Data Fusion of AIRS and CrIMSS Near Surface Air Temperature

P. Kalmus[1], H. Nguyen[1], J. Roman[1], T. Wang[1], Q. Yue[1], Y. Wen[2], J. Hobbs[1], and A. Braverman[1]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, 91109
[2]Department of Geography, University of Florida, Gainesville, FL, 32611

Key Points:

- We have developed a method for fusing any number of two-dimensional remote sensing datasets which estimate the same observable
- We introduce a new daytime and nighttime fused near-surface air temperature product from satellite hyperspectral sounders over CONUS
- The fused product decreases bias and RMSE by 1 K and 25% respectively relative to input datasets, averaged over the domain of the study

Corresponding author: Peter Kalmus, `peter.m.kalmus@jpl.nasa.gov`

Abstract

We present a near surface air temperature (NSAT) fused data product over the contiguous United States using data from the Atmospheric Infrared Sounder (AIRS), on the Aqua platform, and the Cross-track Infrared Microwave Sounding Suite (CrIMSS), on the Suomi National Polar-orbiting Partnership (NPP) platform. We create the fused product using a fast python implementation of Spatial Statistical Data Fusion (SSDF) along with weather station data from NOAA's Integrated Surface Database (ISD) which is used to estimate bias and variance in the input satellite datasets. Our fused NSAT product is produced twice-daily (one daytime and one nighttime estimate per day) and on a 0.25-degree latitude-longitude grid. We provide detailed validation using withheld ISD data and ERA5-Land reanalysis. The fused gridded product has no missing data; has improved accuracy and precision relative to the input satellite datasets, and comparable accuracy and precision to ERA5-Land; and includes accurate uncertainty estimates. Over the domain of our study, the fused product decreases daytime bias magnitude by 1.7 K and 0.5 K, nighttime bias magnitude by 1.5 K and 0.2 K, and overall RMSE by 35% and 15% relative to the AIRS and CrIMSS input datasets, respectively. Our method is computationally fast and generalizable, capable of data fusion from any number of datasets estimating the same quantity. Finally, because our product removes bias, it produces long-term datasets across multi-instrument remote sensing records with improved stationarity for climate trend analysis, even as individual missions and their data records begin and end.

1 Introduction

From the point of view of scientific analysis, satellite remote sensing datasets present several challenges. Many satellite remote sensing datasets are released as "Level 2" (L2) products, geophysical quantities retrieved from directly observed radiances. Instantaneous snapshots are obtained at a great number of spatial and temporal fields of regard, and data coverage can be spatially incomplete due to gores (spaces between orbit tracks determined by orbital and sensor geometry), clouds, downlink limitations, or other issues. Satellite retrievals suffer from uncertainties and errors due to information and algorithm limitations, while uncertainty estimates, if reported at all, are not always reliable. Drifts of orbits and spectral channels, and even sudden changes, make the use of data records from satellites challenging in climate studies by causing bias nonstationarity that must be separated from real signals. While L2 satellite data brings invaluable information to scientific analysis, using it appropriately requires significant expertise and involves serious limitations.

Data fusion is the combining of multiple datasets into a single dataset with better properties than any of the individual input datasets (for a recent review, see Ghamisi et al. (2019)). Here, we demonstrate a data fusion method, called Spatial Statistical Data Fusion (SSDF) that addresses each of the above issues (Nguyen et al., 2012, 2014). We use SSDF to create a fused near-surface air temperature (NSAT) product. NSAT is a critical remote sensing product for climate studies of extreme heat, as well as for many science applications areas of great importance to society such as health, agriculture, urban planning, hydrology and water management, ecology and conservation, and fire management. Our SSDF NSAT product combines two remote sensing data products: L2 NSAT from the Atmospheric Infrared Sounder (AIRS) on the Aqua platform, and L2 NSAT from the Cross-track Infrared Microwave Sounding Suite (CrIMSS) on the Suomi National Polar-orbiting Partnership (NPP) platform, which are furthermore created using two independent retrieval algorithms. We also use information content from in situ weather station networks (NOAA's Integrated Surface Database, or ISD) to determine uncertainties in the two remote sensing datasets which are needed to perform fusion. The fused NSAT product is produced on a twice-daily basis (one daytime and one nighttime estimate per day), and covers the contiguous United States (CONUS) and adjacent parts of North America.

Our fused SSDF NSAT product has the following key advantages over either of the input remote sensing datasets:

1. SSDF fills spatial gaps (e.g., due to orbital gores or clouds);
2. SSDF produces estimates on a regular 0.25-degree spatial grid;
3. SSDF reduces bias and variance;
4. SSDF produces uncertainty estimates that characterize the actual error with more skill than the input datasets;
5. SSDF improves long-term bias stationarity relative to the input datasets, facilitating creation of climate records over changing instrument epochs.

The rest of the paper is organized as follows. We first describe the input datasets and methodology. Then we present the SSDF NSAT product, and the results of validation against withheld ISD surface station data. We also compare the SSDF fused NSAT product to the individual input remote sensing datasets, and to ERA5-Land reanalysis. In the process of validating our SSDF product, we also produce the most thorough validation study to date of the AIRS V7 and CrIMSS-CLIMCAPS V2 NSAT products over CONUS. We conclude with a discussion of advantages, limitations, and potential future work.

2 Data and methods

Performing and evaluating SSDF involves five major steps: (1) Obtaining and filtering input remote sensing datasets that estimate the same quantity; (2) Matching the remote sensing datasets to a reference in situ dataset in space and time; (3) Using these matched data ("matchups") to characterize the input datasets via estimation of their bias and variance relative to the reference estimate; (4) Performing the SSDF calculations; and (5) Validating the results using withheld data from the reference estimate. The method and the specific datasets used in our NSAT dataset are described in the following subsections.

2.1 Satellite NSAT data

The input satellite datasets come from two generations of hyperspectral infrared sounders and retrieval algorithms. The Aqua platform that carries AIRS launched in 2002 in a sun-synchronous polar orbit, with equator crossing times of 1:30 P.M. and 1:30 A.M. for ascending (south to north) and descending (north to south) nodes, respectively. AIRS is an infrared grating spectrometer with 2378 channels, spanning 3.7 to 15.4 μm (Chahine et al., 2006). Power to critical channels of the Aqua satellite's Advanced Microwave Sounding Unit (AMSU)-A2 was lost in September 2016 (Yue et al., 2017), which was used to complement the AIRS instrument in atmospheric temperature and moisture profile retrievals.

We use the AIRS version 7 L2 "infrared-only" temperature retrieval algorithm (Susskind et al., 2014). This retrieval uses the Stochastic Cloud Clearing Neural Network (SCCNN) which is trained to ECMWF fields (Blackwell, 2005) as a first guess, then refines to a final estimate. It also uses information from the satellite's other microwave sounder, AMSU-A1 (Yue et al., 2020). The retrieval uncertainty is estimated via a regression model using eleven retrieval diagnostic quantities as predictors; the regression coefficients are trained on two days of retrievals (9/29/04 and 2/24/07) using ECMWF 3-hour forecasts as a reference dataset (Susskind et al., 2014; Thrastarson et al., 2020). Each individual retrieval has a nominal horizontal resolution of 45 km, and each swath contains 30 retrievals across its width and 45 along track. The product is organized nominally in 240 "orbital granules" per day (AIRS Project, 2020).

The Cross-track Infrared Sounder (CrIS) and the Advanced Technology Microwave Sounder (ATMS) instruments launched onboard the NPP platform in 2012. NPP is in the same orbital plane as Aqua, but at a higher altitude (824 km as opposed to 705 km), with equator crossing times also nominally of 1:30 P.M. and 1:30 A.M. for ascending and descending nodes, respectively. We use the Community Long-term Infrared Microwave Coupled Atmospheric Product System (CLIMCAPS) Version 2 L2 temperature retrieval, which uses an optimal estimation methodology with a first guess from the Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA2) (N. Smith & Barnet, 2020), and information from both instruments. CLIMCAPS uncertainty is estimated and propagated sequentially via error covariance matrices in stages (N. Smith & Barnet, 2019). CLIMCAPS produces a combined infrared and microwave retrieval at two spectral resolutions: Nominal Spectral Resolution (NSR) and Full Spectral Resolution (FSR). We use the CLIMCAPS-SNPP NSR product to create our SSDF product. In what follows, we refer to this product as "CrIMSS-CLIMCAPS" or simply "CrIMSS."

For both instruments, NSAT is obtained from the vertically-resolved temperature profile (100 pressure levels) by interpolation or extrapolation with pressure to the surface pressure for each field of regard (Olsen et al., 2017). The profile temperatures immediately above and below the surface are used for the interpolation, unless the level above is within 5 hPa of the surface pressure. In that case, the two levels above the surface are used. We ingest only L2 NSAT retrievals from AIRS V7 IR-only and CrIMSS-CLIMCAPS products with data quality flags 'good' or 'best' in our data fusion procedure.

2.2 In situ NSAT data

The National Oceanic and Atmosphere Administration (NOAA) Integrated Surface Database (ISD) is a global database of near-surface meteorological observations compiled from over a hundred sources (A. Smith et al., 2011). The record extends back to the 1950s, although new stations have been added on a continual basis as available, improving coverage over time. Today ISD consists of more than 35,000 surface weather stations globally, 14,000 of which remain active. Figure 1 shows the spatial coverage of ISD stations in North America.

We use sub-hourly 2 m NSAT measurements gathered from over 7000 stations in North America as our reference dataset, for bias and variance estimation and for validation. Naturally ventilated screened surface station air temperature measurements are accurate to ± 0.1°C in most circumstances(Harrison & Burt, 2021). ISD data come with a set of ten data quality flags, indicating various problems and levels of quality. We only use ISD data flagged as highest quality, i.e., data must be flagged with either 1 ('Passed all quality control checks') or 5 ('Passed all quality control checks, data originate from an NCEI data source').

2.3 Reanalysis

We also compare the SSDF NSAT results to European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis 5 (ERA5)-Land reanalysis data. The ERA5 is the fifth-generation global atmospheric reanalysis from ECMWF, replacing the ERA-Interim reanalysis which stopped being produced on August 31st, 2019. In addition, newly reprocessed datasets along with recent instruments have been assimilated into the ERA5 that could not be ingested into the ERA-Interim (Hennermann & Berrisford, 2019). We note that some AIRS spectral channels under clear conditions are

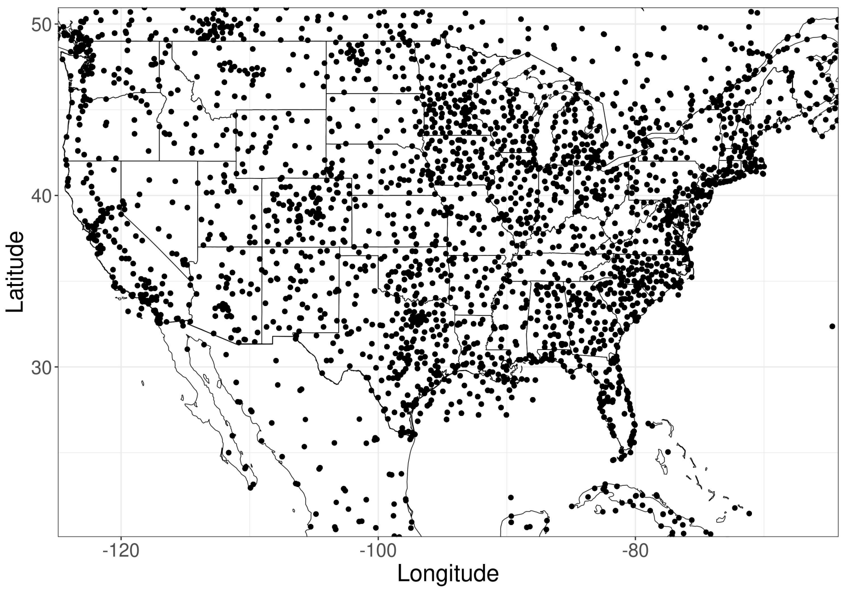

Figure 1: Spatial coverage of the ISD stations over North America.

incorporated into ECMWF reanalysis (Mcnally et al., 2006), but that ISD data are not.

We use hourly ERA5-Land output which is a high-resolution version (∼9 km) of the land component of the ERA5 climate reanalysis. ERA5-Land was chosen over the full ERA5 reanalysis for its finer spatial resolution of 0.1x0.1°. Hourly 2 m air temperature output was selected for our comparison.

2.4 Bias and variance estimation

Biases and variances of input data sources are crucial for proper data fusion. The SSDF methodology assumes input data are unbiased, and weights them by the inverse of their respective variances. This minimizes output errors of the fused estimates. Therefore, data are bias-corrected before SSDF ingestion, and the quality of the final fused product is largely determined by the quality of uncertainty estimates for the inputs.

