Assessing gap-filled Landsat land surface temperature time series data using different observational datasets

Hua Shia\* and George Xianb

aASRC Federal Data Solutions (AFDS), Contractor to the USGS Earth Resources Observation and Science Centre (EROS), Sioux Falls, SD 57198, U.S.A.; bUSGS Earth Resources Observation and Science Centre (EROS), Sioux Falls, SD 57198, U.S.A

\*corresponding author: hshi@contractor.usgs.gov; Tel.: +01-605-594-6050

Word count: 6338

Assessing gap-filled Landsat land surface temperature time series data using different observational datasets

Landsat Analysis Ready Data (ARD)-based time series present challenges in monitoring surface urban heat islands (SUHI) due to rapid changes in land surface temperature (LST) compared to cloud-free satellite observations. This research is to investigate the use of a spatiotemporal gap-filling model as a feasible and cost-effective solution to produce Landsat time-series LST products with both high spatial resolution and temporal frequency. The study identified and filled Landsat ARD thermal times-series data gaps due to missing data, cloud and shadow effects, and data quality. The accuracy of Landsat gap-filled products was assessed using randomly selected clear observations of Landsat and uncertainty products from the gap-filling model, and was evaluated using various existing temperature datasets, including climate data from NOAA Global Historical Climate Network (GHCN) station observations, Daily Surface Weather and Climatological Summaries (DAYMET), and land surface temperature including MODIS, VIIRS, and ECOSTRESS. The result suggests that the gap-filled Landsat LST has significant correlations with existing datasets including field observation and remote sensing data derived from other sensors that have similar monthly and seasonal variation patterns. The uncertainty maps show spatial distributions of uncertainty for gap-filled pixels that have high or low uncertainties. The Landsat gap-filled time-series datasets can be used to measure annual, seasonal, or even monthly landscape thermal conditions, which are useful for SUHI and relevant research, and to perform multi-decade time series LST change analysis under climate change conditions.

Keywords: temporal frequency; gap-filling; Landsat surface temperature; uncertainty analysis, accuracy assessment

# 1. Introduction

Satellite images provide valuable geospatial data for characterizing thermal conditions and support ecosystem and environmental change monitoring, but standard satellite missions always have to live with the trade-offs between spatial resolution and temporal frequency (Luo, Guan, and Peng 2018; Roy and Yan 2020). Remote sensing techniques are increasingly used to combine data from multiple sensors and platforms to create high-quality data products (Bauer 2020). However, these data products may contain gaps due to missing or incomplete data (cloud cover or sensor malfunctions) that can reduce their accuracy and usefulness for scientific research and applications (Wulder et al. 2011). To overcome these limitations, gap-filling (or data fusion) models are widely used in environmental change monitoring, where continuous and accurate observations of environmental variables are crucial (Gao et al. 2006; Roy et al. 2008; Roy and Yan 2020; Zhou, Xian, and Shi 2020).

The growth of multi-sensor integrated datasets provides the opportunity to investigate land surface temperature (LST) dynamics and environmental changes at both high spatial resolution and temporal frequency, but also urges approaches to reduce the inconsistency of data availability. Many gap-filling approaches were developed for predicting missing values related to cloud contamination and Landsat 7 Scan Line Corrector (SLC)-off data (Chen et al. 2011; Yan and Roy 2018; Zhu et al. 2022).Yan and Roy (2018) The main purpose of these models is to fill the gaps in time-series data caused by missing values, sensor failure, or other factors including clouds and cloud shadows, and to estimate the missing values with reasonable accuracy. The accuracy of gap-filling models is essential for environmental change monitoring, as inaccurate or biased estimates (Roy and Yan 2020) may lead to incorrect conclusions and decisions (Stehman et al. 2018; Stehman and Foody 2019; Wulder et al. 2022). To achieve accurate gap-filling, several methods have been proposed, including interpolation, regression, and machine learning algorithms (Zhu et al. 2010; Zhu et al. 2016; Zhou et al. 2022). Despite the progress made in this field, gap-filling models still face challenges, such as uncertainty analysis and accuracy assessment, which need to be addressed to improve their performance for gap-filled products (Wang et al. 2022; Zhu et al. 2022).

