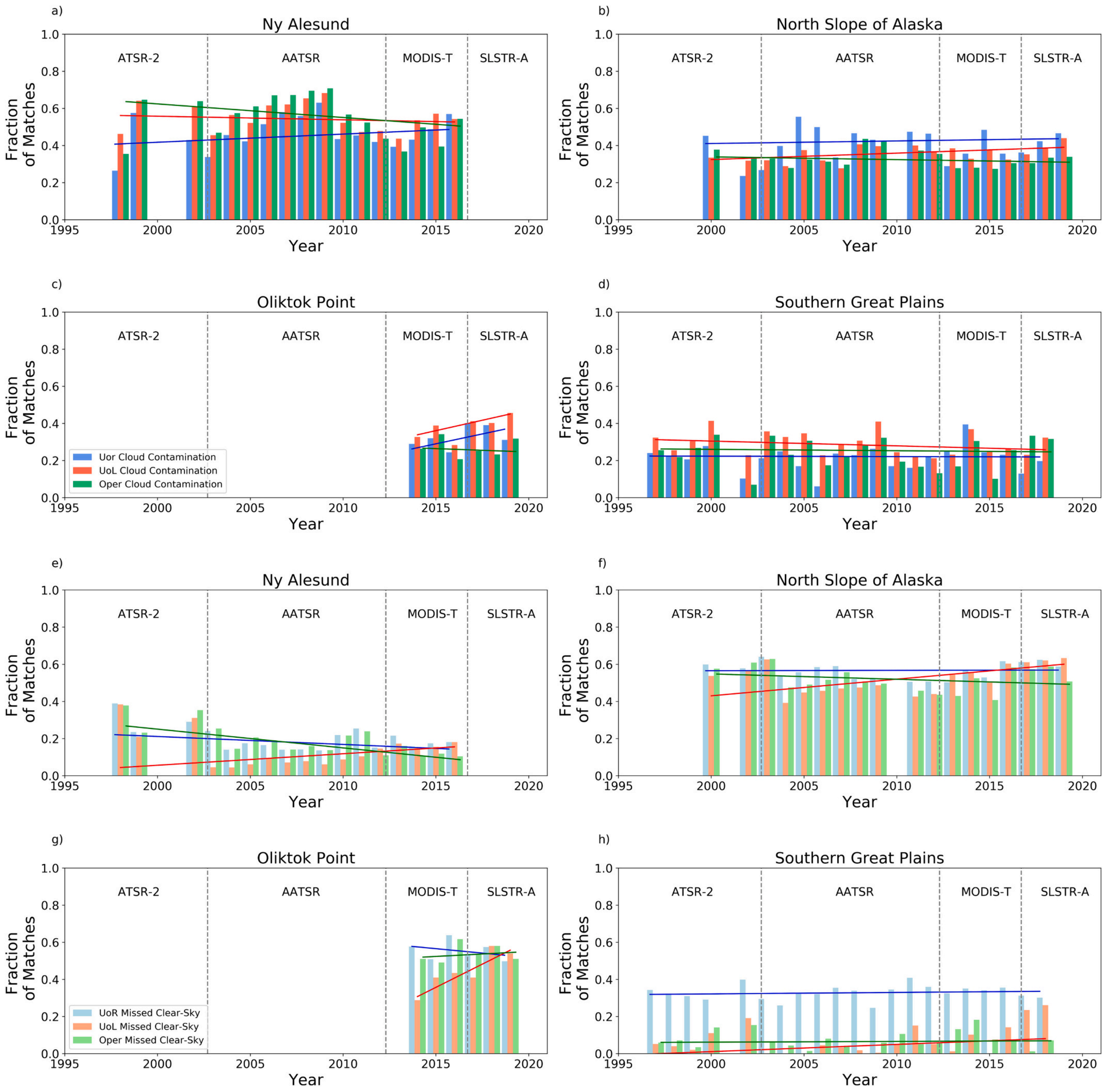

Fig. 7. Annual CC (a-d) and MC (e-h) metrics for the UoR (blue), UoL (red) and Oper (green) cloud detection algorithms at each of the ceilometer locations. Trend lines represent the Monte-Carlo Theil-Sen mean timeseries slopes as described in Section 4.3. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.4. External factors

Cloud mask performance can be affected by external factors such as viewing geometry, cloud type and the total cloud-fraction, which we must first consider before drawing conclusions on the impact of the performance metric stability on the LST timeseries. Fig. 8 shows the relationship between the CC and MC metrics for each algorithm and location with respect to satellite zenith angle, solar zenith angle and the lowest cloud base height. The satellite zenith angle of all matchups is restricted to 0–22 degrees for consistency with the narrower swath width of the ATSR-2 and AATSR instruments. No dependence on satellite zenith angle within this range is apparent for any of the metrics at any of the ceilometer locations.

The solar zenith angle range is site specific. NY is the most northerly site, where the solar zenith angle minimum is 50–60 degrees. NSA and OLI are slightly further south, so solar zenith angles between 40 and 50 degrees also occur in the matchups. For SGP, daytime matches occur with solar zenith angles between 10 and 70 degrees. Twilight conditions are not seen due to the latitude of the ceilometer site and the equatorial overpass time of the satellites. All night-time observations are plotted to the right-hand side of the grey dashed line in Fig. 8 panels e-h. At SGP, UoR CC is lower at night than during the day and UoL MC is higher at night than during the day. At NY, MC increases under twilight conditions for all algorithms.

Table 6
Stability assessment of performance metrics CC and MCS, by sensor and for the entire timeseries. Monte Carlo Theil-Sen slopes and confidence intervals (as described in Section 4.3) are presented. The stability column indicates the stability for the individual sensors (ATSR-2, AATSR, MODIS and SLSTR represented by 1, 2, 3 and 4 respectively) and all (A) sensors in the timeseries. Analysis is only performed for a given sensor where 3+ years of matchups are available.

		Slope (S), Lower Confidence Interval (LCI), Upper Confidence Interval (UCI)															
		Cloud contaminated data			Stable?					Missed clear-sky data			Stable?				
Location	Algorithm	S	LCI	UCI	1	2	3	4	A	S	LCI	UCI	1	2	3	4	A
NY	UoR	0.0044	0.0033	0.0055	🟥	🟥	🟥		🟥	-0.0043	-0.0058	-0.0028	🟥	🟩	🟥		🟥
	UoL	-0.0019	-0.003	-0.0008	🟥	🟥	🟥		🟥	0.0062	0.0042	0.0082	🟥	🟥	🟥		🟥
	Oper	-0.0073	-0.0087	-0.0059	🟥	🟥	🟥		🟥	-0.01	-0.011	-0.0088	🟩	🟩	🟩		🟥
NSA	UoR	0.0014	-0.0005	0.0033	🟥	🟥	🟥	🟥	🟩	0.0002	-0.0019	0.0022	🟩	🟥	🟥	🟩	🟩
	UoL	0.0035	0.0023	0.0046	🟩	🟥	🟥	🟥	🟥	0.009	0.0059	0.012	🟩	🟩	🟥	🟩	🟥
	Oper	-0.0015	-0.0027	-0.0003	🟩	🟥	🟥	🟥	🟥	-0.0029	-0.005	-0.0008	🟩	🟩	🟥	🟥	🟥
OLI	UoR	0.021	0.017	0.025			🟥	🟥	🟥	-0.0096	-0.025	0.0057			🟩	🟩	🟩
	UoL	0.022	0.017	0.028			🟥	🟥	🟥	0.05	0.036	0.064			🟥	🟥	🟥
	Oper	-0.0036	-0.0069	-0.0004			🟥	🟥	🟥	0.0053	-0.0094	0.02			🟥	🟩	🟩
SGP	UoR	-0.0002	-0.006	0.0001	🟥	🟥	🟩		🟩	0.0008	-0.00002	0.0016	🟩	🟥	🟩		🟩
	UoL	-0.0026	-0.0031	-0.002	🟥	🟥	🟥		🟥	0.0039	0.0028	0.005	🟥	🟥	🟥		🟥
	Oper	-0.0007	-0.0012	-0.0002	🟥	🟥	🟥		🟥	0.0004	-0.0003	0.0012	🟥	🟩	🟩		🟩