To estimate bias and variance for satellite footprints, we create an ensemble of "matchups": matched pairs of satellite and ISD station estimates that are close in space and time. For a given period, the matchups are sorted into 240 km (∼two-degree) diameter hexagonal spatial bins based on satellite footprint location, with three-day time bins (day of interest, along with preceding and following days). This binning is the basis for quantifying bias and variance for all satellite footprints in a given space-time cell. We experimented with using longer and shorter time bins to explore the trade off between sample size and capturing rapid changes in conditions affecting retrieval bias, and found that the three-day bin delivered the lowest average biases and variances over CONUS. Before starting, we randomly selected 1% of the ISD matchups to withhold for validation. We chose a relatively small amount to withhold in order to maximize the information content for the SSDF product. In this subsection, the term "ISD" refers to the non-withheld ISD data.

To obtain the matchups we apply the following steps.

1. Given an ISD observation at location $\mathbf{s}$ and time $t^I(\mathbf{s})$, select the AIRS granule (1 of 240) with the closest time to $t^I(\mathbf{s})$.
2. Within this granule, select all L2 retrievals within 100 km of $\mathbf{s}$ and 1 hour of $t^I(\mathbf{s})$.
3. If Step 2 results in more than 1 retrieval, select the one closest in spatial distance.

Note that the Steps in 1-3 will result in a one-to-one match between an ISD observation and a single AIRS footprint. Some ISD observations may have no corresponding AIRS match, in which case we return a null result. We next tessellate a fixed hexagonal spatial grid over CONUS and find the biases and variances using matchups aggregated over 3 days within each grid cell. That is,

I. To compute a bias on day d and mode j (day or night) and in hexagonal grid cell i, we find the set of all valid (i.e., non-null) AIRS-ISD matchups from Steps 1-3 above such that,
   (a) the AIRS data come from mode j,
   (b) the AIRS footprint belongs within the grid cell i,
   (c) the ISD date is in $(d-1, d, d+1)$.

II. The bias and variance for day d, mode j, and grid cell i are then computed using the set of paired ISD-AIRS matchups.

Bias and variance estimation for CrIMSS follows the same procedure. For bias correction, given an instrument observation at location $\mathbf{s}$ on day d and mode j, we compute the corresponding bias within the grid cell which contains $\mathbf{s}$ for day d and mode j, and we subtract it from the instrument's NSAT value. For more detail on the bias and variance estimation process, please refer to the Appendix.

2.5 Data fusion methodology

In this section we review the framework of Spatial Statistical Data Fusion (SSDF; Nguyen et al., 2012) on two satellite NSAT datasets. Remote sensing data in general are heterogeneous. By this we mean that they may have different footprints, measurement error characteristics, and sampling patterns. We account for this by using a spatial statistical model that captures the spatial dependence between the true quantity of interest at a particular location and the observations from all data sources. In particular, the issue of different footprint sizes and shapes is known as a *change-of-support problem* (e.g. Gotway & Young, 2002), and we will address this using SSDF as described in Nguyen et al. (2012).

Consider a discretized domain where $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ is a hidden, real-valued spatial observable. The domain of interest is $\cup\{A_i \subset \Re^d : i = 1, \ldots, N_D\}$, which is made up of N_D fine-scale, non-overlapping, areal regions $\{A_i\}$ with locations $D \equiv \{\mathbf{p}_i \in A_i : i = 1, \ldots, N_D\}$. Nguyen et al. (2012) call these fine-scale regions Basic Areal Units (BAUs), and they represent the smallest resolution at which we will make estimates with the model.

For a given day and mode (d and j using the notation of the previous subsection), denote the vector of NSAT data at all locations by $\mathbf{Z}^k$, where $k = 1$ for AIRS and $k = 2$ for CrIMSS:

$$\mathbf{Z}^k \quad = \quad (Z^k(B_{k1}), Z^k(B_{k2}), \ldots, Z^k(B_{kN_k}))',$$

where $\mathbf{Z}^k$ is N_k-dimensional, B_{kq} is the q-th footprint from the k-th dataset and is made up of BAUs with locations indexed by $D \cap B_{kq}$. We assume that data observed at an arbitrary areal region B follow the "data model" in which the true observable is averaged over the areal region plus an independent error term. That is,

$$Z^k(B) = \frac{1}{|D \cap B|} \left\{ \sum_{\mathbf{s} \in D \cap B} Y(\mathbf{s}) \right\} + \epsilon^k(B); \;\; B \subset \Re^d. \tag{1}$$

where $Y(\cdot)$ is a geophysical observable (here, NSAT) that is common to both datasets, and $\epsilon^k(\cdot)$ is an independent but non-identically distributed Gaussian random variable. That is, we assume that the q-th error in the k-th dataset is distributed as $\epsilon_q^k \sim N(b_q^k, v_q^k)$. In general, b_q^k is not zero, however, in our case b_q^k is assumed to be zero because we performed bias correction as described in the previous subsection, and v_q^k are calculated from the hexagonal-cell-specific mean and variance estimates (see Appendix for details).

Our fused estimate for a region centered at location B_0 is a linear combination of $\mathbf{Z}_1$ and $\mathbf{Z}_2$. That is,

$$\hat{Y}(B_0) \;\; = \;\; \mathbf{a}_1'\mathbf{Z}_1 + \mathbf{a}_2'\mathbf{Z}_2, \tag{2}$$

where $\mathbf{a}_1$ and $\mathbf{a}_2$ are N_1 and N_2 dimensional vectors, respectively. These vectors are unknown and are estimated in a way that minimizes the expected squared error relative to the true observable. That is, we choose $\mathbf{a}_1$ and $\mathbf{a}_2$ to minimize,

$$\begin{aligned} E((Y(B_0) - \hat{Y}(B_0))^2) &= \mathrm{Var}(Y(B_0) - \mathbf{a}_1'\mathbf{Z}_1 - \mathbf{a}_2'\mathbf{Z}_2) \\ &= \mathrm{Var}(Y(B_0)) - 2\mathbf{a}_1'\mathrm{Cov}(\mathbf{Z}_1, Y(B_0)) \\ &\quad -2\mathbf{a}_2'\mathrm{Cov}(\mathbf{Z}_2, Y(B_0)) \\ &\quad -2\mathbf{a}_1'\mathrm{Cov}(\mathbf{Z}_1, \mathbf{Z}_2)\mathbf{a}_2 \\ &\quad +\mathbf{a}_1'\mathrm{Var}(\mathbf{Z}_1)\mathbf{a}_1 + \mathbf{a}_2'\mathrm{Var}(\mathbf{a}_2)\mathbf{a}_2 \end{aligned}$$

subject to the unbiasedness constraint that the elements of $\mathbf{a}_1$ and $\mathbf{a}_2$ add up to 1. That is,

$$1 = \mathbf{a}_1'\mathbf{1}_{N_1} + \mathbf{a}_2'\mathbf{1}_{N_2}, \tag{3}$$

where $\mathbf{1}_{N_k}$ is an N_k-dimensional vector of ones. The solution to the minimization problem in (3) can be found via the method of Lagrange multipliers; but it requires knowledge of the spatial covariance structure $C(B_i, B_j)$, which can be expanded in terms of the BAU covariances:

$$C(B_i, B_j) = \frac{1}{|D \cap B_i||D \cap B_j|} \sum_{\mathbf{u} \in D \cap B_i} \sum_{\mathbf{v} \in D \cap B_j} C(\mathbf{u}, \mathbf{v}). \tag{4}$$

Typically, the covariance structure in kriging-based approaches is estimated from the data, but the formulation in Equation 4 makes estimation intractable for non-linear covariance classes. We make use of the Spatial Mixed Effects model (SME; Cressie & Johannesson, 2008), which assumes that the true observable, here NSAT, can be written as the linear mixed model,

$$Y(\mathbf{s}) = \mathbf{t}(\mathbf{s})'\boldsymbol{\alpha} + \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}). \tag{5}$$

where $\mathbf{t}(\cdot) \equiv (t_1(\cdot), \ldots, t_p(\cdot))'$ is a vector of p known covariates, such as geographical coordinates or other physical variables. The vector of linear coefficients, $\boldsymbol{\alpha}$, is unknown and will be estimated from the data. The middle term captures the spatial dependence as the product of an r-dimensional vector of known spatial basis functions, $\mathbf{S}(\mathbf{s})$, and an r-dimensional Gaussian random variable, $\boldsymbol{\eta}$. Here, we assume that with $\boldsymbol{\eta} \sim N(\mathbf{0}, \mathbf{K})$. Similar to the implementation in Nguyen et al. (2012), we implement these using multi-resolution bisquare basis functions centered at different resolutions of the Inverse Snyder Equal-Area Projection Aperture 3 Hexagon (ISEA3H) type within the Discrete

Global Grid (DGGRID) software (specifically, resolutions 2, 3, and 5 of ISEA3H, for details see Sahr, 2019). The last term, $\xi(\cdot)$, describes the BAU-scale variability of the process. We assume that $\xi(\cdot)$ is an independent Gaussian process with mean zero and variance σ_ξ^2.

The SME model in Equation 5 has useful change-of-support properties, which makes computation of the spatial covariance function straightforward. In particular, Nguyen et al. (2012) shows that

$$\text{cov}(Z(B_i), Z(B_j)) = \mathbf{S}(B_i)'\mathbf{K}\mathbf{S}(B_j) + \sigma_\xi^2 \frac{|D \cap B_i \cap B_j|}{|D \cap B_i||D \cap B_j|} + v_i^k I(i = j), \tag{6}$$

where

$$\mathbf{S}(B_i) \equiv \frac{1}{|D \cap B_i|} \sum_{\mathbf{u} \in D \cap B_i} \mathbf{S}(\mathbf{u}).$$

Notice that Equation 6 allows us to express the covariance between spatial averages *explicitly* in terms of the spatial dependence parameter $\mathbf{K}$. This allows for straightforward estimation of it from footprint data.

Another advantage of the SME model is its scalability. For a general covariance structure, solving for $\mathbf{a}_1$ and $\mathbf{a}_2$ requires inverting a $(N_1 + N_2) \times (N_1 + N_2)$ covariance matrix, which has computational complexity $O((N_1+N_2)^3)$. For large datasets such as AIRS and CrIMSS where the data size is on the order of tens of thousands, this matrix inversion is computationally infeasible. However, the model in Equation 5 implies the following full covariance matrix:

$$\begin{aligned} \boldsymbol{\Sigma} &\equiv \text{var}((\mathbf{Z}^{1\prime}, \mathbf{Z}^{2\prime})') \\ &= \mathbf{S}'\mathbf{K}\mathbf{S} + \mathbf{U}, \end{aligned}$$

where $\mathbf{S}$ is a matrix constructed by appending the spatial function $\mathbf{S}(\cdot)$ over all the footprints in both datasets, $\mathbf{U}$ is the *sparse* covariance matrix for the fine-scale processes $\xi(\cdot)$, and the measurement-error processes $\epsilon^k(\cdot)$ at the given data locations (for more details, see Equation 4 of Nguyen et al., 2012). Using the Sherman-Morrison-Woodbury formula (e.g., Henderson & Searle, 1981), the matrix inverse is given by,

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}^{-1} - \mathbf{U}^{-1}\mathbf{S}'\left(\mathbf{K}^{-1} + \mathbf{S}\mathbf{U}^{-1}\mathbf{S}'.\right)^{-1}\mathbf{S}\mathbf{U}^{-1},$$

Note that the inversion above, and hence the calculation of the coefficients $\mathbf{a}_1$ and $\mathbf{a}_2$ for the fused estimate, is very fast because it only requires inversion of the *sparse* $(N_1 + N_2) \times (N_1 + N_2)$ matrix $\mathbf{U}$, which is typically very sparse, and inversion of $\mathbf{K}$ and $(\mathbf{K}^{-1} + \mathbf{S}'\mathbf{U}^{-1}\mathbf{S})$, both of which are $r \times r$ matrices ($r << N_1 + N_2$).

The methodology described in this section is a scalable variant of Gaussian process prediction (Cressie, 2015). It has been applied to fusion of total column CO_2 concentration (XCO2) from AIRS and OCO-2 and aerosol optical depth from MISR and MODIS (Nguyen et al., 2012, 2014). Hammerling et al. (2012) used another variant called local kriging to produce Level 3 estimates of XCO2 from the GOSAT instrument.

There are two important advantages of Gaussian process prediction over other approaches currently in use such as binning or nearest neighbor interpolation. First, our fused estimates are *best linear unbiased estimates.* That is, the standard errors are guaranteed to be the smallest possible because the estimates are derived through an algorithm that minimizes errors relative to the unknown true process. Such estimates are called best linear unbiased estimates, and are optimal in that sense. It is easily shown that within the class of linear estimators, this method produces the smallest prediction errors. The second advantage is that SSDF provides a statistically principled method for estimating uncertainties (that is, $\text{Var}(\hat{Y}(B_0) - Y(B_0))$). Quantifying and minimizing uncertainties in this manner is crucial for creating data products for scientific analyses that involve making inferences about geophysical observables.