The accuracy of these gap-filling algorithms is crucial for ensuring the reliability and usefulness of the resulting data products (Niclòs et al. 2011; Foody 2020). Accuracy assessment is the process of evaluating the agreement between the estimated values from gap-filling techniques and the true values. Uncertainty analysis is the process of quantifying the variability in the estimated values due to uncertainties in the gap-filling process (Rocchini et al. 2013). The accuracy assessment and uncertainty analysis of gap-filled products are necessary to understand the limitations of the data and to ensure that the data is suitable for use in environmental research and management (Rounsevell et al. 2021; Zhu et al. 2022).

Uncertainty analysis of gap-filled LST products can also provide insights into the reliability of observed data and model outputs in urban environmental change monitoring, which is crucial for decision making and policy development (Rocchini et al. 2013). Accuracy assessment ensures that environmental change monitoring data is reliable and can be used for decision making purposes (Leyk et al. 2018).

One of important environmental change monitoring is surface urban heat island (SUHI) research. In urban heat island research, gap-filling models can be used to complete missing data points in temperature data, which is important for understanding the effects of urbanization on temperature patterns in temporally (Zhou, Xian, and Shi 2020). Uncertainty analysis can also help to quantify the uncertainty associated with temperature data and model outputs, which is important for understanding the reliability of temperature data and for making informed decisions about the impacts of urban heat islands on human health and the environment (Rocchini et al. 2013). Accuracy assessment of gap-filled data is useful for SUHI intensity and heat wave related analysis so that reliable SUHI information can be used to develop effective public health strategies to mitigate the impacts of climate change on human health. The main research problem associated with these models is the input data and models' uncertainty, which can significantly affect the accuracy of gap-filled data.

In this study, we answered questions including what are the current methods used to fill gaps in surface temperature data, what is the uncertainty of input data and models, what is the accuracy of gap-filled products, and what are the limitations of our gap-filling models and how can they be improved. We hypothesize that 1) the use of gap-filling algorithms is a more effective method for filling gaps in Landsat thermal conditions than traditional interpolation methods; 2) the uncertainty in surface temperature data can be quantified using statistical models by incorporating prior knowledge; and 3) the use of gap-filling models and uncertainty analysis can improve the accuracy of urban heat island studies models inputs by reducing the bias and variance in surface temperature datasets.

With the hypotheses, our objectives were 1) to identify and analyse uncertainty of input datasets and gap-filling models for surface temperature time series data; 2) to investigate the impact of gap-filling on the surface temperature time series data; 3) to compare the performance of the gap-filled surface temperature time series data with field observation and other existing LST data for the validation of the accuracy and reliability of the gap-filling technique(s); 4) to provide recommendations for the selection and application of gap-filling techniques for improved urban heat island study.

# 2. Materials and methods

We carried out this research through several steps. We first collected reference datasets from various existing sources with multiple spatial resolutions and temporal frequencies (Section 2.3). For each date of the time series within selected years, reference temperature for each date was taken the same date (or a close date if data were missing) as the gap-filled Landsat LST date (Section 2.4). These reference datasets provided the basis for the accuracy estimates. Then, we conducted the accuracy assessment (Section 2.5) following protocols of consistent estimation required for a statistically rigorous analysis. The statistical parameters of R-Square (R2) and Root Mean Square Error (RMSE) were used. Finally, we analysed the uncertainty from gap-filling models with input Landsat data and the uncertainty from comparison datasets by estimating standard errors using gap-filling models (Section 2.6) and reporting the uncertainty of the users, reference data, and overall accuracies.

## 2.1. Study area

We selected three study areas in the conterminous United States (CONUS) (Section 2.1). These areas are Atlanta, GA; Phoenix, AZ; and Sioux Falls, SD during selected years (1991, 2000, 2016, and 2020). The Atlanta area, one of the largest and most populated urban centres in the U.S. covers four ARD tiles with a total area of 90,000 km2 (Figure 1). It is located in northern Georgia, near the Blue Ridge Mountains. The area has a population of 6,220,106 in 2022 according to the U.S. Census Bureau (<https://www.census.gov/>). The rural landscapes surrounding the city comprise forest, croplands, pastures, hayfields, and water bodies. The area has a humid subtropical climate and monthly mean air temperatures of 6.1 ºC in January and 26.8 ºC in July (NWS 2021). The first six months of 2022 were Atlanta's fourth hottest on record, according to the National Oceanic and Atmospheric Administration (NOAA). The area receives abundant rainfall with an annual average of 1260 mm.