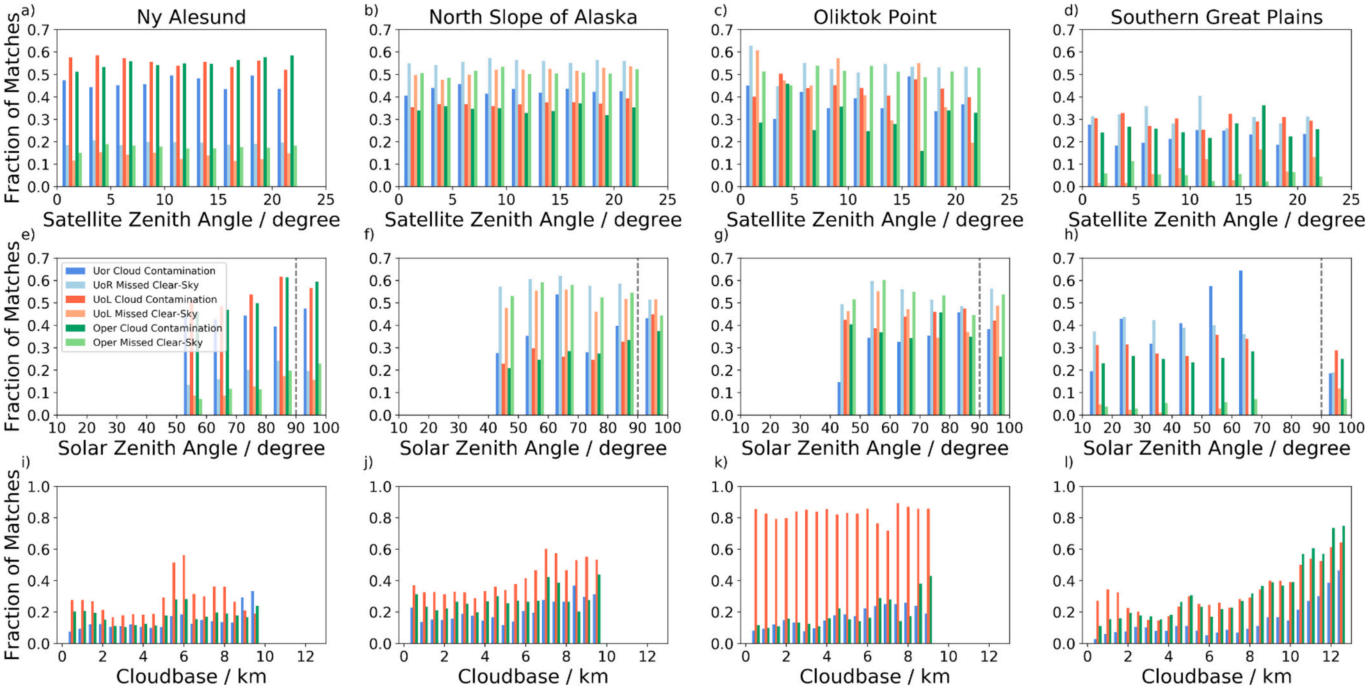

Fig. 8. External factors affecting CC and MC metrics for each of the four ceilometer sites (from left to right: NY, NSA, OLI and SGP). Metrics are plotted as a function of satellite zenith angle (top), solar zenith angle (middle) and the lowest cloud base (bottom). For the solar zenith angle plots (e-h), nighttime matches are all plotted in the 90–100-degree bin to the right of the grey dashed line.

The cloud base range is location dependent, with a maximum of 9 km at the polar locations and 13 km in SGP. This is consistent with the latitudinal variation in tropopause height, which is lower at the poles than at the equator. Except for UoL at OLI (which shows poor performance for all cloud base heights), the cloud contaminated fraction of observations typically increases with cloud base height. Intuitively this makes sense as the cloud optical thickness of high-level cirrus cloud is typically lower than cumulus or stratus cloud features, reducing the cloud signal in the observations.

Considering temporal evolution of the external factors that influence cloud detection; the range of solar zenith angles in matchups at each ceilometer site should be consistent given no orbital drift in the satellites used to generate the CDR. Cloud type and amount has changed over time (Mao et al., 2019; Norris, 2005) so we must account for this in our stability analysis. Cloud fraction varies on the two-to-three-year timescale at all sites, with a slightly decreasing (but highly variable) trend with time apparent for SGP (not shown).

Cloud fraction (CF) and CC are positively correlated for all three algorithms at NY at the 95 % confidence interval with r values of 0.79, 0.66 and 0.55 for UoR, UoL and Oper respectively. CF and CC are also positively correlated for UoR at NSA ($r = 0.62$), for Oper at OLI ($r = 0.86$) and for UoL and Oper at SGP ($r = 0.78$ and $r = 0.66$ respectively). We can account for the CF variability in the stability analysis by multiplying CC by 1-CF, prior to calculating the Monte-Carlo Theil-Sen slopes. This has no impact on the stability of the results presented in Table 6 and we therefore conclude that temporal variability in CF is not a major contributor to instability in the CC metric.

5.5. LST impacts

Having established that the cloud mask performance metrics can be unstable for reasons that are not obviously linked to external controlling factors (section 5.4) it is important to consider the impact of these instabilities on the LST timeseries. To do this we calculate the 'true' LST for each in-situ site and seasonal average, taking clear-sky observations from the 'S' group only (where cloud contamination is less likely in the satellite path), limiting matches to those where the satellite ground pixel was no more than one removed from the ground pixel containing the ceilometer. These constraints allow a 2 % (most likely) contamination by cloud (Table 4) for the matches with one ground pixel shift. This is considered an acceptable trade off to increase the number of satellite-ceilometer matches available for the analysis. For clear-sky data, using a mid-tropospheric matching height of 6 km (Section 3.3) results in fewer matches with no ground pixel shift, as the satellite viewing geometry needs to be close to nadir for this occur. For cloudy matches, a cloud base below 6 km increases the occurrence of pixels with no ground shift across a wider range of off-nadir satellite viewing angles. In each case we calculate the anomaly by subtracting the seasonal climatology in each location, calculated using the same constraints on clear-sky pixels as the LST calculated from the satellite observations. In this analysis, we do not the omit years where some seasons are underrepresented due to gaps in the data record as we are now focusing on LST anomalies.

The same process is undertaken to calculate the anomalies in the 'algorithm defined' LST. This consists of LST from the pixels correctly identified as clear-sky by the satellite cloud detection algorithm and the cloudy pixels that have been falsely flagged as clear-sky. Stability is assessed by subtracting the timeseries of annual average anomalies for the true LST time series, from the algorithm-specific time series and fitting a linear trend to the result. The results are shown in Fig. 9 and Table 7. OLI is omitted due to the relatively short duration of the data record (8 years) compared to the GCOS definition of LST stability, measured in kelvin per decade.