2.6 Dataset preparation for validation

We validate our SSDF product using a randomly chosen reserved 1% of the ISD dataset. We match up SSDF, AIRS, CrIMSS, and ERA5 estimates to withheld ISD data using a 100 km and 1 hour matchup criterion (see Section 2.4 for more detail). This matchup procedure generates multiple paired datasets: ISD-AIRS, ISD-CrIMSS, ISD-SSDF, and ISD-ERA5. These matchup datasets might differ in their coverage; for instance, an SSDF estimate might be matched to an ISD observation at a location where there are no nearby AIRS or CrIMSS estimates. Therefore, to mitigate the effect of biases due to differing spatial and temporal coverage in these matchup pairs, we also require that SSDF estimates are also close to (within the same matchup distance and time) at least one datum from the comparison dataset. This allows us to compare, for example, AIRS and SSDF(AIRS) datasets which have the same number of samples, all of which are collocated in space and time within the matchup criterion.

The choices of a 1% test ISD dataset and this matchup scheme results in over 4000 AIRS-SSDF sample pairs and over 13,000 CrIMSS-SSDF sample pairs for 2013, a typical year.

3 Results

3.1 SSDF product overview

We produced fused NSAT using two satellite input datasets over North America between 25 N and 50 N, from November 28 2012, when CrIMSS-CLIMCAPS first becomes available, through the end of 2020. During this time period, there were 34 days and 36 nights with no AIRS data (approximately half of which occurred in 2020), and 24 days and 28 nights with no CrIMSS-CLIMCAPS data. In the cases with only one input satellite dataset, the SSDF product is created from only the single dataset, thus creating a continuous record. There was one day/night period (November 7, 2020) without either AIRS or CrIMSS-CLIMCAPS data; we did not create SSDF product for this day.

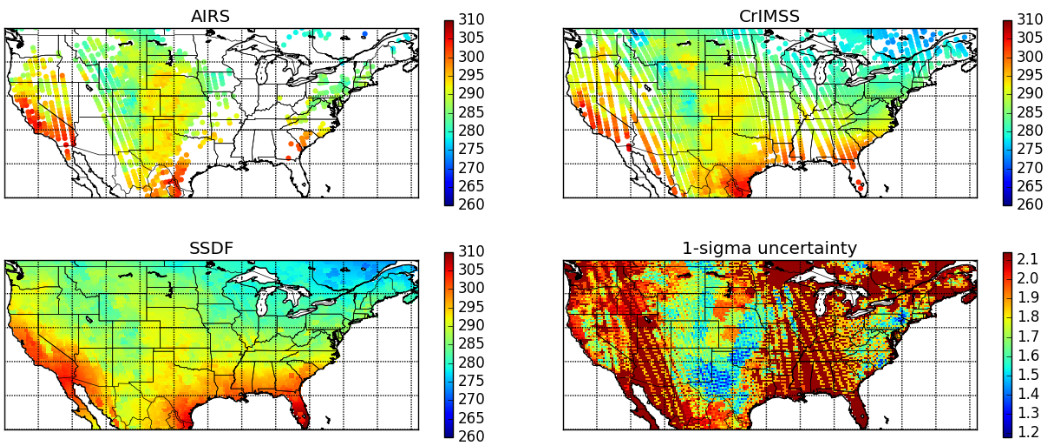

Figure 2: Sample data fusion satellite NSAT inputs, SSDF NSAT results, and uncertainty estimates for 2015 October 31, day. The top two plots show maps of the input satellite NSAT data ingested into the SSDF product, with AIRS on the left and CrIMSS on the right. The bottom-left plot shows the SSDF fusion results. The bottom-right plot shows the uncertainty estimates on the SSDF fusion results at the 1-sigma level. All units are Kelvin.

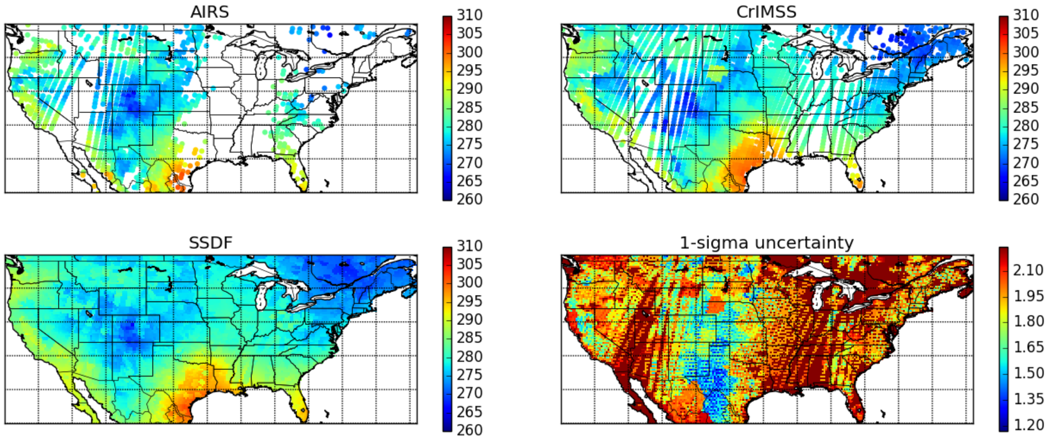

Figure 3: Same as Figure 2 but for night. All units are degrees K.

Figures 2 and 3 provide maps representing one arbitrarily chosen day and night of the SSDF product. For both the day and night cases, the top two plots show maps of the input satellite data ingested into the SSDF product, with AIRS on the left and CrIMSS on the right; the bottom left plot shows the SSDF fusion results; and the bottom right plot shows the uncertainty estimates on the SSDF fusion results at the 1-sigma level. These sample maps demonstrate how our SSDF method fills in missing data in the input datasets by exploiting spatial correlations to provide a complete gap-filled, gridded product. They also provide a first look at the SSDF uncertainty estimates. Note that the estimated uncertainties are higher in regions that contain no observations, contain observations from only a single input dataset, or in which the two input datasets have relatively poor agreement.

3.2 Bias, standard deviation, and RMSE comparison

We now turn to validation against withheld ISD reference data to quantify improvement in the SSDF products. We examine bias, standard deviation, and RMSE, calculated from the withheld matchups, of AIRS, CrIMSS, ERA5-Land, and the corresponding matched SSDF data. In what follows, analyze daytime and nighttime separately, as daytime and nighttime biases differ significantly.

We first show maps of bias, RMSE, and standard deviation relative to the 1% of withheld (testing-only) ISD reference data, based on the matchups aggregated into the hexagonal bins. Figure 4 shows maps of bias (retrieval - ISD) for AIRS, CrIMSS, and SSDF, for the 2013-2020 period in total, and for day-only and night-only. Individual bias estimates for retrieval-ISD pairs are aggregated into 2-degree hexagonal cells.

Overall, in the mean over CONUS and over the entire time period, SSDF provides a reduction in the magnitude of daytime bias of 1.7 K and 0.5 K relative to AIRS and CrIMMS, respectively. At night, SSDF is essentially unbiased in the mean over the domain and provides a reduction in the magnitude of bias of 1.5 K and 0.2 K relative to AIRS and CrIMMS, respectively.

AIRS shows a strong cold bias in daytime over the mountainous West, which is also present in CrIMSS, although less severe. AIRS shows a near-constant warm bias over the entire Eastern CONUS at night, while CrIMSS shows a sharp warm bias over small regions of the mountainous West at night. SSDF corrects all of these biases (through the bias-correction procedure described above) and produces estimates with lower biases than either of its input satellite data sets over the domain.

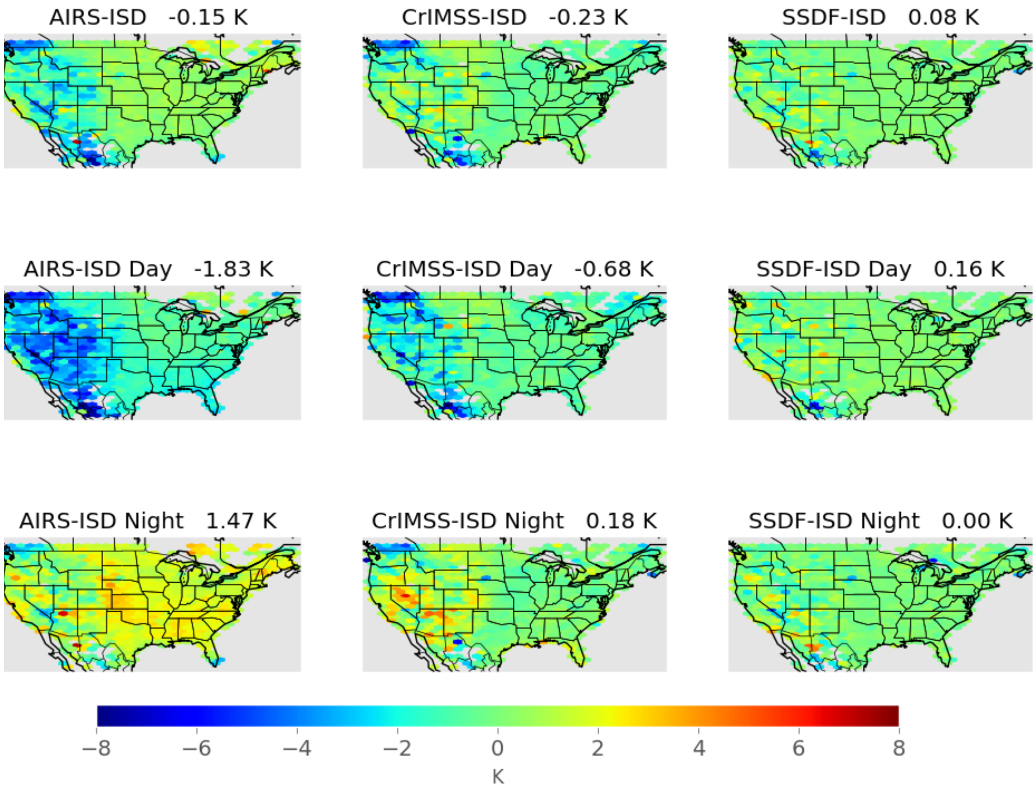

Figure 4: Maps of bias (retrieval - ISD) over the product period of 2013-2020, created against the withheld ISD test data, for AIRS (first column), CrIMSS-CLIMCAPS (second column) and SSDF (third column), for both day and night together (top row), for day only (second row) and for night only (third row). Individual bias estimates for retrieval-ISD matchup pairs are aggregated over 2-degree hexagonal cells. The mean bias over CONUS for the entire time period is shown in the title for each map.

Figures 5 and 6 show maps of standard deviation and RMSE for AIRS, CrIMSS and SSDF, for the 2013-2020 period, and for daytime only and nighttime only. Standard deviation and RMSE tell a similar story to that of bias. Overall, in the mean over CONUS and over the entire time period, SSDF provides a reduction in RMSE of 35% and 15% compared to AIRS and CrIMSS, respectively.

CrIMSS has high RMSE over the mountainous West in both day and night, but low RMSE over the eastern two-thirds of the continent. Similarly, AIRS has relatively high RMSE over the entire domain, but especially over the mountainous West. Mountainous regions pose particular challenges for remote sensing of surface quantities, and of NSAT in particular, which can vary greatly depending on e.g., north-facing versus south-facing mountain surfaces. Furthermore, variations in topographic features between ISD stations and their matched remote sensing retrievals can lead to random errors, increasing RMSE and variance estimates. However, SSDF NSAT shows a clear decrease in bias over all regions, including in the mountainous western CONUS, although there is potential for improvement in the SSDF product over the West.

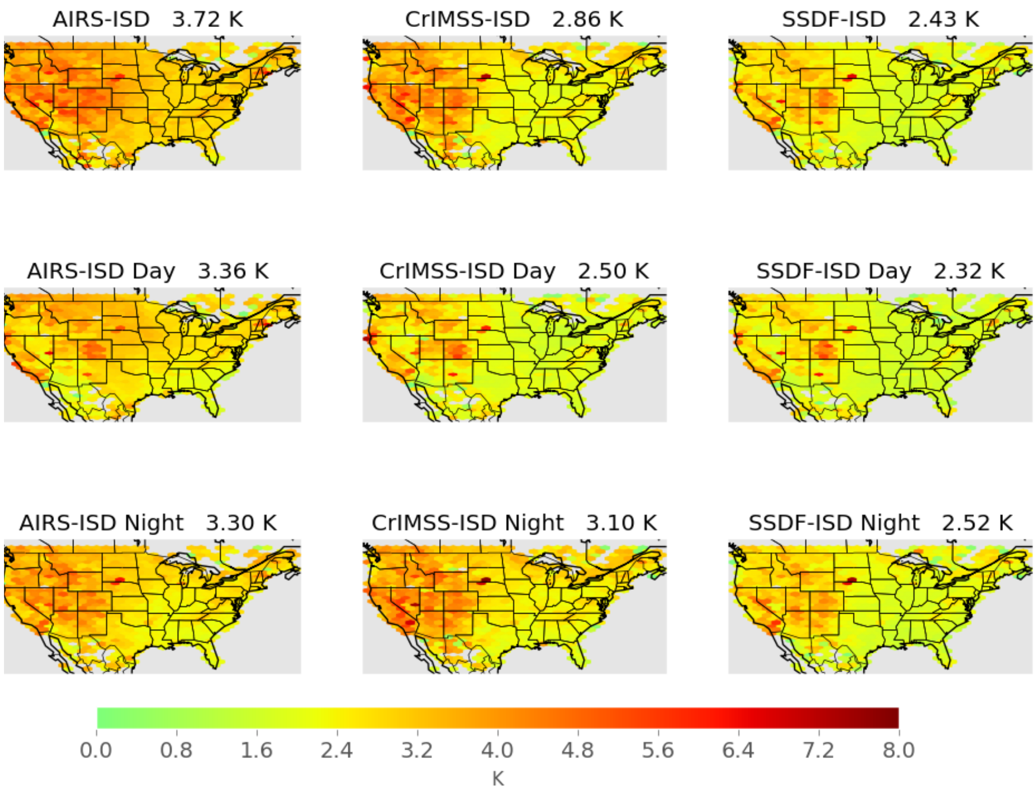

Figure 5: Standard deviation maps. The nine panels are similar to those in Figure 4 but for standard deviation.