The second study area is the Sioux Falls metropolitan area and surrounding rural areas in South Dakota, United States. The area is within one ARD tile and has a spatial extent of 22,500 km2 (Figure 1). The city of Sioux Falls has grown at a rapid pace since the late 1970s, with the city’s population increasing from 81,000 in 1980 to 208,884 in 2022 (<https://www.census.gov/>). It is the 130th largest city in the US but the largest city in South Dakota. The rural landscapes surrounding the city comprise croplands, pastures, and hayfields, with patches of forests concentrated in parks, bottomlands, shelterbelts, and farmsteads. Within the sub-humid continental temperate climate zone, Sioux Falls has warm, humid summers and cold winters with most precipitation occurring between April and September (yearly average about 840 mm). The monthly mean air temperatures vary from −16.7 to −4.4 °C in winter (December– February) and from 16.1 to 30.0 °C in summer (June– August). The area is also known for its strong winds which can reach up to 56 km per hour.

*Figure 1 near here*

The Phoenix area is within two Landsat ARD tiles with a spatial extent of 45,000 km2 (Figure 1). Phoenix is the most populous city of Arizona, with 1,644,409 residents as of 2022 (<https://www.census.gov/>). It is the fifth most populous city and the most populated state capital in the country and the only U.S. state capital with a population of more than one million residents. Phoenix lies near the confluence of the Gila and Salt rivers and is situated at the northern edge of the Sonoran Desert, an arid ecological zone whose characteristic plant is the nationally protected saguaro cactus. The area has a typical arid subtropical climate. The metropolitan area is known as the "Valley of the Sun" due to its location in the Salt River Valley. The area has monthly mean air temperatures of 35 ºC in July and 12 ºC in January, has a large temperature difference between day and night, and receives only 185 mm annual average rainfall (NWS 2021).

## 2.2. Reference data and extracting strategy

We selected four existing reference datasets for the study (Table 1). The first one is NOAA Global Historical Climatology Network daily (GHCN) (Figure 1 and Table 2), an integrated database of daily climate summaries from land surface stations across the globe (Menne et al. 2017). GHCN is made up of daily climate records from numerous sources that have been integrated and subjected to a common suite of quality assurance reviews. NOAA National Centres for Environmental Information (NCEI) provides numerous daily variables, including maximum and minimum temperature, total daily precipitation, snowfall, and snow depth. About half the stations only report precipitation. Both record length and period of record vary by station and cover intervals ranging from less than a year to more than 175 years. The second one is Moderate Resolution Imaging Spectroradiometer (MODIS), onboard the NASA Terra and Aqua Earth Observing System satellites, which provides daily multiple LST products. The most recently Collection 6 (C6) MODIS LST includes three refinements over bare soil surfaces compared to the Collection 5 (C5) MODIS LST product (Duan et al. 2017; Hulley and Hook 2021). The third one is VIIRS-derived data products that are used to measure cloud and aerosol properties, ocean colour, ocean and LST, ice movement and temperature, fires, and Earth's albedo. Climatologists use VIIRS data to improve our understanding of global climate change archived and distributed through the Oak Ridge National Laboratory (ORNL) (Hulley and Hook. 2018). The fourth reference dataset is the Daily Surface Weather and Climatological Summaries (DAYMET) dataset (Thornton et al. 2021), which provides gridded estimates of daily weather parameters for North America, including daily continuous surfaces of minimum and maximum temperature, precipitation occurrence and amount, humidity, shortwave radiation, snow water equivalent, and day length.

*Table 1 and Table 2 near here.*

We selected climate data from all available GHCN stations (Table 2), which have full temperature records for 1991, 2000, 2016, and 2020, as extracting points and converted the points to Landsat resolution (30 m × 30 m) from all pixels in the CONUS ARD grid system as a mask to spatially match the other existing LST datasets. The total sample consisted of 180 points within seven CONUS ARD tiles in three selected study areas. We chose this extracting strategy to prioritize four desirable strategy criteria: (1) probability based; (2) simple to implement; (3) easy to compare for multiple spatial resolution LST datasets; (4) extracting the same locations to GHCN stations. Another motivation for implementing the extracting strategy is that GHCN is field observation and air temperature but more accurate than other existing remote sensing derived LST. Also, we employed a weekly composite of MODIS and VIIRS data, which offers improved quality due to the composite algorithm mitigating cloud impact. For example, the daily MODIS and VIIRS data often contain large portion of missing values, potentially leading to misleading comparison results. Additionally, we used GHCN station locations to create 3\*3 Landsat pixel (30 m) masks to get average values from gap-filled Landsat LST and extract DAYMET (1000 m), MODIS and VIIRS (1000 m), and ECOSTRESS (70 m) to be comparable within the same land cover class. The gap-filled uncertainty layer is used for analysis.