Based on the relatively short timeseries of matches available, the UoL and Oper algorithms are stable in NSA and the UoR and Oper algorithms in SGP, to within the GCOS threshold requirement of 0.3 K per decade. However, some caution must be applied here in the interpretation of these results due to the large interannual variability (of order 2–4 K) in the anomaly difference timeseries and the inevitable dependence of the stability on dataset length (max 22 years). These timeseries also indicate a site-specific bias in the LST timeseries dependent on the cloud screening algorithm used. In NY, this is of order 2 K for UoL and 4 K for UoR. In NSA the bias is larger for UoL (~3.5 K) and for both sites close to 0 K for Oper. In SGP, UoR has a clear cold bias of ~5 K.

Table 7
LST stability in kelvin per decade arising from cloud detection instability, for the UoR, UoL and Oper algorithms at the NY, NSA and SGP ceilometer locations.

	NY	NSA	SGP
UoR	0.4	−0.73	0.21
UoL	0.38	0.18	0.5
Oper	0.36	0.01	0.29

5.6. Characterising CC and MC observations

LST stability requires a balancing of the two mechanisms by which cloud detection can fail: 1) the omitted clear-sky data due to over-screening and 2) the cloud contaminated pixels missed in the cloud screening process. The approach taken in constructing these matches allows us to look at the characteristics of both the missed clear-sky observations and the unscreened cloud for each of the evaluated algorithms. We do this first for the missed clear-sky observations, selecting the ceilometer clear-sky observations in group S, limiting these matches to where the ground pixel for the satellite path contains the ceilometer location (no shift), where we are confident that the cloud contamination likelihood is 0 % under these conditions. We plot the PDF of the difference between the LST in each pixel that is wrongly flagged as cloud minus the seasonal LST climatology (Fig. 10). Each location, algorithm and season is considered independently. For the cloud contamination equivalent plot, we take all matches in group U and identify those that are not correctly flagged by each of the evaluated algorithms. No limit is placed on the ground shift in the matched pixel as the cloud height is matched directly (we are certain that the cloud is in both the ceilometer and satellite field of view). LST is retrieved for these pixels (which will include the effect of the cloud on the retrieved surface temperature) and we plot the PDF of the difference between these LSTs and the climatological LST (Fig. 11). In all cases we fit a Gaussian distribution to the PDF.

The mean value and number of observations corresponding to each PDF are provided in Table 8 for the missed clear-sky pixels (corresponding to Fig. 10), and Table 9 for the cloud contaminated pixels (corresponding to Fig. 11). Considering first the missed clear-sky pixels; in NY the largest numbers of missed pixels during the day occur in AMJ and JAS with the largest proportion missed by UoR. Despite having the largest number of pixels missed, the temperature difference of these pixels is smaller relative to the climatology for UoR than for UoL (−1.58 K compared with −3.37 K in AMJ, −0.07 K compared with −4.61 K in JAS). This is important, as the smaller the temperature bias in the missed observations, the smaller the impact of excluding these values when calculating an average LST over a grid cell.

For Oper the temperature difference is negative in AMJ (−2.28 K) and positive in JAS (1.33 K). At night, the largest number of missed pixels occur in JFM and OND for all algorithms. In these months cloud and snow-covered ground surfaces are more frequently of a similar

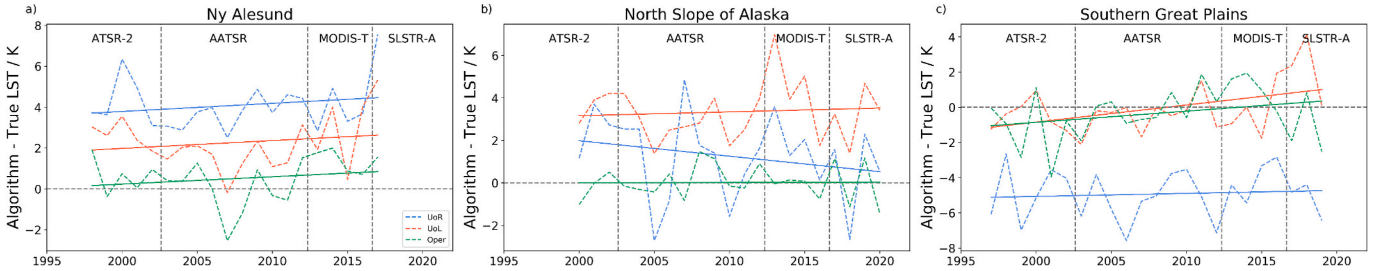

Fig. 9. LST stability for the satellite-ceilometer matches at NY, NSA and SGP. Timeseries show the 'algorithm-specific' LST anomalies minus the 'true' LST anomalies. Stability is determined using a linear fit to the resulting difference.

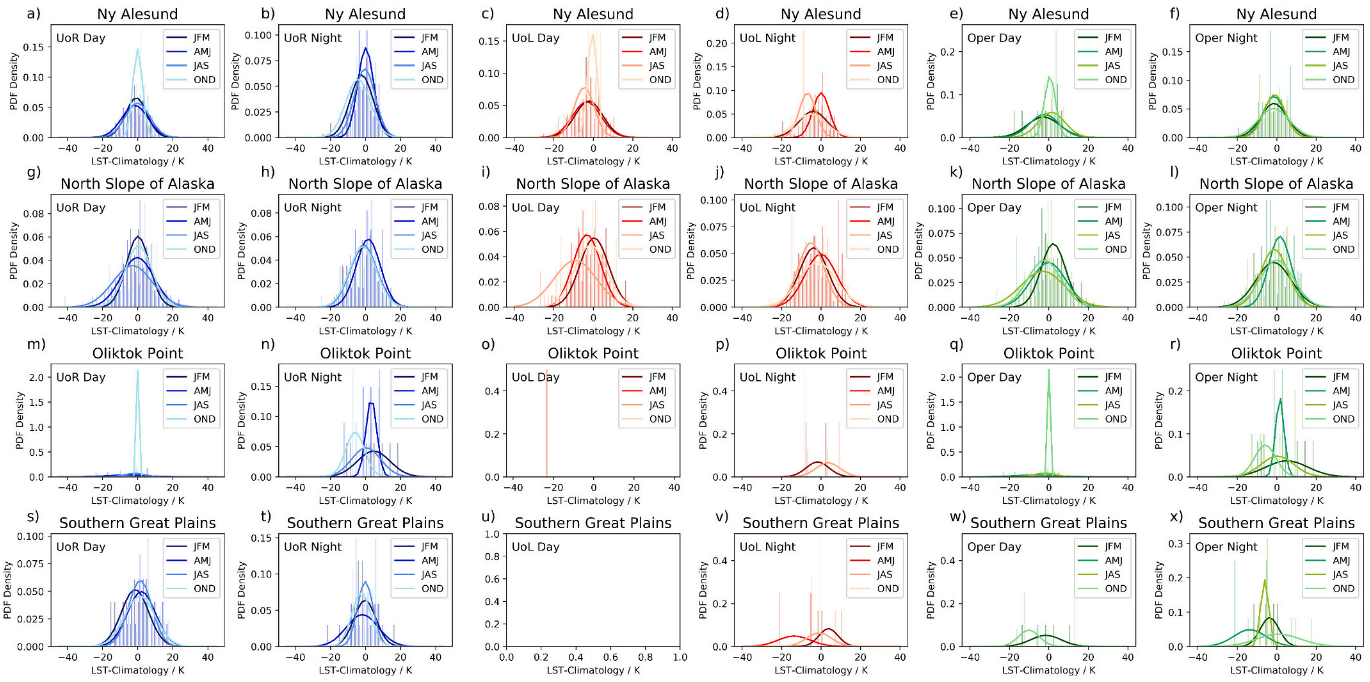

Fig. 10. Seasonal probability density functions for the missed clear-sky pixels in the ceilometer-satellite matches for each location. Results are presented independently for each algorithm: UoR (blue), UoL (red) and Oper (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