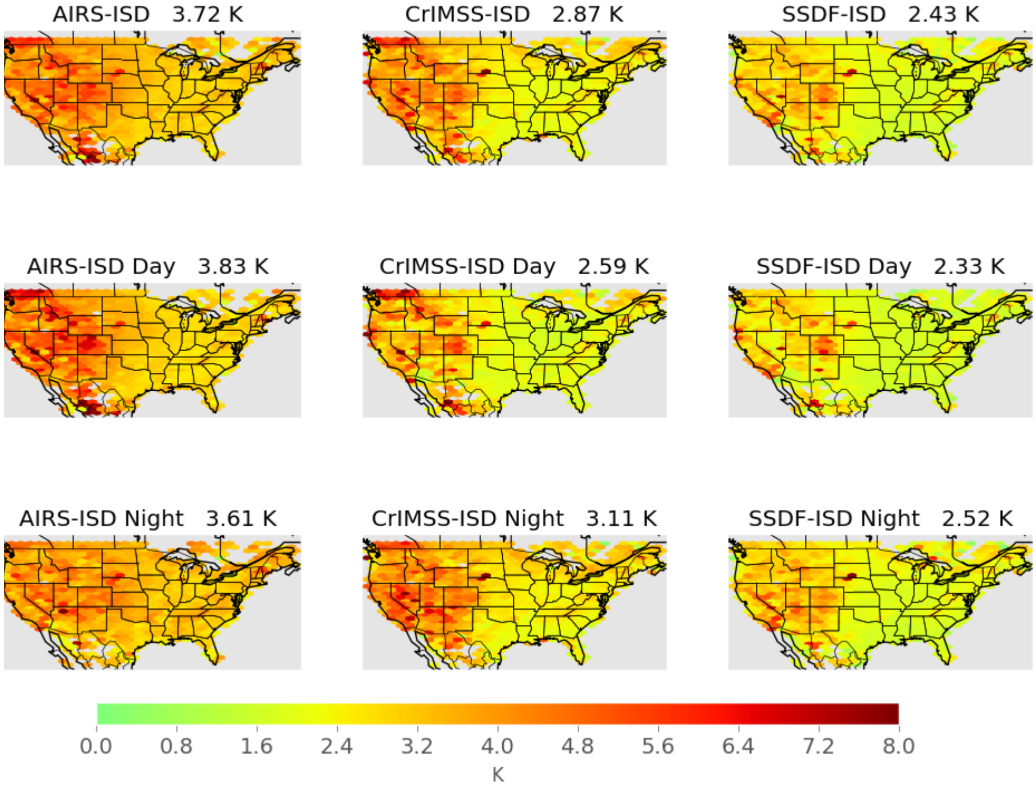

Figure 6: RMSE maps. The nine panels are similar to those in Figure 4 but for RMSE.

We repeated this analysis over CONUS and the 2013-2020 period for the SSDF product created with AIRS alone, without CrIMSS. We found similar improvements in bias, standard deviation, and RMSE. The mean bias of the AIRS-only SSDF product over the entire domain was -0.08 K for daytime only, and -0.03 K for nighttime only. The overall RMSE was 2.52 K, 4% higher than the overall RMSE of the SSDF product created from both AIRS and CrIMSS.

Figure 7 shows histograms of the NSAT error (retrieval/reanalysis - ISD) for the year 2013, over CONUS only. The three comparison datasets (AIRS, CrIMSS, and ERA5-Land) were matched separately to SSDF outputs, to ensure that the SSDF product and each corresponding comparison dataset are considering the same scenes. The SSDF error histograms are symmetric with a single mode and peak at 0 for both day and night, which is consistent with the errors being unbiased. The AIRS histogram exhibits a cold bias during the day and a warm bias at night. CrIMSS has a similar day/night bias shift, but of a smaller magnitude. A cold bias over land, particularly at higher temperatures, has been previously noted for both input datasets (Yue et al., 2020, 2021), although there have been few validation studies (Ferguson & Wood, 2010; Sun et al., 2021). The SSDF product exhibits smaller mean biases and RMSEs than either input dataset. On average, over both input datasets, daytime and nighttime, SSDF decreases mean bias magnitude by 81% and mean RMSE by 23% relative to the input datasets.

Next, we examine the seasonality of bias and RMSE. Figure 8 shows the mean bias (retrieval/reanalysis – ISD) by month split into day/night to examine seasonality. There is a significant cold bias during the day for AIRS and CrIMSS that switches to a warm bias at night. During the day, AIRS has a smaller bias during winter months (Dec/Jan/Feb) and a larger bias during summer months (Jun/Jul/Aug). This is switched during nighttime where a larger warm bias is observed during winter and a smaller warm bias is observed during summer. These AIRS biases are of course also apparent in Figure 7. The SSDF product is relatively unbiased for both day and night. The SSDF bias magnitude is slightly larger during the day than night. From May to December, the SSDF product has a smaller bias at night than does ERA5-Land while during the day the reanalysis and the SSDF mean biases are of similar magnitude.

Figure 9 shows mean RMSE (retrieval/reanalysis – ISD) by month split by day/night, i.e., the mean RMSE values calculated in 2-degree spatial bins. RMSE is largest for AIRS, particularly during the day. Generally, RMSE is higher in winter and lower in summer. During the day, the ERA5-Land has the lowest RMSE. At night, the SSDF RMSE is comparable and sometimes lower than the ERA5-Land RMSE.

We next examine relative performance in hot and cold extremes. Figure 10 shows the mean bias (retrieval/reanalysis – ISD) by ISD percentile of the ISD matchups. The error bars are the standard error of the mean at the 95 percent confidence level. The lighter shade of every color is the matched SSDF corresponding to the comparison dataset. All retrievals and reanalysis do best in the mean state (25th to 75th percentile). At the extremes, each of the datasets being compared to ISD have warm biases for low values (1st through the 15th percentile) and cold biases for high values (85th through the 99th); in other words, these datasets dampen out capture cold or warm extremes represented in the ISD. The SSDF product captures the extremes better than the input datasets, AIRS and CrIMSS. However, the reanalysis generally does best, having the smallest bias regardless of percentile, and is better at capturing the extremes.

We next examine performance at extremely high elevations. Figure 11 shows mean biases (retrieval/reanalysis – ISD) aggregated by ISD elevation. At around 2500 meters, mean biases increase with elevation in the SSDF product, AIRS, CrIMSS, and reanalysis. Daytime mean biases at these high elevations are larger in SSDF, although

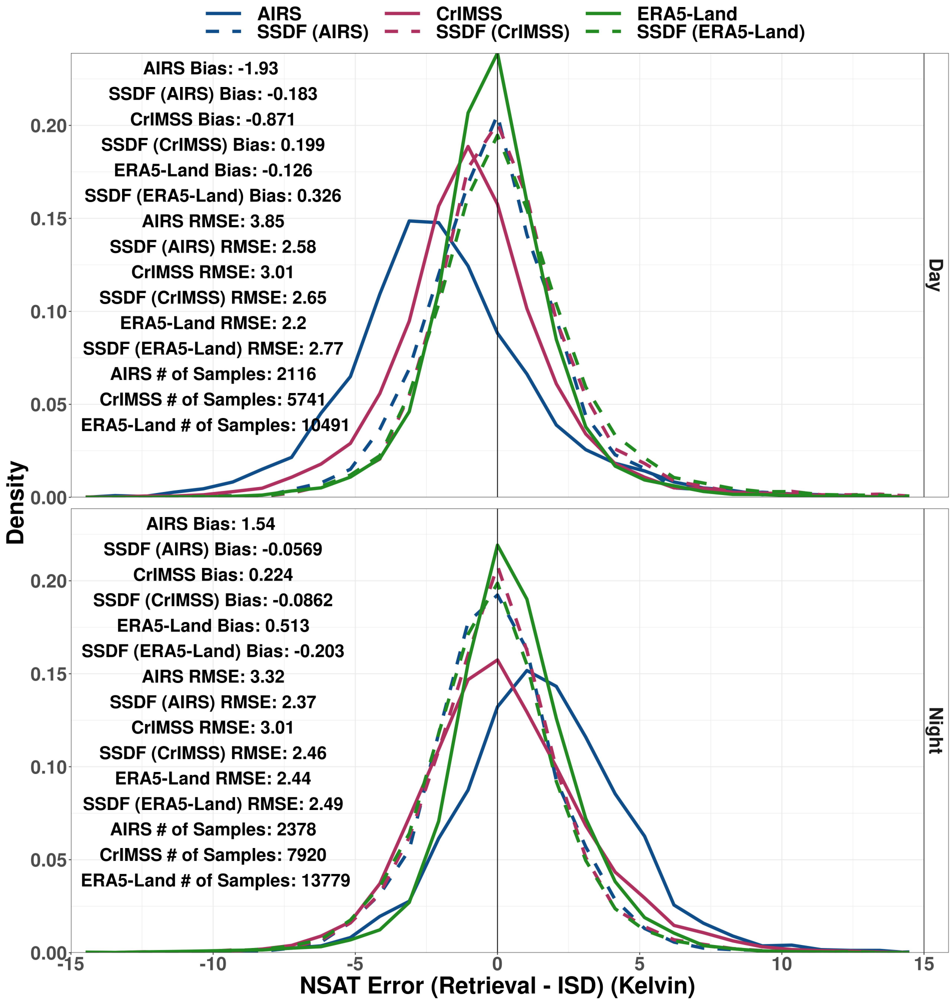

Figure 7: Histograms of errors for day (top) and night (bottom) for 2013 over CONUS, for AIRS (blue), CrIMSS (red) and ERA5-Land (green). The dashed line is the SSDF subset matched to the other datasets. Mean statistics of bias, RMSE, and the number of samples are provided.

we note that the sample size is small. At night, SSDF shows lower mean biases than AIRS, CrIMSS, or ERA5-Land at high elevations.

In order to increase the sample size for high-elevation cases, Figure 12 shows the mean biases aggregated by ISD elevation for elevations higher than 2000 meters over the period 2012-2020. During the day, the SSDF bias exceeds AIRS and CrIMSS, consistent with Figure 11. We hypothesize that this excess bias in SSDF for a very small number of data points at very high elevations is caused by the bulk-binning method for bias estimation. As Figure 11 shows, both remote sensing datasets exhibit a cold bias during the daytime at lower elevations. Because the two-degree hexagonal bins for bias estimation are dominated by lower elevations (as the problematic high elevations are high mountain surfaces), and because both remote sensing dataset biases switch signs from cold bias to warm bias at approximately 2500 m, the cold bias correction calculated from the bulk bins ends up exacerbating the warm bias from the input datasets at the highest elevations. In a future version of SSDF, we will improve the

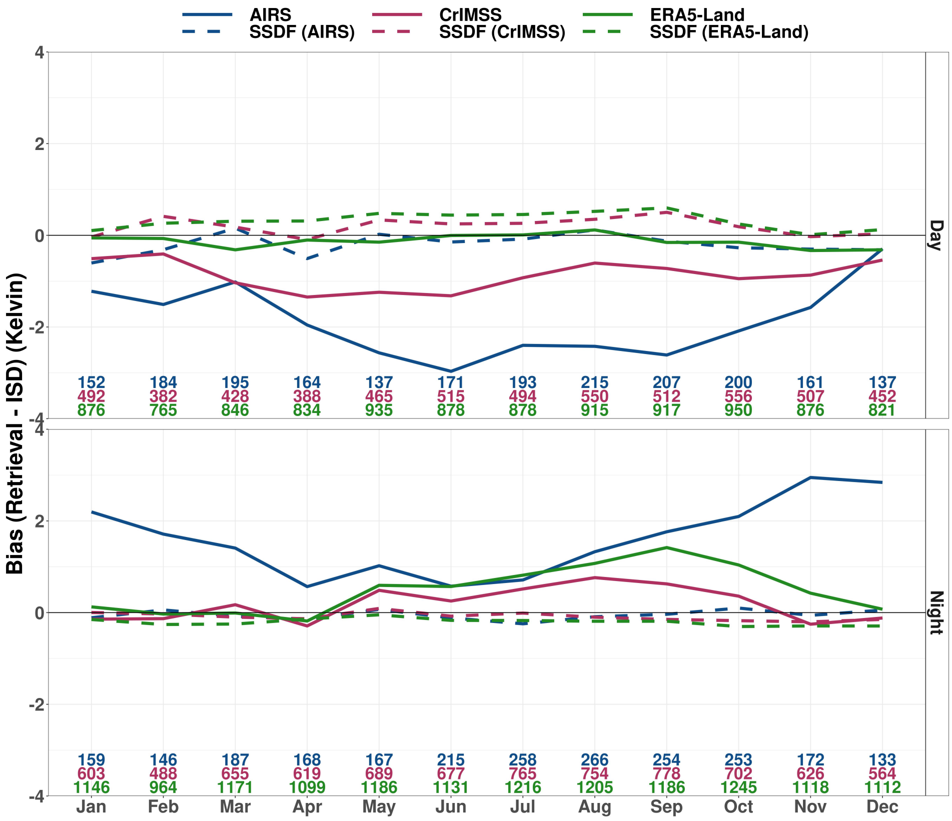

Figure 8: Mean bias as a function of month for day (top) and night (bottom) for 2013 over CONUS. Numbers at the bottom indicate the number of data points, and are color-coded according to dataset.

bias estimation of the input datasets, which could mitigate or eliminate this bias at the very small number of estimates elevations above 2500 m.