## 2.3. Summary of gap-filling method

As summarized in the introduction, we have developed a new method of time series gap-filling that is designed for multi-sensor and multi-time data harmonization (Figure 2). This method uses pixels from the orbit overlap region to fill data gaps based on time series similarity, which retains the observation variation. Model assembling procedures were used to estimate stable predictions that are not only robust to occasionally cloud-contaminated training data but also allowed us to estimate the uncertainty of the predictions (Zhou, Xian, and Shi 2020). In this study all overlap regions in the study areas have more than 21 clear observations, while non-overlap regions often have insufficient data. Thus, the 7-parameter linear harmonic model is used to replace those cloud-contaminated data, which maintains the details of seasonality in the training data (Equation (1)). The procedure for each target pixel that includes clustering the training data, stratifying random selection for the target pixel, and predict the full time series LST via linear regression (equation (2)). For detailed information see (Zhou, Xian, and Shi 2020).

(1)

where is the modeled time series for a single pixel location in the overlap region; describes the mean of over the time series; and are coefficients for harmonic component *m*; *t* is day of year; and *L* is the length of the time period (*L* = 365.25). Parameter *M* (*m* = *3*) determines the highest frequency used for modelling.

(2)

where and represent the linear parameter to be estimated and represents the error terms. and are sampled training data, and the target pixel time series at y are clear observation dates.

*Insert figure 2 here.*

## 2.4. Accuracy assessment

Accuracy assessment is the procedure used to quantify product quality. Attempts have been made to quantify limiting factors resulting from the Landsat low temporal availability of data used for generating high frequency LST information at regional level. Sub-pixel fractional error matrices are introduced as a more appropriate way for assessing the accuracy of mixed pixels. For classification with coarse spatial resolution data, limitations of the classification method produce a maximum achievable accuracy defined as the average percent fraction of dominant land cover of all pixels in the mapped area. We used a combination of station data, climate data (DAYMET), MODIS and VIIRS LST data, and Landsat data to validate the accuracy of the gap-filled products (Figure 2). Specifically, we used the following methods. (1) Station validation: We selected NOAA GHCN station observation data on temperature and intensity at 20 randomly selected sites in the study areas. We compared these data to the corresponding data in the gap-filled product to assess the accuracy of the gap-filling techniques. (2) Landsat validation: We chose raw Landsat data to validate the accuracy of the gap-filled products over time. We compared the Landsat data from multiple time periods to the gap-filled product to Landsat data. We estimated temperature and UHI intensity using image analysis software and compared the results to the gap-filled product. (3) DAYMET LST, MODIS and VIIRS LST data: We also used these data to evaluate the gap-filled products. The GHCN station data is a point dataset, and DAYMET data is rasterized based on GHCN with 1 km resolution. The MODIS and VIIRS dataset have a spatial resolution of 1 km with 8-day composites. ECOSTRESS is a new dataset with limited products.

We compared the gap-filled data with the original data to assess the accuracy of the gap-filled products using two metrics to evaluate the accuracy: root mean square error (RMSE) and the coefficient of determination (R2). The uncertainty analysis was conducted to quantify the variability in the gap-filled data. We used the bootstrap method to simulate the variability in the gap-filled data and calculated the 95% confidence intervals for the gap-filled data. The following equations were used to calculate statistical parameters of RMSE and R2:

(3)

where: differences, squared, and N is sample size.

(4)

where: N is samples, X is the predictor variable, and Y is the response variable in this regression model.