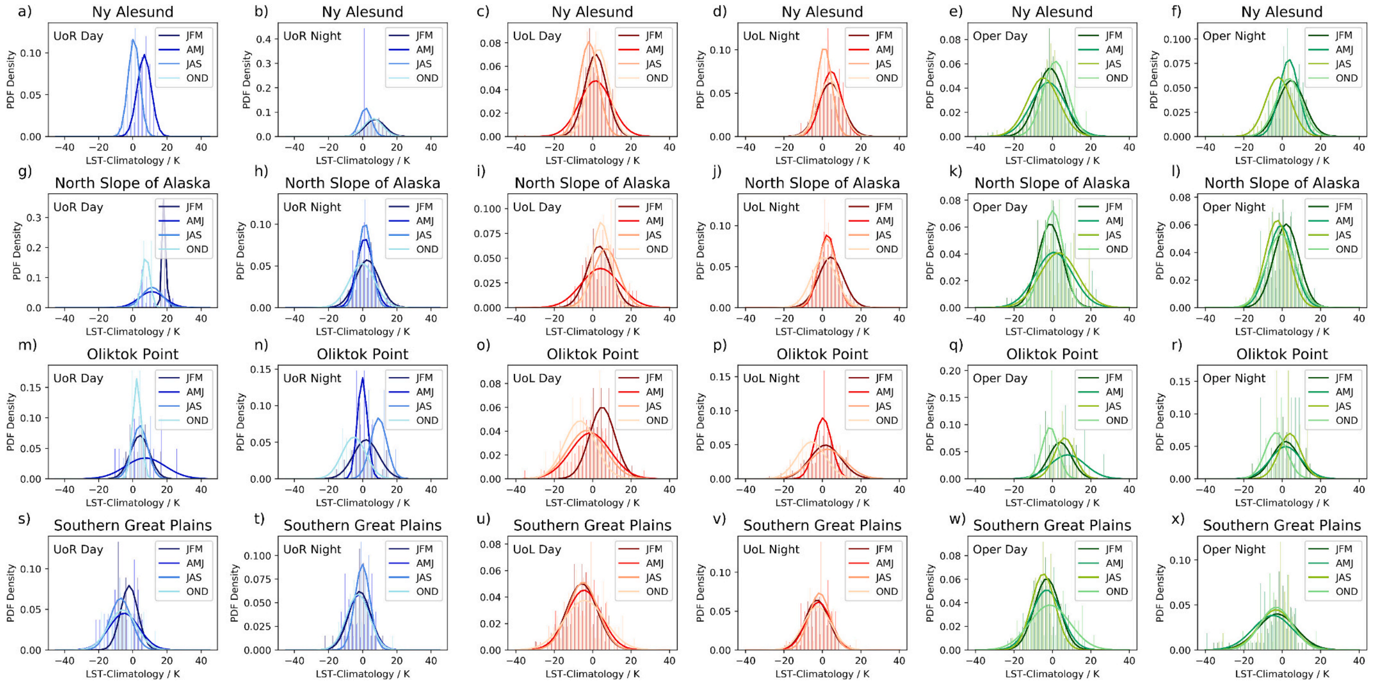

Fig. 11. Seasonal probability density functions for the cloud contaminated pixels in the ceilometer-satellite matches for each location. Results are presented independently for each algorithm: UoR (blue), UoL (red) and Oper (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

temperature, making it harder to differentiate cloud from snow where only thermal channels are available. The cloud masks have a tendency to over-screen, the impact of which is typically exclusion of pixels colder than the climatology, with mean values of −2.15 K and − 4.24 K for UoR in JFM and OND respectively, and − 4.04 K and − 5.19 K for UoL. For Oper, the mean temperature difference for the missed pixels is smaller, −1.43 K and − 1.55 K in JFM and OND.

NSA follows a similar pattern for daytime over-screening, with AMJ and JAS having the highest numbers. UoR and Oper miss the largest number of matches, with very similar statistics in the missed matches for both. In AMJ the temperature bias is small relative to the climatology (−0.18 K and − 0.21 K for UoR and Oper respectively), but larger in JAS (−2.98 K and − 3.18 K). For UoL, despite over-screening fewer observations, the difference in these observations from the climatology is

Table 8
Missed clear-sky PDF mean values and number of missed observations for day (bold) and night.

		PDF Mean								Number of Observations							
		JFM		AMJ		JAS		OND		JFM		AMJ		JAS		OND	
NY	UoR	**−0.8**	−2.15	**−1.58**	0.14	**−0.07**	−0.43	**−0.004**	−4.24	**36**	125	**161**	17	**89**	18	**5**	160
	UoL	**−2.25**	−4.04	**−3.37**	−6.9	**−4.61**	−6.9	**−0.5**	−5.19	**16**	59	**84**	11	**27**	11	**3**	85
	Oper	**−2.76**	−1.43	**−2.28**	−1.39	**1.33**	−1.08	**0.54**	−1.55	**8**	120	**77**	8	**56**	23	**4**	167
NSA	UoR	**0.16**	−1.08	**−0.18**	1.6	**−2.98**	−1.02	**0.5**	−1.45	**60**	101	**132**	27	**97**	39	**25**	94
	UoL	**0.45**	−3.72	**−3.24**	−0.16	**−9.18**	−5.14	**−1.41**	−1.77	**24**	90	**59**	20	**55**	23	**18**	69
	Oper	**2.2**	−1.58	**−0.21**	1.6	**−3.18**	−1.26	**−2.29**	0.04	**20**	74	**106**	14	**94**	31	**14**	86
OLI	UoR	**−7.32**	4.67	**−5.96**	3.32	**−1.72**	0.36	**0.005**	−5.54	**1**	8	**5**	3	**7**	5	**2**	5
	UoL	**–**	−1.79	**−22.9**	–	**–**	3.47	**–**	−8.74	**0**	2	**1**	0	**0**	4	**0**	1
	Oper	**−7.32**	5.13	**−6.45**	1.54	**−2.08**	0.36	**0.005**	−5.85	**1**	6	**5**	2	**7**	5	**2**	5
SGP	UoR	**−1.48**	−0.3	**2.13**	−1.8	**1.45**	−0.11	**1.25**	−2.42	**22**	15	**22**	15	**32**	6	**37**	7
	UoL	**–**	3.73	**–**	−13.64	**–**	−1.07	**–**	−1.13	**0**	3	**0**	2	**0**	3	**0**	1
	Oper	**−1.54**	−3.61	**–**	−13.64	**13.18**	−6.08	**−10.16**	0.94	**5**	4	**0**	2	**1**	8	**6**	6

Table 9
Cloud PDF mean values and number of cloud contaminated observations for day (bold) and night.