Figure 9: Mean RMSE as a function of month for day (top) and night (bottom) for 2013 over CONUS. Numbers at the bottom indicate the number of data points, and are color-coded according to dataset.

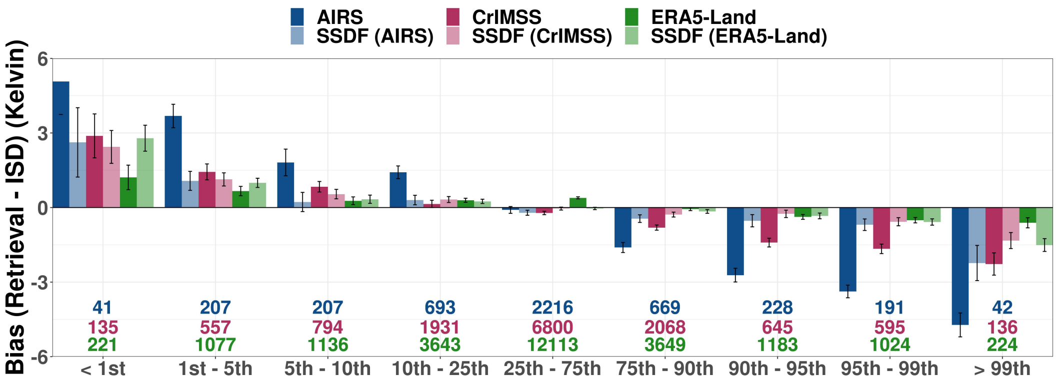

Figure 10: Mean biases as a function of ISD percentile for 2013 over CONUS. Numbers at the bottom indicate the number of data points, and are color-coded according to dataset.

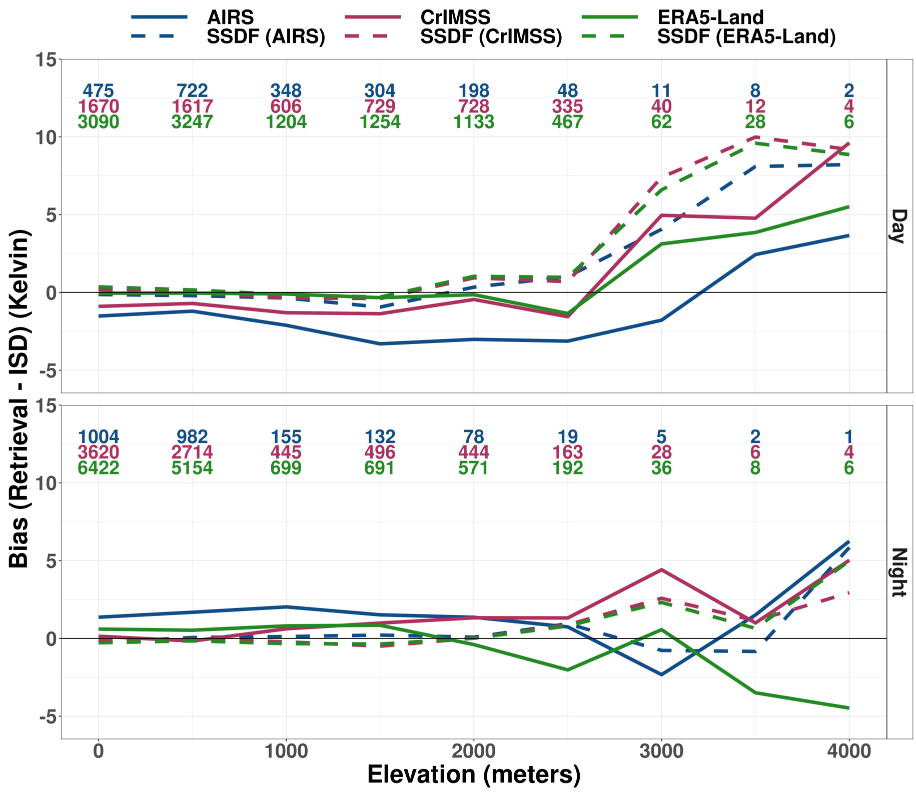

Figure 11: Mean biases as a function of ISD elevation for day (top) and night (bottom) for 2013 over CONUS. Numbers at the top indicate the number of data points, and are color-coded according to dataset.

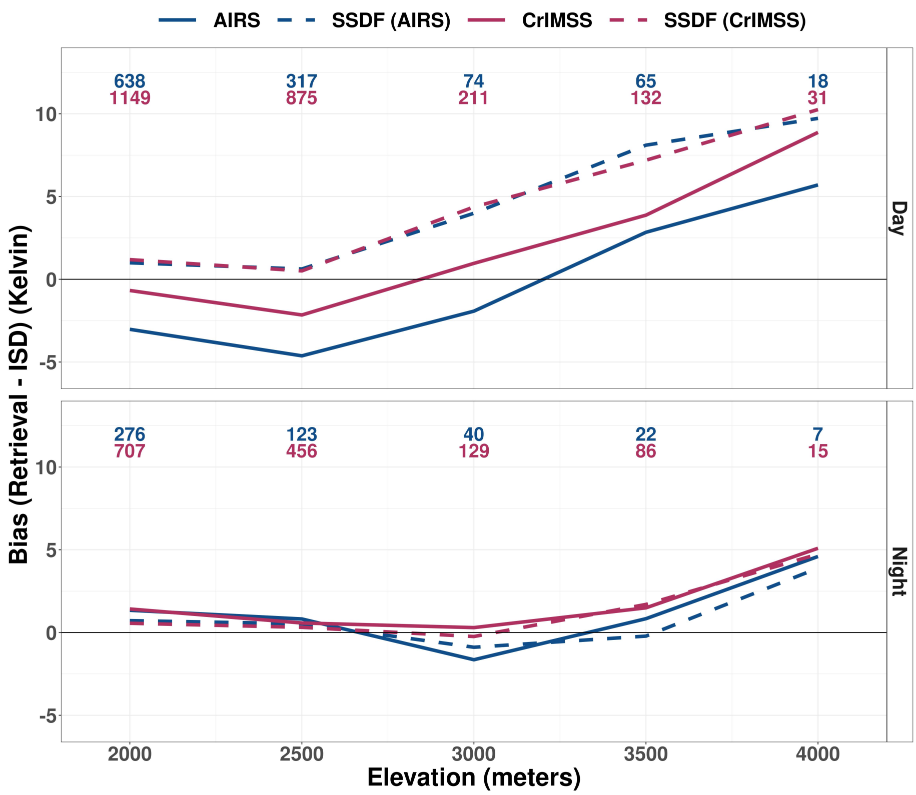

Figure 12: Mean biases as a function of ISD elevation for day (top) and night (bottom) over CONUS from 2012-2020 for AIRS, CrIMSS, and SSDF. Numbers at the top indicate the number of data points, and are color-coded according to dataset.

3.3 Validation of uncertainty estimates

The SSDF algorithm provides a mean (prediction/estimate) and standard deviation (uncertainty) of the conditional distribution of true NSAT, given the available inputs; this distribution is termed the predictive distribution. In what follows, this is a Gaussian distribution, centered at the SSDF estimate. This information can be used to construct prediction intervals for the true NSAT. Here we provide a summary and probabilistic assessment of the SSDF predictive distribution along with related information from the AIRS V7 and CrIMSS-CLIMCAPS V2 products. In the notation that follows, we use the subscript i in place of the areal unit notation B_i.

- In addition to each SSDF NSAT estimate, $\hat{Y}_i$, the algorithm also provides the conditional standard deviation of the predictive distribution, denoted $\hat{\sigma}_{\hat{Y},i}$.
- The AIRS V7 NSAT retrieval, $Z_{1,i}$, is accompanied by a corresponding uncertainty estimate, denoted $\hat{\sigma}_{Z,1,i}$ (Susskind et al., 2014). This estimate results from a regression model for predicting the absolute retrieval error given several predictors available from the retrieval.
- The CrIMSS-CLIMCAPS V2 retrieval, $Z_{2,i}$, also has a corresponding uncertainty estimate, denoted $\hat{\sigma}_{Z,2,i}$ (N. Smith & Barnet, 2020). This estimate results from a linear approximation of the posterior standard deviation of the true state given the observed radiances for a single footprint and is an output of the optimal estimation (OE) approach used in CLIMCAPS.

Figure 13 shows histograms of these uncertainty estimates: $\hat{\sigma}_{Z,1}$, $\hat{\sigma}_{Z,2}$, and $\hat{\sigma}_{\hat{Y}}$ across the CONUS data record. The solid line shows uncertainty estimates from AIRS (blue) and CrIMSS (red) while the dashed shows the corresponding matched SSDF uncertainty estimates. CrIMSS has a peak around 1.2 K with a narrow distribution; AIRS V7 has a peak between 1.5 and 2 K with a wide distribution. SSDF uncertainty histograms peak around 2 K.

These uncertainty estimates are properties of distributions, whereas we define error e_i as a realization of a random variable that represents the difference between an estimate and the true state. For example, the error for SSDF is $e_{\hat{y},i} = \hat{Y}_i - Y_i$, where Y_i is the ISD validation for colocation i. If the predictive distribution is assumed to be Gaussian, the empirical coverage of intervals of the form

$$\hat{Y}_i \pm c\,\hat{\sigma}_{\hat{Y},i},$$

can be assessed for the ISD matchups. In the case of an unbiased estimate, "well-calibrated" uncertainty estimates, and a Gaussian distribution; intervals with $c = 1$ should cover the true state Y_i about 68% of the time, and about 95% of the time for $c = 2$.

Figure 14 shows scatterplots of the joint distribution of the uncertainty estimate (x-axis) and the observed error (retrieval-ISD). There are many cases for AIRS and CrIMSS where the uncertainty estimate grossly underestimates the true error; over 15% of the time for both datasets and for day and night, the true error is more than three times greater than the uncertainty estimate. However, this occurs about 3% of the time with SSDF in the day and fewer than 5% of the time at night. Overall, the CrIMSS uncertainty estimates are distributed too narrowly, and with a peak too low, to capture the true error. The AIRS uncertainty estimates also peak at a value below the peak of the error distribution, although the uncertainty estimate distribution is much wider, including a very long tail of high uncertainty estimates.

In general, SSDF uncertainty estimates are consistent with statistical expectations under Gaussian assumptions. For example, one would expect one-sigma uncertainty estimates to cover a standard error distribution 68% of the time, and we see

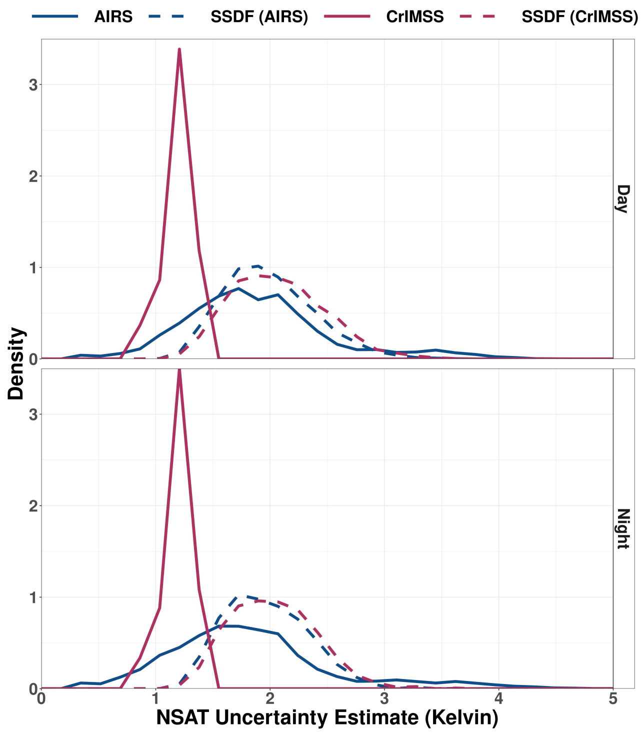

Figure 13: Histograms of uncertainty estimates for day (top) and night (bottom) for 2013 over CONUS.

that the SSDF uncertainty estimates do so roughly 65% of the time in daytime. Similarly, one would expect the estimates to cover 95% and over 99% at the 2- and 3-sigma levels, with SSDF covering about 90% and 97% during daytime.