## 2.5. Uncertainty analysis

There are many sources of uncertainty in Landsat gap-filling processing, such as clear observation training collection, QA band issues, and the modelling approach issues. The gap-filled Landsat LST is generally based on clustering of spatial entities within a spectral space. One major concern is the use of seasonal models to predict the variability of LST into several discrete dates within seasons. This type of approach is often inappropriate given the continuous values by regression model, which usually provides overestimated prediction in the high end and underestimated prediction in the low end. This leads to uncertainty in the products resulting from the use of remote sensing data. Based on the assumption that any gap-filled Landsat LST has an associated error and/or uncertainty of unknown magnitude, the statistical quantification of uncertainty analysis should be a core part of scientific research. In this study we analysed uncertainty layers from Landsat gap-filling models and reviewed recent attempts to take explicitly into uncertainty when mapping LST. We used the Landsat gap-filling uncurtaining layers that calculated the standard deviation (SD) of iterations for each prediction as an indicator of uncertainty (Figure 2). Standard deviation is often used to quantify the level of uncertainty in a set of measurements. In the context of gap-filling, it is used to indicate the degree of uncertainty of gap-filled LST for the pixel. The magnitude of the standard deviation is directly proportional to the level of uncertainty of the pixel. In other words, the larger the standard deviation is, the more uncertainties of the pixel exist. The results have the same structure as ARD tiles with all missing data filled, using the following formula:

(5)

where: is SD, n is the number of filled LST, x is filled LST, and is mean of filled LST.

## 2.6. Comparison analysis

The synthesis and comparison of spatial data collected or derived at different spatial resolutions with various uncertainty factors is challenging work in geospatial statistics (Gotway and Young 2002). The geostatistical concept of support belongs to one of spatial scales, in which it is a property of a variable used in comparison analysis. The support can be as small as a point or as large as the full extent of the study area. In this study, the spatial unit, which is based on GHCN station with Landsat 9 pixels mask, used to sample the multiple LST and climate datasets have also been considered in the support of the data (Figure 2). Comparison results (RSME, standard deviation, spatial autocorrelation, and others) can sometimes be estimated using geostatistical support-effect models. We present the case studies of LST illustrating some of these differences and how conclusions about zonal mean of temperature may be affected. We identify the influence of GHCN observational air temperature on statistical results as a local zonal mean temperature. The way to avoid the station air temperature is by careful construction of sampling design and analysis.

# 3. Results

## 3.1. Landsat ARD gap-filled LST

The study conducted in the three selected cities in 1991, 2000, 2016, and 2020 demonstrated that the Landsat ARD gap-filled products can better differentiate the performances of the spatiotemporal gap-filling model with improved training data strategy. The top row of Figures 3 (Atlanta), 4 (Sioux Falls), and 5 (Phoenix) shows the examples of the gap-filled Landsat LST from different dates. The original Landsat LST only covered part of the tiles (bottom row) and was contaminated by scattered clouds. By closely comparing these pair maps, the gap-filled Landsat LST showed a similar value range to the Landsat LST observations in clear areas. Both observations and gap-filled showed high LST in urban (or bare) areas and low LST along forest or water areas. The spatial distribution patten of gap-filled Landsat LST is associated well with urban land and most of high values are located within the city limits in Atlanta and Sioux Falls.

The gap-filling LST results are promising when compared to the spatial patterns of the GHCN station dataset. The gap-filled LST showed more spatial detail and variation than other LST data because these products, including MODIS/VIIRS and ECOSTRESS are at 500 m (daily) and 1000 m (8-day composite) while the gap-filling procedure is at 30-m scale. Temporally, the gap-filled products revealed more detailed seasonal and monthly variations, which can potentially be used to study seasonal SUHI change.