		PDF Mean								Number of Observations							
		JFM		AMJ		JAS		OND		JFM		AMJ		JAS		OND	
NY	UoR	**–**	7.23	**6.65**	1.65	**0.63**	1.68	**–**	6.26	**0**	426	**159**	1	**247**	97	**0**	671
	UoL	**1.58**	4.13	**1.46**	4.53	**−1.86**	0.98	**3.04**	3.81	**135**	1048	**586**	32	**910**	225	**39**	1537
	Oper	**−1.15**	4.66	**−2.08**	3.48	**−4.74**	−2.0	**1.63**	3.44	**151**	593	**746**	27	**464**	189	**24**	972
NSA	UoR	**18.2**	2.65	**11.1**	1.21	**11.3**	1.24	**7.69**	−0.43	**2**	187	**28**	26	**30**	65	**2**	260
	UoL	**3.42**	4.36	**4.11**	2.58	**6.95**	2.42	**4.52**	0.69	**69**	293	**118**	65	**102**	53	**32**	382
	Oper	**−1.08**	2.07	**0.89**	−0.73	**3.32**	−2.64	**0.19**	−0.75	**69**	204	**73**	36	**54**	69	**31**	304
OLI	UoR	**4.15**	1.93	**7.33**	−0.25	**4.07**	9.33	**2.4**	−5.15	**10**	34	**9**	6	**13**	11	**5**	32
	UoL	**4.98**	1.54	**−1.1**	0.27	**−3.16**	2.55	**−6.56**	−5.76	**46**	127	**71**	22	**126**	71	**22**	138
	Oper	**4.02**	1.72	**8.0**	1.83	**6.38**	3.97	**−1.2**	−3.25	**8**	21	**11**	4	**20**	6	**5**	16
SGP	UoR	**−2.12**	−1.76	**−5.12**	−2.42	**−7.16**	−0.18	**−6.05**	−2.72	**10**	29	**12**	33	**10**	31	**19**	30
	UoL	**−5.61**	−2.84	**−4.51**	−2.05	**−5.01**	−1.53	**−4.2**	−2.67	**77**	80	**85**	50	**55**	43	**38**	75
	Oper	**−2.83**	−2.71	**−2.98**	−4.91	**−4.7**	−3.31	**−1.32**	−3.18	**51**	63	**64**	54	**49**	54	**40**	47

larger (−3.24 K and − 9.18 K). The larger differences would result in a bigger discrepancy between the true LST averaged over a given timeframe and the sub-sampled LST due to over-screening. At night, matches are more frequently missed in JFM and OND but the temperature difference of these missed observations is smaller than observed in NY (−1.08 and − 1.02 K for UoR, −3.72 and − 1.77 K for UoL and − 1.58 and 0.04 K for Oper).

Under sampling in OLI and SGP is much less frequent than for NY and NSA. All algorithms have between 0 and 8 missed pixels in any given season so it is difficult to draw conclusions on the type of pixel likely to be erroneously masked as cloud in this case. For UoL and Oper the numbers of missed clear-sky pixels are similar in SGP. For UoR, the number of missed pixels is slightly higher (mostly during the day) with a range in the mean temperature difference of these missed matches of between −1.48 and 2.13 K across the year.

We consider now the cloud contaminated observations that remain unscreened by each of the evaluated algorithms. In NY, the largest numbers of cloud-contaminated observations occur in AMJ and JAS for daytime matches. The absolute numbers are significantly larger for UoL (586 and 910) and Oper (746 and 464) than they are for UoR (159 and 247). The temperature difference of these cloud contaminated observations relative to the climatology is algorithm dependent. For UoR the temperature differences are positive (6.65 K and 0.63 K in AMJ and JAS). For UoL the sign of the temperature difference is different for the two seasons (1.46 and − 1.86 K) and for Oper the difference is negative (−2.08 and − 4.74 K). At night, the largest number of cloud-contaminated pixels occur in JFM and OND. UoL has significantly more cloud-contaminated observations than Oper (1048 and 1537 compared with 593 and 972) but very similar positive temperature biases in the undetected cloud (4.13 and 3.81 K for UoL and 4.66 and 3.44 K for Oper). UoR has fewer cloud contaminated observations (426 and 671), but they have a larger positive temperature bias (7.23 and 6.26 K).

In NSA, UoR has relatively few cloud-contaminated observations during the day, but those that do occur are significantly warmer than the climatology, with the mean temperature difference ranging between 7.69 and 18.2 K. For UoL the mean value of the cloud-contaminated observations is also warmer than the climatology (3.42–6.95 K). At night, more cloud-contaminated observations occur, particularly during JFM and OND. For UoR and UoL the temperature differences are generally positive, but smaller than the daytime case (−0.43 to 2.65 K for UoR and 0.69 to 4.36 K for UoL). For Oper the temperature difference is positive in JFM (2.07 K) and negative throughout the remainder of the year (−0.73 to −2.64 K).

In OLI, UoL has the largest number of cloud-contaminated daytime matches. The maximum number occurs in JAS with a negative temperature bias of −3.16 K, but the largest temperature bias is in OND (−6.56 K). For UoR and Oper, fewer cloud-contaminated matches occur, but tend to have a positive temperature bias. At night, the cloud contaminated temperature bias is negative for all algorithms in OND (−3.25 to −5.76 K), but positive during the remainder of the year with the exception of UoR in AMJ.

In SGP, the numbers of cloud-contaminated matches are more consistent across all seasons for each of the algorithms. The mean temperature of these contaminated matches is always colder than the climatology with larger cold biases during the day than at night. For UoR, the number of cloud contaminated matches is lowest, but the mean temperature difference largest during the day (−2.12 to −7.16 K compared with −4.2 to −5.61 K for UoL and − 1.32 K to −4.7 K for Oper). At night, cold biases from cloud contamination are generally smaller for UoR and UoL, but larger for Oper.

Considering the cloud detection stability results (Table 6), our understanding of the missed clear sky observations (Fig. 10) and cloud contaminated matches (Fig. 11) we can evaluate the LST stability in results section 5.5. From the cloud detection stability results only two timeseries were stable with respect to both missed clear sky observations and cloud contamination: UoR in an NSA and SGP. These were stable with respect to the full-time series, but not for every individual sensor within the time series. From the LST stability analysis we had four stable time series: UoL and Oper at NSA and UoR and Oper at SGP.

Stability with respect to both cloud detection metrics is required for stability in the LST time series. In addition to this stability, there is another aspect that is also important for LST retrieval; namely stability in the surface temperature of the missed clear-sky pixels (relative to the climatology) over time, and likewise stability in the temperature of the cloud contaminated pixels in relation to the climatology, over time. For example, the same fraction of cloud contaminated observations may occur every year over the duration of a time series but if the temperature profile of those cloud contaminated pixels changes with respect to the climatology, stability in the LST time series cannot be guaranteed.

Variability can also occur as a result of sampling frequency when there is seasonal variation in the temperature bias of missed clear-sky pixels or cloud contaminated pixels relative to the climatology. For example, if the number of daytime missed clear-sky pixels increased over the time series relative to nighttime missed clear-sky matches this might alter the LST bias with time. Large natural variability in the annual average LST anomaly also makes detection of the impact of cloud masking stability in the LST signal very challenging.

6. Discussion

Understanding dataset stability is necessary for accurate assessment of uncertainties in temporal trends, but the stability of pre- or post-processing steps (e.g. cloud detection) in the provision of CDRs are rarely considered. Nonetheless they are important as they have the potential to introduce non-geophysical trends into the data, which may affect comparisons with reference datasets (e.g. in-situ data) that do not require the same screening. We demonstrate here the complexity involved in assessing cloud detection performance metric stability, given that validation data for cloudy/clear status cannot be assumed perfect. Metrics can vary between sensors as the result of external (geophysical) factors such as changes in cloud amount or due to changes in screening performance, e.g. tuning differences in the cloud screening, between sensors (non-geophysical). Metrics can also be location dependent (biome, cloud regime, viewing geometry) and these factors can affect different cloud screening algorithms in different ways. Achieving temporal stability in cloud masking performance requires both within-sensor stability and across-sensor stability.