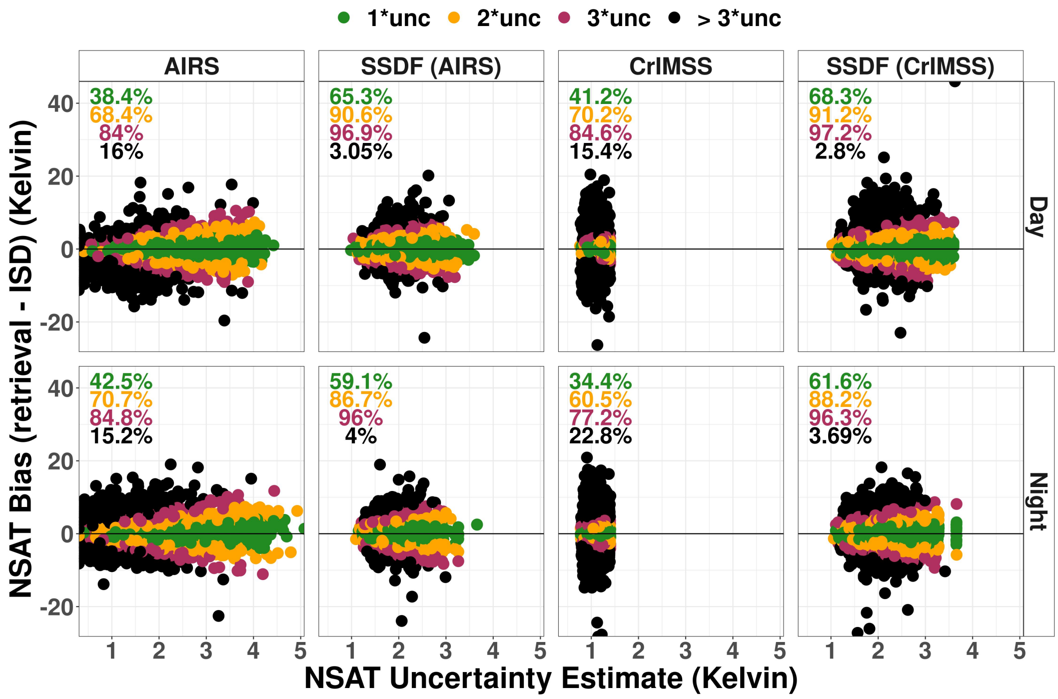

Figure 14: Observed errors (retrieval - ISD) versus uncertainty estimates for day (top) and night (bottom) for 2013 over CONUS. The colors show whether the range of each observed error was within the uncertainty bound, as described in the text: 1×uncertainty (green, should cover the true state about 68% of the time), 2×uncertainty (orange, should cover the true state about 95% of the time), 3×uncertainty (red, should cover the true state about 99% of the time) or > 3×uncertainty (black).

3.4 Empirical distribution consistency

The ISD record provides a sample of the empirical distribution of NSAT over CONUS. Here, we assess the relative consistency of the SSDF empirical distribution versus the other products against the ISD reference distribution. Figure 15 shows an example of the empirical cumulative distribution (ECDF) for the ISD (pink) and AIRS (blue). While it is almost certainly the case that the products' ECDFs deviate from the ISD reference distribution in some subtle ways, we evaluate their relative consistency with ISD through a series of hypothesis tests. Figure 16 shows the difference between the ECDF of the retrieval/reanalysis to the ECDF of ISD. The AIRS ECDF has the largest difference to the ISD ECDF, particularly during the Day.

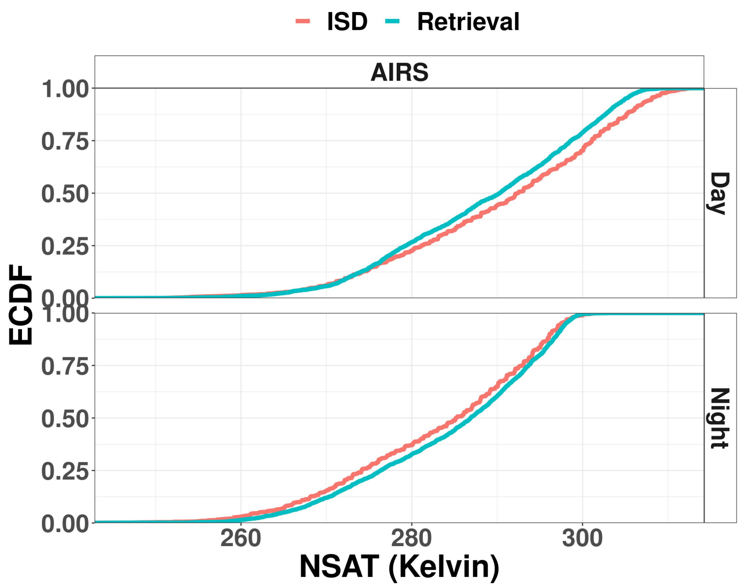

Figure 15: ECDF for AIRS (blue) and ISD (pink) for day (top) and night (bottom) for 2013 over CONUS.

The SSDF estimates are tested against each of the other products (AIRS, CrIMSS, ERA5-Land) for night and day conditions. Each assessment is carried out using a randomization or resampling test (Wilks, 2006). For this test, the null hypothesis is that the empirical distributions of SSDF and the comparison product deviate equally from the ISD reference distribution. The alternative hypothesis is that either SSDF or the comparison product have an empirical distribution that is closer to the ISD reference distribution. For this procedure, the test statistic is computed as the difference in two-sample Kolmogorov-Smirnov (KS) statistics for the products versus ISD.

For each instance of the test, we have a collection of matched triples $\{\hat{\mathbf{Y}}, \mathbf{Z}_k, \mathbf{Y}\}$; where $\hat{\mathbf{Y}} \equiv \{\hat{Y}_i\}; i = 1, \ldots, n$ are the SSDF estimates, $\mathbf{Z}_k \equiv \{Z_{k,i}\}; i = 1, \ldots, n$ are the comparison products, and $\mathbf{Y} \equiv \{Y_i\}; i = 1, \ldots, n$ are the ISD NSAT. As above, $k = 1$ for AIRS, $k = 2$ for CrIMSS, and here $k = 3$ for ERA5-Land. Then, test k has a test statistic

$$\gamma_k = \delta(\hat{\mathbf{Y}}, \mathbf{Y}) - \delta(\mathbf{Z}_k, \mathbf{Y}),$$

where δ is the traditional two-sample KS statistic. The KS statistic is the maximum difference in the two ECDFs being compared. Thus, the test statistic γ_k for the current test is a *difference* of ECDF deviations. A negative value is an indication that the SSDF distribution is closer to ISD than the comparison product.

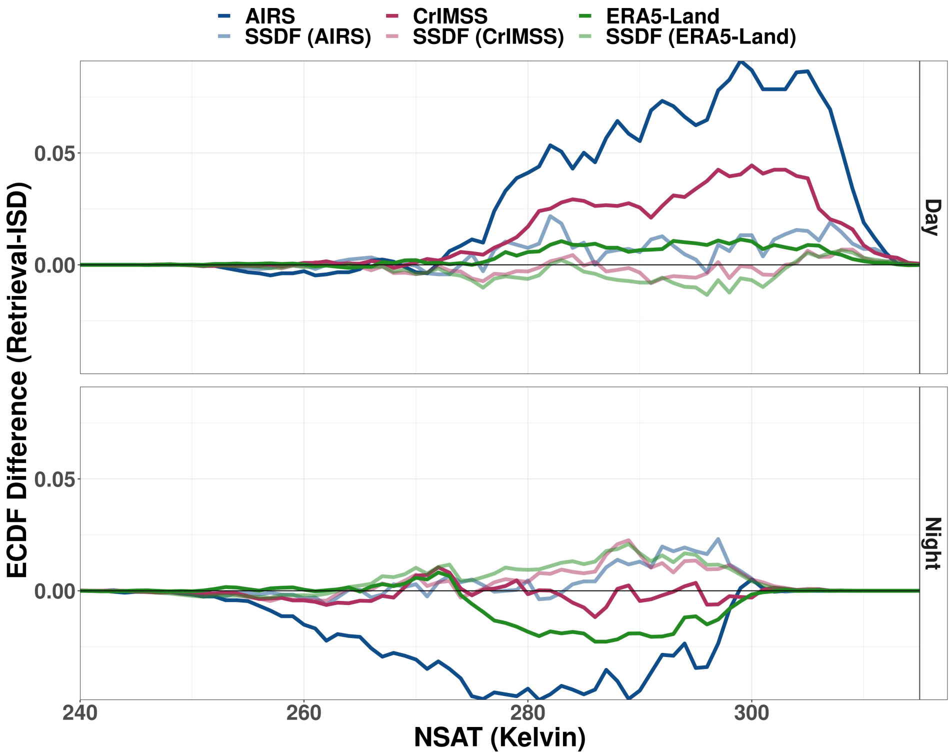

Figure 16: The ECDF difference between the retrieval/reanalysis and the ISD color coded for day (top) and night (bottom) for 2013 over CONUS.

The distribution of the test statistic under the null hypothesis can be established through a resampling procedure. The procedure should preserve the inherent dependence of the matched triples, but the assignment of the two comparison groups can be shuffled randomly. A null distribution is generated by repeating these steps $m = 1, \ldots M$ times:

1. Define shuffled data vectors $\mathbf{W}_{m,1}$ and $\mathbf{W}_{m,2}$.
2. For each validation matchup ($i = 1, \ldots, n_k$), assign $W_{i,m,1} = \hat{Y}_i$ and $W_{m,2,i} = Z_{k,i}$ with probability 0.5; otherwise assign $W_{m,1,i} = Z_{k,i}$ and $W_{i,m,2} = \hat{Y}_i$. This effectively shuffles the labels for SSDF and the comparison product for each matchup.
3. Compute the test statistic for the randomized samples,

$$\gamma_{0,m,k} = \delta(\mathbf{W}_{m,1}, \mathbf{Y}) - \delta(\mathbf{W}_{m,2}, \mathbf{Y}),$$

The distribution of $\gamma_{0,m,k}$ provides the null distribution of the test statistic for each test. Figure 17 displays the test statistics γ_k along with density plots of the null distributions of test statistics $\gamma_{0,m,k}$ for $M = 20,000$ resampled datasets for each test. A two-sided p-value can be computed for each test as

$$p_k = \frac{1}{M} \sum_{m=1}^{M} I_\gamma(|\gamma_{0,m,k}| > |\gamma_k|),$$

where I_γ is an indicator function.

The p-values for each of the resampling tests of SSDF versus other products are displayed as text in Figure 17. All tests, except the night comparison of SSDF and CrIMSS, yield p-values of 0, indicating a significant difference in consistency with the ISD reference distribution. These results can also be seen visually as the observed test statistics γ_k, shown as vertical lines, lie well outside the corresponding null distributions. The tests indicate SSDF is more consistent with ISD than AIRS for both day and night conditions, as well as a favorable result for SSDF versus CrIMSS for day and versus ERA5-Land at night. The positive test statistic for SSDF versus ERA5-Land during the day indicates the reanalysis is more consistent with ISD in this case.

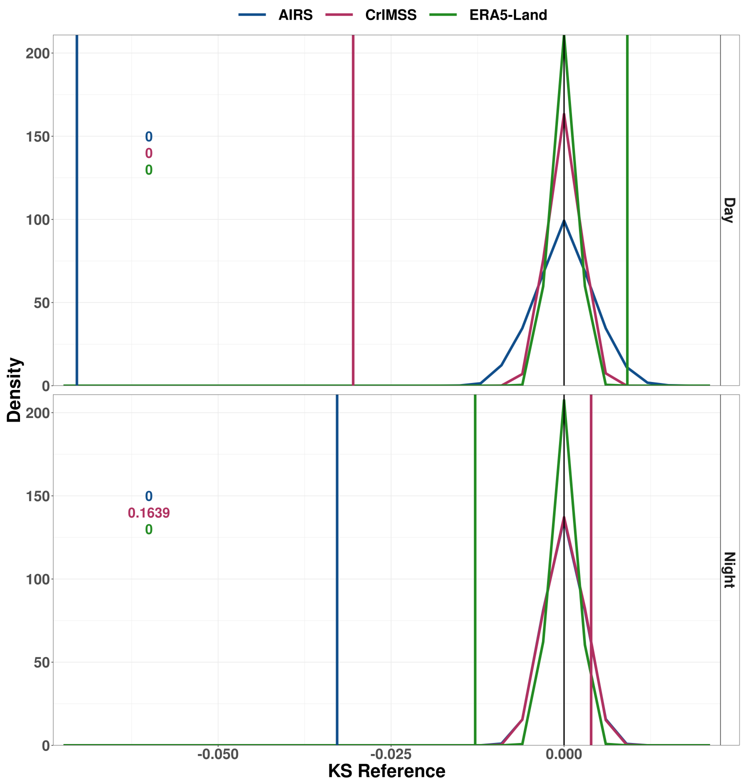

Figure 17: Histogram of the KS statistic for AIRS (blue), CrIMSS (maroon) and ERA5-Land (green), for day (top) and night (bottom) for 2013 over CONUS. The corresponding p-value is color-coded on the left side.