*Figures 2, 3, and 4 near here.*

## 3.2. Accuracy assessment

We used randomly selected GHCN observation data as the validation datasets, in which over 350 days per year of air temperature records are available. By comparing the 3x3 pixel average LST with weather station records, we found that the gap-filled Landsat LST data had good agreement with the data from GHCN in all three study areas. Figures 5, 6, and 7 represent the time series of observed and gap-filled Landsat LST in 36 GHCN field stations from 1991, 2000, 2016, and 2020 with different Landsat sensors within three study areas. The GHCN air temperature is also displayed as a reference of seasonal temperature pattern in the three selected study areas. The gap-filled values followed well with observations except for several scattered outliers in specific dates and stations, while both observed and gap-filled summer LSTs are higher than the station air temperatures. The outliers in Figure 6 with very low LST came from Landsat 8 images that were mostly covered by clouds, but pixel QA only flagged part of the images as cloud or cloud shadow. Similarly, the low LST at the peak of the summer in Figure 6 also came from cloud-contaminated observations (Landsat 5 and 7 image). Figure 7 shows that the LST gap-filled model didn’t perform well in winter months because of snow cover and limited clear observation for training. To assess the accuracy quantitatively, the RMSE and R2 of observed LST with GHCN and gap-filled Landsat LST with GHCN were calculated (Tables 3, 4, and 5). Generally, the gap-filled Landsat LST has a better range of RMSE than the Landsat LST product accuracy, indicating the algorithms have successfully filled the LST. The RMSE of the gap-filled Landsat LST data ranged from 1.5 to 4.2°C (see Table 3) in Atlanta, 1 to 5 °C in Sioux Falls (Table 4), and 0.5 to 3°C in Phoenix (Table 5). The gap-filled data has a significant relationship with GHCN validation data. We compared both original clear Landsat LST (in brackets is the number of clear observations in Tables 3, 4, and 5 and gap-filled Landsat LST with GHCN air temperature. The correlation between GHCN and gap-filled LST is similar with (or lower than) GHCN and the original Landsat clear observation if the number of gap-filled LST close or equal the number of original clear observation, the average R2 values of all stations is about 0.81 and 0.68 (p<0.05) in Atlanta (Table 3) , 0.74 and 0.81 (p<0.05) in Sioux Falls (Table 4), 0.82 and 0.74 (p<0.05) in Phoenix (Table 5). Some of the gap-filled Landsat LST show large variations, while observations are consistent at low or high values. As the GHCN station data were randomly selected and expanded to 9 Landsat pixels, we went through most gap-filled images and found that some cloud or cloud shadow pixels were labelled as clear observations by the Landsat pixel QA band, which were not filled and could lead to the consistently high or low values of gap-filled LST.

*Tables 3, 4, and 5 near here*

*Figures 5, 6, and 7 near here*

The gap-filled Landsat LST products, on average, slightly overestimate LSTs in urban and bare areas and underestimate LSTs in other land cover in all three study areas. Atlanta (Figure 6), located in a subtropical humid climate condition, has relatively small seasonal or annual temperature variations compared with other study areas. Phoenix (Figure 8), in a subtropical arid land climate condition, has more clear observations for training, resulting in better performance of the gap-filling model than in Sioux Falls (Figure 7), which is located in a moderate temperate continental climate condition with snow cover in winter and has less training data available.

## 3.3. Uncertainty analysis

We also analysed the uncertainty of the gap-filled Landsat surface temperature data using a Monte Carlo simulation. The simulation randomly sampled the input data from their respective probability distributions and calculated the resulting distribution of the gap-filled data. The simulation showed that the uncertainty of the gap-filled Landsat LST data was primarily determined by the uncertainty in the input Landsat data, followed by the uncertainty in the auxiliary data and model parameters. Figure 9 shows the spatial distribution of the yearly number of clear observations of Landsat (top) and annual mean of gap-filling models uncertainty map (bottom) in 2016 and 2020 in the three study areas. Table 6 gives the maximum, mean, and minimum number of clear observations and uncertainty values for the same times. Table 6 also shows that 1) in Atlanta, 55% of pixels have prediction uncertainties of ±0.65 °C in 2016 and ±0.97 °C in 2020 with annual mean LST of 19.2 °C in 2016 and 18.4 °C in 2020; 2) in Sioux Falls, over 50% of prediction uncertainties are around ±0.94°C in 2016 and ±2.08°C in 2020 with annual mean LST of 13.9 °C in 2016 and 13.6 °C in 2020; 3) in Phoenix, about 60% of prediction uncertainties are about ±0.76°C in 2016 and ±0.68°C in 2020 with annual mean LST of 29.2 °C in 2016 and 29.9 °C in 2020. The gap-filled products from the gap-filling model have better results in Atlanta and Phoenix, except for individual outliers. The uncertainty is high in Sioux Falls because it has fewer clear observation for training data due to snow/ice cover in wintertime. The uncertainty is low in Phoenix because it has more observations for training data than the other two areas.

*Table 6 near here,*

We used 9-pixel masks to calculate the mean of prediction uncertainty by using GHCN data (Figure 9).