Where reference data exist, we have demonstrated here a novel methodology for identifying a plausible range in the CC and MC metrics. The range can then be used to specify the uncertainty in the cloud detection metrics, essential in the calculation of stability metrics. We demonstrate that this stability can be calculated to within a high level of accuracy, which is far beyond the accuracy within which LST stability can be calculated. The large interannual variability in LST makes it difficult to assess the impact of the cloud masking stability as the signal is masked by natural variability over the short time duration of the existing satellite LST CDRs as matched to ceilometer data.

Cloud affecting a single pixel retrieval can result in a significantly biased LST. Typically, this temperature bias is assumed to be cold, but we have shown that warm biases from cloud can occur in polar regions. Most users do not use the per-pixel retrieved products directly, but use gridded data, which is likely already to have undergone some form of averaging, and they may then further average the data for the purpose of their analysis if they are looking at longer-term variability, seasonal effects or larger-scale spatial variability. The process of averaging the data in space and time has the effect of diluting the bias, making the net impact of cloud contamination on the calculated metric dependent on the cloud contaminated fraction of the data as shown in this manuscript.

Quantifying the uncertainty associated with cloud contamination is therefore difficult. Returning to the single pixel example, a cloud contaminated pixel will have a retrieved LST with a large uncertainty attributable to cloud. However, a neighbouring correctly-classified clear-sky pixel will have no uncertainty associated with the cloud detection. Therefore, specifying an uncertainty associated with a binary classification, rather than a gradated scale is complex. Correct attribution of the uncertainty would require certainty on the accuracy of the classification, and indeed if this was known, the errors in the classification would be corrected. Assigning a 'generic' uncertainty at the pixel level would result in an overestimation for correctly classified pixels and likely an underestimation for the erroneously classified pixels in an attempt not to penalise the correct classification too severely. If the uncertainty cannot be well defined at the pixel level, then it cannot be propagated through into gridded, higher-level products. One possible approach is to define an uncertainty on the gridded product, as is done for the sampling uncertainty (Bulgin et al., 2016b), which would be a function of the cloud contaminated fraction and would require some dependence on latitude and/or biome (and possibly other external factors).

Returning to the question of how to quantify cloud contamination uncertainty at the per-pixel level (enabling propagation through to all products), one approach could be to base this on the probability that a pixel is cloud contaminated, thereby providing a gradated scale rather than a binary classification. This information is available from the UoR and UoL cloud detection algorithms, but not all cloud-screening methodologies provide such information, particularly when they are based on threshold testing. Where such information does exist, challenges would remain in quantifying the potential impact of cloud contamination on the retrieved LST. For example, cloud type would be a determining factor here; thick, convective cloud will have a larger impact on the retrieved LST than thin, broken or partially-transparent cirrus cloud.

The assessment of cloud masking stability presented here is limited geographically by the availability of readily available reference data. The sites evaluated in this manuscript were limited to four (three of which are in polar regions) due to: 1) the limited number of ceilometer sites with long data records and 2) the limited number of ceilometers with sufficient attenuation heights to observe cirrus cloud. Both are required to assess long-term temporal stability and make a fair comparison between in-situ and satellite observations (as satellites see the full atmospheric column). These four sites are insufficient to characterise the variability in cloud masking stability likely to arise in different regions of the world due to varying atmospheric and cloud regimes. More data are required to extend this type of analysis and increase global representivity. One approach to increasing the viability of this type of analysis across the globe could be satellite-following ceilometers (with attenuation heights sufficient to detect cirrus cloud). By removing the limitation of a fixed nadir viewing angle of the ceilometer in the matching process, the number of clear-sky matches would increase. This is a forward-looking solution but doesn't address historical data gaps from in-situ cloud-viewing ceilometers.

This paper has demonstrated that temporal instability in the performance of cloud detection algorithms can impact the stability of long-term LST trends. Research and development of cloud-detection methods has been on-going throughout the satellite era (more than 40 years), during which time no perfect automated methodology has been found. The definition of 'perfect' is also subjective; dependent on the application/purpose of the cloud mask and difficult to define where the answer isn't clear-cut e.g. where exactly is the edge of a cloud? The solution to this problem therefore lies in thoroughly understanding the mechanisms of cloud masking instability and quantifying their impact sufficiently to estimate corrections when calculating trends in LST, or at least estimating the consequent uncertainty in the observed LST changes.

It would be premature to derive such a correction at this stage due to the geographical limitations of the analysis presented here, imposed by the lack of appropriate in-situ data available to validate cloud mask stability. A more geographically complete understanding of the mechanisms could be achieved using shorter-term ceilometer records (covering only one or two of the most recent satellite sensors contributing to the CDR discussed in this manuscript), as newer ceilometers are more likely to have sufficient attenuation heights to detect cirrus cloud. New in-situ sensors, such as satellite-tracking ceilometers (as discussed above) would remove ambiguity in the matches between the satellite and ceilometer that arise with the sensors seeing a different atmospheric path. Both could be used to improve our understanding of how/why cloud detection algorithms fail and the spatial variability in these failures.

At the current time, the most appropriate course of action is making data producers aware of this source of instability so that they can take steps to promote its quantification (for their particular product/application and in lobbying for new in-situ instruments specifically designed to facilitate cloud mask validation) and include this in the uncertainty information provided with their data. The second step is to alert users to this source of long-term instability in the data, promoting the use of the uncertainty information provided with data products to account for this instability when calculating long-term trends in LST, and applying the same principles to other geophysical variables that are reliant on cloud detection prior to retrieval. These are the top priorities in the near-term for exploiting the results presented in this paper.

7. Conclusions

We have demonstrated in this paper that cloud detection methodologies cannot be assumed to be temporally stable and that inconsistencies in their performance can result in instabilities as large as $+/-$ 0.73 K per decade, which is 0.43 K larger than the threshold stability target as set in the GCOS requirements. This assessment is of relevance to LST data users as cloud masking instability affects the calculation of uncertainties in long-term trends. To the best of our knowledge, cloud detection stability has not been previously assessed for any other target climate variable, so the conclusions drawn here may apply more widely to other satellite-derived CDRs that are reliant on cloud detection as a pre-processing step. Further assessments of a similar nature should be made following updates to cloud detection methodologies in LST CDR generation (for example for v2.0 of the LST CCI dataset using an updated version of the UoL algorithm). The ability to carry out these assessments would be greatly enhanced by the development of satellite-following ceilometry.

Funding

This study has received funding from the European Space Agency (ESA) within the framework of the Land Surface Temperature project under the Climate Change Initiative (LST_cci), grant number 4000123553/18/I-NB.

CRediT authorship contribution statement

Claire E. Bulgin: Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Ross I. Maidment:** Software, Formal analysis, Data curation. **Darren Ghent:** Writing – review & editing, Funding acquisition, Data curation. **Christopher J. Merchant:** Writing – review & editing, Methodology, Conceptualization.

Data availability

Data will be made available on request.

References

Ackerman, S.A., Strabala, K.I., Menzel, W.P., Frey, R.A., Moeller, C.C., Gumley, L.E., 1998. Discriminating clear sky from clouds with MODIS. J. Geophys. Res. Atmos. 103, 32141–32157. https://doi.org/10.1029/1998JD200032.