3.5 Long-term stationarity

We next assess the stationarity in the bias of the SSDF dataset. First, we examine the annual mean bias over the entire record relative to the withheld ISD reference data. Figure 18 shows the annual mean bias (both day and night) for both the input datasets, as well as for two periods of SSDF: the pre-CrIMSS period (2003 to 2011, inclusive) and the post-CrIMSS period (2013 to 2020, inclusive). Shading shows two standard deviations of these annual bias estimates, with the two SSDF periods calculated separately. We exclude 2002 as this year only includes 4 months of AIRS data, and we exclude 2012 as this year was a mixture of AIRS-only and AIRS-plus-CrIMSS.

These summary data clearly show that SSDF significantly improves both the mean annual bias, and the standard deviation in mean annual bias, relative to the input datasets. The mean of these annual bias estimates are -0.10°C, -0.23°C, and 0.02°C for AIRS, CrIMSS, and SSDF respectively, from 2003 to 2020 inclusive for AIRS and SSDF and from 2013 to 2020 inclusive for CrIMSS. However, these data also suggest a step change in SSDF mean annual bias in the pre-CrIMSS and post-CrIMSS period. The mean of the SSDF mean annual bias estimates in the pre-CrIMSS and post-CrIMSS periods are -0.020°C and 0.076°C, respectively, a shift of about 0.1°C. This shift is small compared to the biases in the input remote sensing datasets, and the apparent downward trend in the AIRS dataset. Over-correction with the addition of the CrIMSS dataset might be an artifact of the bias estimation bulk-binning procedure. This small step change in bias does not occur in the AIRS-only SSDF product over the full AIRS record. Future versions of SSDF will use improved uncertainty quantification methods to estimate input dataset biases, which could mitigate or eliminate this small shift in annual mean bias in transitioning from the AIRS-only SSDF product to the two-instrument product. In the meantime, the first version of our product creates a more coherent and stable climate record than the two input datasets taken separately.

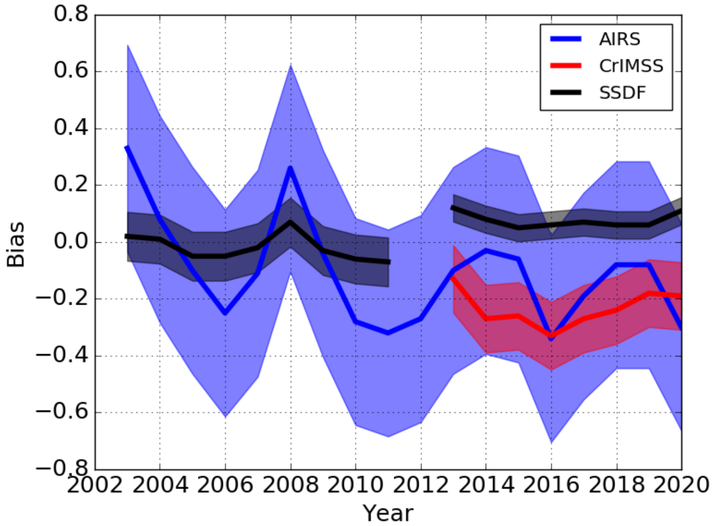

Figure 18: Annual mean bias for each year of the data record, for the SSDF product and each of the two remote sensing input products, relative to the withheld ISD data. Shading shows two standard deviations of these annual bias estimates, with the two SSDF periods calculated separately.

Figure 19 shows the histogram of the SSDF uncertainty estimates for 2011 (black) and 2013 (red). The mean uncertainty is provided as text. The histograms are comparable. The 2011 (single instrument only) histogram is shifted slightly to the right suggesting higher uncertainty estimates with one instrument compared to two. Indeed, the mean SSDF uncertainty estimate is 2.15/2.21 (Day/Night) during 2011 and decreases to 2.12/2.09 in 2013. However, this is to be expected as the additional information from CrIMSS provides greater certainty for SSDF.

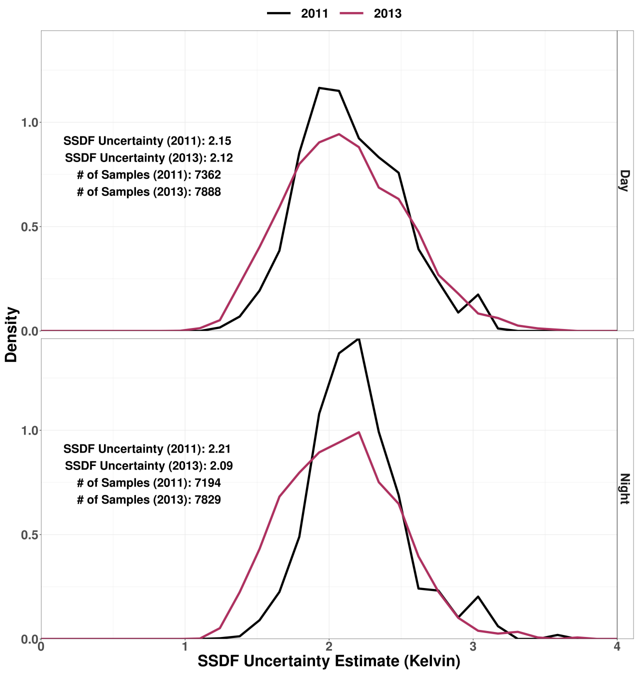

Figure 19: SSDF uncertainty histogram for 2011 (black) and 2013 (red) aggregated by day (top) and night(bottom). Summary statistics of mean SSDF uncertainty are provided as text on the upper left.

4 Discussion and conclusion

We have produced a new fused NSAT product over CONUS, from November 2012 through December 2020, using Spatial Statistical Data Fusion of AIRS and CrIMSS remotely sensed NSAT. We also provided detailed validation using withheld ISD data, and comparison to ECMWF ERA5-Land reanalysis. Remote sensing data provides information to span the spatial domain, in situ data provides the information to correct the remote sensing data, and SSDF provides the means to join them into something greater than the sum of their parts. The fused gridded product has no missing data (apart from one day and night without either AIRS or CrIMSS-CLIMCAPS input data); has improved accuracy and precision relative to the input satellite datasets, has comparable accuracy and precision relative to to ERA5-Land and indeed significantly lower nighttime bias than ERA5-Land; and includes estimates that are more consistent with the observed errors relative to in situ ISD observations. To summarize, our NSAT SSDF pilot product is comparable in precision and accuracy to the cutting-edge ERA5-Land reanalysis, but it is a direct observational product that does not involve physical modeling. Furthermore, unlike reanalysis it could support near-real-time product creation for operational applications.

SSDF is general and could be applied to any number of datasets estimating the same observable. It could be applied across a wide range of satellite observables, such as atmospheric composition, water vapor profiles, or vapor pressure deficit (the difference between the water vapour pressure and the saturation water vapour pressure), so long as uncertainty estimates of the input datasets can be obtained. We emphasize that the quality of the SSDF product depends on the quality of the bias and variance estimates of the input datasets.

Our plans for future work include improving the bias and variance estimation using simulation-based uncertainty quantification (Hobbs et al., 2017; Braverman et al., 2021). Simulation-based uncertainty quantification has the potential to further improve the overall quality of the SSDF product. It could also mitigate or eliminate the two issues our validation has uncovered: increased bias at a small number of data points at elevations in excess of 2500 m, and a 0.1 K shift in annual mean bias when transitioning from the AIRS-only version (2002-2012) to the two input (AIRS+CrIMSS) SSDF version (2012-2020).

We also plan to create an NSAT SSDF product over global land areas, expanding beyond CONUS, and apply the SSDF method to other hyperspectral surface products (e.g., vapor pressure deficit). Finally, we plan to develop SSDF products for satellite instruments that sample observables at different points in the diurnal cycle, to enable fusion of datasets from polar-orbiting and inclined platforms to make optimal use of all available remote sensing.

Open Research

The SSDF NSAT dataset described in this paper is available at `http://dx.doi.org/10.5067/CPXNAPA2WSQ8`.

Publicly available data were obtained from the NASA Atmospheric Infrared Sounder and the Suomi-NPP projects, the NOAA Integrated Surface Databse, and the European Centre for Medium-Range Weather Forecasts reanalysis.

Aqua AIRS V7 is available from the NASA GES DISC repository (AIRS Project, 2019). The retrieved surface air temperature (TSurfAir), the corresponding error estimate for TSurfAir (TSurfAirErr), and the corresponding quality flag (QC) (TSurfAir_QC) were obtained for the standard IR-only product.

CrIMSS-CLIMCAPS V2 is available from the NASA GES DISC repository (Barnet, 2019). Near surface temperature (surf_air_temp), the corresponding QC flag (surf_air_temp_qc), and the corresponding error estimate (surf_air_temp_err) were obtained from the NSR product.

NOAA ISD NSAT data is available using the rnoaa R package.

ECMWF ERA5-Land gridded hourly 2 m temperature means are available from the Copernicus Climate Change Service (C3S) Climate Data Store (Copernicus 2017).

Acknowledgments

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). We thank Evan Fishbein, Evan Manning, Erik Fetzer, and Bjorn Lambrigtsen for helpful discussions.

Appendix A Matchups and bias estimation

In this section, we will elaborate in detail our procedure for matching between ISD and the instruments' observations, and the consequent bias estimation process. For clarity, we establish the following notation. Let $\mathbf{s}$, $\mathbf{u}$, and $\mathbf{v}$ be latitude-longitude locations; e.g., $\mathbf{s} = (lat, lon)$. On a given day (or night) let $Z^k(\mathbf{u})$ be the value of the k-th instrument's near-surface temperature retrieval centered at $\mathbf{u}$. and focus on a single ISD station at location $\mathbf{s}$ during a single period. Let $t_1^I(\mathbf{s}), \ldots, t_M^I(\mathbf{s})$ be the times at which observations are acquired at this station during the period. These time points may be irregularly spaced, and M can change from station to station. The ISD measurements are $Z^I(\mathbf{s}, Z_m^I(\mathbf{s}))$, $m = 1, \ldots, M$.

Let $t^k(\mathbf{u})$ be the acquisition times associated with the k-th instrument's footprints centered at location $\mathbf{u}$. In principle, $\mathbf{u}$ ranges over all footprint locations for the appropriate instrument during the entire period, but in practice these locations are grouped by granules. We denote granule number during the current period by $g = 1, \ldots, 120$, and the set of footprints belonging to granule g by $\mathcal{G}_g^k$. The time associated with $\mathcal{G}_g^k$ is τ_g^k. To ease the computational burden, $\mathbf{u}$ ranges only over locations in the single granule with time that is closest to $t_m^I(\mathbf{s})$.

A matchup associates the location and time of an ISD value, $\left(\mathbf{s}, t^I(\mathbf{s})\right)$, with the location and time of the k-th instrument's footprint in the period: $\left(\mathbf{u}^*, t^k(\mathbf{u}^*)\right)$. The *matchup function* is,

$$
\begin{aligned}
&\mathbb{M}^k\left(\mathbf{s}, t_m^I(\mathbf{s})\right) = \left(\mathbf{u}^*, t^k(\mathbf{u}^*)\right), \\
&\mathbf{u}^* = \underset{\mathbf{u}}{\operatorname{argmin}} \left\{ ||\mathbf{u} - \mathbf{s}||, \, \mathbf{u} \in \left(\mathcal{G}_{g^*}^k \cap \mathcal{U}^{time} \cap \mathcal{U}^{space}\right) \right\}, \\
&g^* = \underset{g}{\operatorname{argmin}} \left\{ \left|\tau_g^k - t_m^I(\mathbf{s})\right| \right\}, \\
&\mathcal{U}^{time} = \left\{ \mathbf{u} : \left|t^k(\mathbf{u}) - t_m^I(\mathbf{s})\right| \leq 1 \text{ hour} \right\}, \; \mathcal{U}^{space} = \left\{ \mathbf{u} : ||\mathbf{u} - \mathbf{s}|| \leq 100 \text{ km} \right\}.
\end{aligned}
$$

Note that, for a given instrument and period, there will only be one granule that satisfies the criterion provided by g^*.