*Figures 8 and 9 near here.*

The factors affecting gap-filled Landsat LST product accuracy could come from several sources including violations of the fundamental assumptions underlying the approach, Landsat swath overlap areas, seasonal linear regression algorithm, errors in the collection training dataset by using the QA band, and Landsat thermal data quantity and quality (Figure 10). For example, the gap-filled products have higher value of accuracy assessment in Phoenix due to high availability of clear observations for training data (Figure 10C). The gap-filled products have lower value of accuracy assessment in Sioux Falls due to snow/ice cover even it is a clear observation, because of the lack of clear observation, the gap-filling modelling approach assumes that the spectral behaviour of LST through time can only be represented with seasonal harmonic models that are used for training all dates during periods of time with limited training data. Therefore, the model output is expected to be better in winter when training data is extracted from warm months (dates) and the regression model that has a high uncertainty in snow cover dates. Figures 9B show high uncertainty values in the early dates of the year (mostly in January and February in Sioux Falls). We also checked the original Landsat images for these pixels having high uncertainty value by dates and location in Figures 8 and 9. Most of these pixels have cloud cover or no clear observation available. All these factors, such as snow cover and limited training data with seasonal model, depend on the regional geography and vary across different climate zones. For example, the quantity of available cloud- and snow-free Landsat data varies considerably in these three study areas and is shown by the presence of swath overlap zones of Landsat acquisition footprints within ARD tiles. How the algorithm responds to these variations in input data can impact the accuracy of the gap-filled products. For the gap-filling model, the selection of training data and accuracy of training data labels may both exert considerable influence on accuracy. Decisions made with respect to the filtering and balance of the pool of seasonal training data points may have different effects on accuracy of LST. Finally, using overlapped and clear observations from adjacent ARD tiles in training models improved the models and removed a source of error that arose from the gap-filling model design.

## 3.4. Comparison analysis

We compared the gap-filled Landsat LST with three other large-scale time series LST products including MODIS, VIIRS, ECOSTRESS, and rasterized DAYMET climate data. We also compared these LST datasets with GHCN station observations. We created 3x3 Landsat pixels based on GHCN points as a centre pixel to match coarse LST datasets. The comparison shows that the gap-filled Landsat LST matches the GHCN air temperature better than the others, except for DAYMET data. We did compare DAYMET with GHCN, and they are well matched with R2 = 0.99 in most cases. We utilized MODIS and VIIRS 8-day composites due to their superior data quality compared to daily data, primarily because of the impact of cloud cover. Specifically, the daily MODIS and VIIRS data often contain numerous missing values, even on clear-sky days. This makes it challenging to compare them with gap-filled Land Surface Temperature (LST) data at a 3x3 30 m pixel size based on GHCN locations. By filtering out cloud pixels from the daily ECOSTRESS data using pixel QA information, we found that there isn’t enough usable data available. Additionally, ECOSTRESS data is only available for the year 2020. Table 7 shows R2 and RMSE from the selected GHCN station locations. In general, gap-filled LST attains the accuracy that is broadly consistent with accuracy of these other existing LST datasets. For example, in the daily data comparison, gap-filled has low or similar RMSE to VIIRS in 2020 and ECOSTRESS has the highest RMSE. One reason is the gap-filled and ECOSTRESS LSTs have daily data, but VIIRS and MODIS LSTs are 8-day composite data that have better results with cloud and bad pixels fixed. For monthly and seasonal, all datasets have similar patterns with GHCN except ECOSTRESS.

*Table 7. near here*

# 4. Discussion

This study presented a comprehensive accuracy assessment, uncertainty analysis, and multi-dataset comparison for evaluating the gap-filled Landsat LST data. The multi-disciplinary history provided conveys the complexity of the issues encountered in vali-dating and comparing multiple source spatial data and the widespread interest in solutions that have been developing over several decades. This is now an active area of re-motely sensed data fusion and gap-filling research, and much novel research work has recently been developed (Roy and Yan 2020; Zhu et al. 2022).

Our results show that the product using gap-filling techniques has high accuracy for estimating annual, seasonal, and even monthly thermal condition for UHI and trends analysis. It also shows that the gap-filling algorithm is effective in filling missing values in remote sensing data. However, the accuracy of the gap-filled data varied depending on the test site with type of climate zones. The application of models was more accurate in arid areas (Arizona) than in humid regions (Georgia). However, the model outcomes in South Dakota are similar with these in other two locations in summer but are not good in winter months due to snow/ice cover. The RMSE values were lower in the urban site than in the other land covers, while there was no significant difference of RMSE in different geographical regions. The correlation coefficient was higher in the arid region than in the humid region. The uncertainty analysis showed that the variability in the gap-filled data was higher in the semi-arid region of Arizona than in the humid region of Georgia. The uncertainty affecting gap-filled Landsat LST product accuracy could arise from several sources including modelling approach, seasonal linear regression algorithm, errors in the collection training dataset by using the QA band, and Landsat thermal data quantity and quality. The comparison analysis reveals that gap-filled Landsat LSTs are more accurate in monthly and seasonal estimates.