Ackerman, S., Frey, R., Strabala, K., Liu, Y., Gumley, L., Baum, B., Menzel, P., 2010. Discriminating Clear-Sky from Cloud with MODIS Algorithm Theoretical Basis Document (MOD35), vol. No. Version 6.1. Cooperative Institute for Meteorological Satellite Studies, University of Wisconsin, Madison.

Aldred, F., Good, E., Bulgin, C., Rayner, N., 2023. User Requirements Document (CCI Land Surface Temperature No. LST-CCI-D1.1-URD v3).

Anderson, M.C., Hain, C.R., Jurecka, F., Trnka, M., Hlavinka, P., Dulaney, W., Otkin, J. A., Johnson, D., Gao, F., 2016. Relationships between the evaporative stress index and winter wheat and spring barley yield anomalies in the Czech Republic. Clim. Res. 70, 215–230.

Bento, V.A., DaCamara, C.C., Trigo, I.F., Martins, J.P.A., Duguay-Tetzlaff, A., 2017. Improving land surface temperature retrievals over mountainous regions. Remote Sens. 9, 38. https://doi.org/10.3390/rs9010038.

Berry, D.I., Corlett, G.K., Embury, O., Merchant, C.J., 2018. Stability assessment of the (A)ATSR Sea surface temperature climate dataset from the European Space Agency Climate Change Initiative. Remote Sens. 10, 126. https://doi.org/10.3390/rs10010126.

Birks, A.R., 2007. Improvements to the AATSR IPF Relating to Land Surface Temperature Retrieval and Cloud Clearing over Land (AATSR Technical Note). Rutherford Appleton Laboratory, Chilton, Didcot.

Bulgin, C.E., Sembhi, H., Ghent, D., Remedios, J.J., Merchant, C.J., 2014. Cloud-clearing techniques over land for land-surface temperature retrieval from the Advanced Along-Track Scanning Radiometer. Int. J. Remote Sens. 35, 3594–3615. https://doi.org/10.1080/01431161.2014.907941.

Bulgin, C.E., Embury, O., Corlett, G., Merchant, C.J., 2016a. Independent uncertainty estimates for coefficient based sea surface temperature retrieval from the Along-Track Scanning Radiometer instruments. Remote Sens. Environ. 178, 213–222. https://doi.org/10.1016/j.rse.2016.02.022.

Bulgin, C.E., Embury, O., Merchant, C.J., 2016b. Sampling uncertainty in gridded sea surface temperature products and Advanced Very High Resolution Radiometer (AVHRR) Global Area Coverage (GAC) data. Remote Sens. Environ. 177, 287–294. https://doi.org/10.1016/j.rse.2016.02.021.

Bulgin, C., Merchant, C., Ghent, D., Klüser, L., Popp, T., Poulsen, C., Sogacheva, L., 2018. Quantifying uncertainty in satellite-retrieved land surface temperature from cloud detection errors. Remote Sens. 10, 616. https://doi.org/10.3390/rs10040616.

Bulgin, C.E., Merchant, C.J., Ferreira, D., 2020. Tendencies, variability and persistence of sea surface temperature anomalies. Sci. Rep. 10, 7986. https://doi.org/10.1038/s41598-020-64785-9.

Bulgin, C.E., Embury, O., Maidment, R.I., Merchant, C.J., 2022. Bayesian cloud detection over land for climate data records. Remote Sens. 14, 2231. https://doi.org/10.3390/rs14092231.

Coppo, P., Ricciarelli, B., Brandani, F., Delderfield, J., Ferlet, M., Mutlow, C., Munro, G., Nightingale, T., Smith, D., Bianchi, S., Nicol, P., Kirschstein, S., Hennig, T., Engel, W., Frerick, J., Nieke, J., 2010. SLSTR: a high accuracy dual scan temperature radiometer for sea and land surface monitoring from space. J. Mod. Opt. 57, 1815–1830. https://doi.org/10.1080/09500340.2010.503010.

Ding, H., Xu, L., Elmore, A.J., Shi, Y., 2020. Vegetation phenology influenced by rapid urbanization of the Yangtze Delta region. Remote Sens. 12. https://doi.org/10.3390/rs12111783.

Donlon, C., Berruti, B., Buongiorno, A., Ferreira, M.-H., Féménias, P., Frerick, J., Goryl, P., Klein, U., Laur, H., Mavrocordatos, C., Nieke, J., Rebhan, H., Seitz, B., Stroede, J., Sciarra, R., 2012. The Global Monitoring for Enviroment and Security (GMES) Sentinel-3 mission. Remote Sens. Environ. 120, 37–57. https://doi.org/10.1016/j.rse.2011.07.024.

Duguay-Tetzlaff, A., Bento, V.A., Goettsche, F.M., Stoeckli, R., Martins, J.P.A., Trigo, I., Olesen, F., Bojanowski, J.S., da Camara, C., Kunz, H., 2015. Meteosat land surface temperature climate data record: achievable accuracy and potential uncertainties. Remote Sens. 7, 13139–13156. https://doi.org/10.3390/rs71013139.

Embury, O., Merchant, C.J., Good, S.A., Rayner, N.A., Høyer, J.L., Atkinson, C., Block, T., Alerskans, E., Pearson, K.J., Worsfold, M., McCarroll, N., Donlon, C., 2024. Satellite-based time-series of sea-surface temperature since 1980 for climate applications. Sci. Data 11, 326. https://doi.org/10.1038/s41597-024-03147-w.

European Space Agency, 2023. Sentinel-3 SLSTR Technical Guide (Online Guide). European Space Agency.

Foster, G., Rahmstorf, S., 2011. Global temperature evolution 1979-2010. Environ. Res. Lett. 6, 044022. https://doi.org/10.1088/1748-9326/6/4/044022.

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. J. Atmos. Ocean. Technol. 25, 1057–1072. https://doi.org/10.1175/2008JTECHA1052.1.

Ghent, D.J., Corlett, G.K., Göttsche, F.-M., Remedios, J.J., 2017. Global land surface temperature from the Along-Track Scanning Radiometers. J. Geophys. Res. Atmos. 122, 12,167–12,193. https://doi.org/10.1002/2017JD027161.

Ghent, D., Veal, K., Trent, T., Dodd, E., Sembhi, H., Remedios, J., 2019. A new approach to defining uncertainties for MODIS land surface temperature. Remote Sens. 11, 1021. https://doi.org/10.3390/rs11091021.

Global Climate Observing System, 2022. The 2022 GCOS Implementation Plan (No. GCOS-244). WMO, Geneva.

Good, E.J., Ghent, D.J., Bulgin, C.E., Remedios, J.J., 2017. A spatiotemporal analysis of the relationship between near-surface air temperature and satellite land surface temperatures using 17 years of data from the ATSR series. J. Geophys. Res.-Atmos. 122, 9185–9210. https://doi.org/10.1002/2017JD026880.

Good, E.J., Aldred, F.M., Ghent, D.J., Veal, K.L., Jimenez, C., 2022. An analysis of the stability and trends in the LST_cci land surface temperature datasets over Europe. Earth Space Sci. 9. https://doi.org/10.1029/2022EA002317 e2022EA002317.