For a given ISD station (indexed by location $\mathbf{s}$) in the current period, p, we create the sets of matchup values for the k-th instrument as follows,

$$
\mathcal{A}^k(p, \mathbf{s}) = \left\{ Z^I\left(\mathbf{s}, t_m^I(\mathbf{s})\right), \, Z^k\left(\mathbb{M}^k\left(\mathbf{s}, t_m^I(\mathbf{s})\right)\right) \right\}_{m=1}^{M(p,\mathbf{s})}
$$

for all ISD time points at $\mathbf{s}$ indexed by $m = 1, \ldots, M(p, \mathbf{s})$. p is identified by a date and a mode (day/night) indicator, e.g., $p = (d, j) =$ (2013-01-01, day). $M(p, \mathbf{s})$ is the number of ISD station values in period p at location $\mathbf{s}$. There is at most one AIRS and one CrIMSS footprint associated with each station-time, but the same footprint can be associated with more than one station-time. Thus, $\mathcal{A}^k(p, \mathbf{s})$ may contain multiple elements if there is more than one ISD measurement during period p at location $\mathbf{s}$. They may also be empty if there are no matching AIRS or CrIMSS footprints.

After creating $\mathcal{A}^k(p, \mathbf{s})$ for all periods and ISD locations, we create supersets of matchup value pairs by combining across three-day moving windows, by mode:

$$
\mathcal{A}^{kj}(d, \mathbf{s}) = \mathcal{A}^k(d-1, j, \mathbf{s}) \cup \mathcal{A}^k(d, j, \mathbf{s}) \cup \mathcal{A}^k(d+1, j, \mathbf{s}), \;\; \mathcal{A}^{kj}(d) = \bigcup_{\mathbf{s}} \mathcal{A}^{kj}(d, \mathbf{s}).
$$

$j \in \{$day,night$\}$. We chose the three-day time window after experimenting with shorter and longer windows. Shorter windows did not provide adequate sample sizes while longer windows failed to capture weather-related changes. Ideally, window duration would be as short as possible since longer time windows result in larger variance estimates in the fused data, relative to withheld ISD data.

The final step before actually computing estimated bias and variance for each AIRS and CrIMSS footprint is to tessellate a 240 km (approximately two degrees), hexagonal spatial grid over CONUS. We do this by creating a discrete global grid using the DGGRID software package (Sahr et al., 2003; Sahr, 2019). One of the centers, for example, is at 87.72550324 W, 40.7908839 N, near Watseka, Illinois; this center uniquely determines the tessellated grid. All elements of $\mathcal{A}^{kj}(d)$ are sorted in to these grid cells based on the instrument's footprint locations. Formally, let $i \in 1, \ldots, L$ index grid cell centers, and let $1_i(\mathbf{u}) = 1$ if $\mathbf{u}$ lies inside cell i, and zero otherwise. For grid cell i, mode j, and date d, set

$$\mathcal{A}_i^{kj}(d) = \left\{ \left\{ Z^I\left(\mathbf{s}, t_m^I(\mathbf{s})\right),\, Z^k\left(\mathbf{u}_{m\mathbf{s}}^*, t^k(\mathbf{u}_{m\mathbf{s}}^*)\right) \,:\, 1_i(\mathbf{u}_{m\mathbf{s}}^*) = 1 \right\}_{m=1}^{M(d,j,\mathbf{s})} \right\}_{all\ \mathbf{s}},$$

where $M(d, j, \mathbf{s})$ is the number of time points acquired by the ISD station at $\mathbf{s}$ on day d in mode j, L is the total number of hexagonal grid cells, and we write $\mathbf{u}_{m\mathbf{s}}^*$ to emphasize its dependence on m and $\mathbf{s}$ via the matchup functions.

The bias assigned to all footprints from the k-th instrument observed on day d in mode j belonging to grid cell i is,

$$\mathrm{b}_{dji}^k = \frac{1}{|\mathcal{A}_i^{kj}(d)|} \sum_{all\ \mathbf{s}} \sum_{m=1}^{M(d,j,\mathbf{s})} \left[Z^k\left(\mathbf{u}_{m\mathbf{s}}^*, t^k(\mathbf{u}_{m\mathbf{s}}^*)\right) - Z^I\left(\mathbf{s}, t_m^I(\mathbf{s})\right) \right] 1_i\left(\mathbf{u}_{m\mathbf{s}}^*\right).$$

The corresponding variance assigned to all footprints observed on day d in mode j belonging to grid cell i is,

$$\mathrm{v}_{dji}^k = \frac{1}{|\mathcal{A}_i^{kj}(d)|} \sum_{all\ \mathbf{s}} \sum_{m=1}^{M(d,j,\mathbf{s})} \left[Z^k\left(\mathbf{u}_{m\mathbf{s}}^*, t^A(\mathbf{u}_{m\mathbf{s}}^*)\right) - Z^I\left(\mathbf{s}, t_m^I(\mathbf{s})\right) - \mathrm{b}_{dji\cdot}^k \right]^2 1_i\left(\mathbf{u}_{m\mathbf{s}}^*\right),$$

Subtracting the biases from the satellite footprints yields bias-corrected data. Denote an footprint acquired by the k-th instrument on day d in mode j, centered at location $\mathbf{u}$, by $Z_{dj}^A(\mathbf{u})$, where we suppress the argument $t^A(\mathbf{u})$ since, for a given date and mode, location and time are confounded. The bias-corrected value is denoted by $Z_{dj}^{k*}(\mathbf{u})$ as follow:

$$Z_{dj}^{k*}(\mathbf{u}) = Z_{dj}^A(\mathbf{u}) - \mathrm{b}_{dji^*}^A, \quad i^* = \underset{i}{\mathrm{argmax}}\, 1_i(\mathbf{u}),$$

with associated variance $\mathrm{v}_{dji^*}^k$.

References

AIRS Project. (2019). *Aqua/AIRS L2 Standard Physical Retrieval (AIRS-only) V7.0.* Goddard Earth Sciences Data and Information Services Center (GES DISC). Greenbelt, MD, USA. (Accessed: 2019-2021) doi: 10.5067/VP1M6OG1X7M1

AIRS Project. (2020). *About the data.* Retrieved 2021-12-26, from `https://airs.jpl.nasa.gov/data/about-the-data/granules/`

Barnet, C. (2019). *Sounder SIPS: Suomi NPP CrIMSS Level 2 CLIMCAPS normal spectral resolution: Cloud cleared radiances v2.* Goddard Earth Sciences Data and Information Services Center (GES DISC). Greenbelt, MD, USA. (Accessed: 2019-2021) doi: 10.5067/CNG0ST72533Z

Blackwell, W. (2005). A neural-network technique for the retrieval of atmospheric temperature and moisture profiles from high spectral resolution sounding data. *Geoscience and Remote Sensing, IEEE Transactions on*, *43*(11), 2535-2546. doi: 10.1109/TGRS.2005.855071

Braverman, A., Hobbs, J., Teixeira, J., & Gunson, M. (2021). Post hoc uncertainty quantification for remote sensing observing systems. *SIAM/ASA Journal on Uncertainty Quantification*, *9*(3), 1064–1093.

Chahine, M. T., Pagano, T. S., Aumann, H. H., Atlas, R., Barnet, C., Blaisdell, J., ... Zhou, L. (2006). AIRS: Improving weather forecasting and providing new data on greenhouse gases. *Bulletin of the American Meteorological Society*, *87*(7), 911–926.

Cressie, N. (2015). *Statistics for spatial data.* John Wiley & Sons.

Cressie, N., & Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 209–226.

Ferguson, C. R., & Wood, E. F. (2010). An evaluation of satellite remote sensing data products for land surface hydrology: Atmospheric infrared sounder. *Journal of Hydrometeorology*, *11*(6), 1234–1262.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., ... others (2019). Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, *7*(1), 6–39.

Gotway, C. A., & Young, L. J. (2002, June). Combining incompatible spatial data. *Journal of the American Statistical Association*, *97*, 632-648.

Hammerling, D. M., Michalak, A. M., & Kawa, S. R. (2012). Mapping of CO2 at high spatiotemporal resolution using satellite observations: Global distributions from OCO-2. *Journal of Geophysical Research: Atmospheres*, *117*(D6).

Harrison, G., & Burt, S. D. (2021). Quantifying uncertainties in climate data: measurement limitations of naturally ventilated thermometer screens. *Environmental Research Communications*.

Henderson, H., & Searle, S. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, *23*, 53-60.

Hennermann, K., & Berrisford, P. (2019). *ERA5 data documentation, ECMWF.*

Hobbs, J., Braverman, A., Cressie, N., Granat, R., & Gunson, M. (2017). Simulation-based uncertainty quantification for estimating atmospheric CO2 from satellite data. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), 956–985.

Mcnally, A. P., Watts, P. D., A. Smith, J., Engelen, R., Kelly, G. A., Thépaut, J. N., & Matricardi, M. (2006). The assimilation of airs radiance data at ecmwf. *Quarterly Journal of the Royal Meteorological Society*, *132*(616), 935–957. Retrieved from `http://dx.doi.org/10.1256/qj.04.171` doi: 10.1256/qj.04.171

Nguyen, H., Cressie, N., & Braverman, A. (2012). Spatial statistical data fusion for remote sensing applications. *Journal of the American Statistical Association*, *107*(499), 1004-1018. Retrieved from `https://doi.org/10.1080/01621459.2012.694717` doi: 10.1080/01621459.2012.694717

Nguyen, H., Katzfuss, M., Cressie, N., & Braverman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, *56*(2), 174–185.

Olsen, E. T., Fishbein, E., Manning, E., & Maddy, E. (2017). AIRS/AMSU/HSB Version 6 L2 product levels, layers and trapezoids. *Jet Propulsion Laboratory, Pasadena, CA, USA.*

Sahr, K. (2019). *Dggrid version 7.0: User documentation for discrete global grid software.*

Sahr, K., White, D., & Kimerling, A. J. (2003). Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, *30*(2), 121–134.

Smith, A., Lott, N., & Vose, R. (2011). The integrated surface database: Recent developments and partnerships. *Bulletin of the American Meteorological Society*, *92*(6), 704–708.

Smith, N., & Barnet, C. D. (2019). Uncertainty characterization and propagation in the Community Long-Term Infrared Microwave Combined Atmospheric Product System (CLIMCAPS). *Remote Sensing*, *11*(10), 1227.

Smith, N., & Barnet, C. D. (2020). CLIMCAPS observing capability for temperature, moisture, and trace gases from AIRS/AMSU and CrIS/ATMS. *Atmospheric Measurement Techniques*, *13*(8), 4437–4459.

Sun, J., McColl, K. A., Wang, Y., Rigden, A. J., Lu, H., Yang, K., ... Santanello Jr, J. A. (2021). Global evaluation of terrestrial near-surface air temperature and specific humidity retrievals from the atmospheric infrared sounder (airs). *Remote Sensing of Environment*, *252*, 112146.

Susskind, J., Blaisdell, J. M., & Iredell, L. (2014). Improved methodology for surface and atmospheric soundings, error estimates, and quality control procedures: the atmospheric infrared sounder science team version-6 retrieval algorithm. *Journal of Applied Remote Sensing*, *8*(1), 084994. Retrieved from http://dx.doi.org/10.1117/1.JRS.8.084994 doi: 10.1117/1.JRS.8.084994

Thrastarson, H. T., Manning, E., Kahn, B., Fetzer, E., Yue, Q., Wong, S., ... others (2020). AIRS/AMSU/HSB Version 7 Level 2 product user guide. *Jet Propulsion Laboratory, California Institute of Technology: Pasadena, CA, USA*, 83–92.

Wilks, D. S. (2006). *Statistical methods in the atmospheric sciences* (Second ed.). Academic Press.

Yue, Q., Lambrigtsen, B., et al. (2017). AIRS V6 test report supplement: Performance of AIRS+AMSU vs. AIRS-only retrievals. *Jet Propulsion Laboratory, California Institute of Technology: Pasadena, CA, USA.* Retrieved from https://docserver.gesdisc.eosdis.nasa.gov/repository/Mission/AIRS/3.3_ScienceDataProductDocumentation/3.3.5_ProductQuality/V6_Test_Report_Supplement_Performance_of_AIRS+AMSU_vs_AIRS-Only_Retrievals.pdf

Yue, Q., Lambrigtsen, B., et al. (2020). AIRS V7 L2 performance test and validation report. *Jet Propulsion Laboratory, California Institute of Technology: Pasadena, CA, USA.* Retrieved from https://docserver.gesdisc.eosdis.nasa.gov/public/project/AIRS/V7_L2_Performance_Test_and_Validation_report.pdf

Yue, Q., Lambrigtsen, B., et al. (2021). Version 2 CLIMCAPS-Aqua retrieval product performance test report. *Jet Propulsion Laboratory, California Institute of Technology: Pasadena, CA, USA.* Retrieved from https://docserver.gesdisc.eosdis.nasa.gov/public/project/Sounder/CLIMCAPS.V2.Test.Report.Aqua.pdf