He, X., Wang, D., Gao, S., Li, X., Chang, G., Jia, X., Chen, Q., 2024. The anisotropy of MODIS LST in urban areas: a perspective from different time scales using model simulations. ISPRS J. Photogramm. Remote Sens. 209, 448–460. https://doi.org/10.1016/j.isprsjprs.2024.02.012.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Hocking, J., Rayer, P., Rundle, D., Saunders, R., Matricardi, M., Geer, A., Brunel, P., Vidot, J., 2015. RTTOV v11 Users Guide, vol. No. 1.4. EUMETSAT. NWPSAF-MO-UD-028.

Hocking, J., Rayer, P., Rundle, D., Saunders, R., Matricardi, M., Geer, A., Brunel, P., Vidot, J., 2019. RTTOV v12 Users Guide, vol. No. 1.3. EUMETSAT. NWPSAF-MO-UD-037.

Hollmann, R., Merchant, C.J., Saunders, R., Downy, C., Buchwitz, M., Cazenave, A., Chuvieco, E., Defourny, P., de Leeuw, G., Forsberg, R., Holzer-Popp, T., Paul, F., Sandven, S., Sathyendranath, S., van Roozendael, M., Wagner, W., 2013. The ESA Climate Change Initiative: satellite data records for Essential Climate Variables. Bull. Am. Meteorol. Soc. 94, 1541–1552. https://doi.org/10.1175/BAMS-D-11-00254.1.

IDEAS+AATSR QC Team, 2016. (A)ATSR Third Reprocessing Dataset User Summary. Telespaio VEGA, UK.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J.J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., Suttie, M., 2019. The CAMS reanalysis of atmospheric composition. Atmos. Chem. Phys. 19, 3515–3556. https://doi.org/10.5194/acp-19-3515-2019.

Ji, L., Senay, G., Velpuri, N., Kagone, S., 2019. Evaluating the temperature difference parameter in the SSEBop model with satellite-observed land surface temperature data. Remote Sens. 11. https://doi.org/10.3390/rs11161947.

Kayet, N., Pathak, K., Chakrabarty, A., Sahoo, S., 2016. Spatial impact of land use/land cover change on surface temperature distribution in Saranda Forest, Jharkhand. Model. Earth Syst. Environ. 2, 127. https://doi.org/10.1007/s40808-016-0159-x.

Kogler, C., Pinnock, S., Arino, O., Casadio, S., Corlett, G., Prata, F., Bras, T., 2012. Note on the quality of the (A)ATSR land surface temperature record from 1991 to 2009. Int. J. Remote Sens. 33, 4178–4192. https://doi.org/10.1080/01431161.2011.645085.

Li, Z.-L., Tang, B.-H., Wu, H., Ren, H., Yan, G., Wan, Z., Trigo, I.F., Sobrino, J.A., 2013. Satellite-derived land surface temperature: current status and perspectives. Remote Sens. Environ. 131, 14–37. https://doi.org/10.1016/j.rse.2012.12.008.

Lieberherr, G., Wunderle, S., 2018. Lake surface water temperature derived from 35 years of AVHRR sensor data for European Lakes. Remote Sens. 10, 990. https://doi.org/10.3390/rs10070990.

Mao, K., Yuan, Z., Zuo, Z., Xu, T., Shen, X., Gao, C., 2019. Changes in global cloud cover based on remote sensing data from 2003 to 2012. Chin. Geogr. Sci. 29, 306–315. https://doi.org/10.1007/s11769-019-1030-6.

Masuoka, E., Fleig, A., Wolfe, R.E., Patt, F., 1998. Key characteristics of MODIS data products. IEEE Trans. Geosci. Remote Sens. 36, 1313–1323. https://doi.org/10.1109/36.701081.

Maturilli, M., Herber, A., 2017. Ceilometer Cloud Base Height from Station Ny-Ålesund from August 1992 to July 2017, Reference List of 290 Datasets 290 Datasets. https://doi.org/10.1594/PANGAEA.880300.

Merchant, C.J., Harris, A.R., Maturi, E., Maccallum, S., 2005. Probabilistic physically based cloud screening of satellite infrared imagery for operational sea surface temperature retrieval. Q. J. R. Meteorol. Soc. 131, 2735–2755. https://doi.org/10.1256/qj.05.15.

Merchant, C.J., Embury, O., Bulgin, C.E., Block, T., Corlett, G.K., Fiedler, E., Good, S.A., Mittaz, J., Rayner, N.A., Berry, D., Eastwood, S., Taylor, M., Tsushima, Y., Waterfall, A., Wilson, R., Donlon, C., 2019. Satellite-based time-series of sea-surface temperature since 1981 for climate applications. Sci. Data 6, 223. https://doi.org/10.1038/s41597-019-0236-x.

Metropolis, N., Ulam, S., 1949. The Monte Carlo method. J. Am. Stat. Assoc. 44, 335–341. https://doi.org/10.2307/2280232.

Morris, V.R., 2016. Ceilometer Instrument Handbook (No. DOE/SC-ARM-TR–020, 1036530). https://doi.org/10.2172/1036530.

Morris, V., Zhang, D., Ermold, B., 1996. Ceil. https://doi.org/10.5439/1181954.

Na, Q., Cao, B., Qin, B., Mo, F., Zheng, L., Du, Y., Li, H., Bian, Z., Xiao, Q., Liu, Q., 2024. Correcting an off-nadir to a nadir land surface temperature using a multitemporal thermal infrared kernel-driven model during daytime. Remote Sens. 16. https://doi.org/10.3390/rs16101790.

Norris, J.R., 2005. Multidecadal changes in near-global cloud cover and estimated cloud cover radiative forcing. J. Geophys. Res. Atmos. 110. https://doi.org/10.1029/2004JD005600.

Peeling, J.A., Chen, C., Judge, J., Singh, A., Achidago, S., Eide, A., Tarrio, K., Olofsson, P., 2024. Applications of remote sensing for land use planning scenarios with suitability analysis. IEEE J. Select. Top. Appl. Earth Observat. Remote Sens. 17, 6366–6378. https://doi.org/10.1109/JSTARS.2024.3370379.

Perry, M., Ghent, D.J., Jiménez, C., Dodd, E.M.A., Ermida, S.L., Trigo, I.F., Veal, K.L., 2020. Multisensor thermal infrared and microwave land surface temperature algorithm intercomparison. Remote Sens. 12, 4164. https://doi.org/10.3390/rs12244164.

Riffler, M., Lieberherr, G., Wunderle, S., 2015. Lake surface water temperatures of European alpine lakes (1989-2013) based on the Advanced Very High Resolution Radiometer (AVHRR) 1 km data set. Earth Syst. Sci. Data 7, 1–17. https://doi.org/10.5194/essd-7-1-2015.

Sen, P.K., 1968. Estimates of the regression coefficient based on Kendall's tau. J. Am. Stat. Assoc. 63, 1379–1389. https://doi.org/10.1080/01621459.1968.10480934.

Simpson, J.J., Harris, A.R., Merchant, C.J., Murray, M.J., Allen, M.R., 2000. Development and validation of new surface temperature retrievals cloud classification algorithms and an evaluation of the diurnal cycle. In: Harris, R.A. (Ed.), First MSG RAO Workshop, ESA Special Publications. ESA Publications Division, Noordwijk, the Netherlands, pp. 141–144.

Simpson, J.J., McIntire, T.J., Stitt, J.R., Hufford, G.L., 2001. Improved cloud detection in AVHRR daytime and night-time scenes over the ocean. Int. J. Remote Sens. 22, 2585–2615. https://doi.org/10.1080/01431160119916.

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). J. Atmos. Ocean. Technol. 17, 595–615. https://doi.org/10.1175/1520-0426(2000)017<0595:CCOTOI>2.0.CO;2.