Several limitations are found to the current study. First, the validation data sources used in our study have their own uncertainties, and these uncertainties may propagate into our assessment of the accuracy and uncertainty of the gap-filled Landsat LST. Additionally, NOAA GHCN records air temperatures that are different from gap-filled and other LST datasets. Second, we only evaluated our own gap-filling techniques. Third, we only used Landsat data and our results may not be applicable to other remote sensing datasets. Fourth, our study only evaluated the accuracy and uncertainty of the gap-filled products at the 9 pixel (30x30 m) and two Landsat ARD tile levels (5000x5000 m). Future studies may need to investigate the accuracy and uncertainty at the continental or global scale.

Gap-filling can introduce uncertainty into the final products because of combining factors including input data and models selection (Friedl et al. 1995; Murphy et al. 2004; Zhou, Xian, and Shi 2020; Rounsevell et al. 2021). Depending on the method used, gap-filling can be a complex and computationally intensive process. Gap-filling requires access to multiple dates of data, and the availability of these data may be limited in some regions or for certain time periods because of missing data, cloud and shadow, and snow and ice cover (Gao et al. 2006; Zhu et al. 2022). The accuracy of gap-filling can be affected by spatial and temporal variability in the data. For example, filling in missing data in a forested area may be more difficult than in an area with lower temperature. Different methods may be more or less suitable for different types of data or for different applications (Rocchini et al. 2013; Leyk et al. 2018; Stehman and Foody 2019). Gap-filling relies on the quality of the surrounding data to estimate missing values (Zhang et al. 2020; Zhou, Xian, and Shi 2020). If the surrounding data is poor quality or affected by noise, the accuracy of the gap-filled data may be compromised. Validating the accuracy and reliability of the gap-filled data can be challenging, as there may not be ground-based measurements or other independent sources of data available for comparison except air temperature from weather station observations. Also, as it is difficult to determine the true value of missing data, it can make it difficult to evaluate the accuracy of the gap-filling method.

Future work may need to address some of the limitations of this study by focusing on investigating the accuracy and uncertainty of gap-filled products using other remote sensing datasets and at different spatial and temporal scales. Additionally, the impact of the gap-filling techniques on the accuracy of downstream analyses, such as thermal condition, vegetation indices and land cover classifications, should be investigated. Future studies may need to investigate the accuracy and uncertainty at the regional or global scale. Other topics may need further attention: 1) Machine learning techniques, such as deep learning and artificial neural networks, which could offer better accuracy and un-certainty estimates for gap-filled products. 2) Investigation of the impact of gap-filling techniques on downstream analyses for other landscapes in addition to urban. 3) Evaluation of the accuracy and uncertainty of gap-filling products over longer time periods and larger spatial scales. 4) Development of auto-standardized methods for evaluating gap-filled products by using other existing remote sensing derived LST products. Standardized methods for evaluating the accuracy and uncertainty of gap-filled products would facilitate comparison and benchmarking of different techniques and products.

# 5. Conclusion

# Gap-filling accuracy assessment, uncertainty analysis, and comparison analysis are essential to ensure the reliability of gap-filled Landsat LST products. This paper presents the results of accuracy assessment, uncertainty analysis, and comparison analysis of a new time series gap-filled Landsat LST that is modelled from multi-sensor and multi-time Landsat data harmonization. Landsat LST observations within the ARD tiles without gap filling are not adequate to represent temporal frequency of surface thermal conditions in a time series, resulting in either overestimates or underestimates of their seasonal or annual temporal means. The Landsat LST with gap-filling substantially added temporal density for daily Landsat LST records and can be used to calculate monthly and seasonal Landsat LST. This increased frequency Landsat time-series LST provides an optional temperature source for SUHI monitoring, assessment, and trend analysis. The gap-filled Landsat LST has significant correlations with air temperature recorded from gridded weather records, suggesting similar daily, monthly, and seasonal variation patterns between the two datasets. The data can be used in longtime time-series SUHI and intensity annual, seasonal, even monthly change analysis. Furthermore, we demonstrate that widespread uncertainty is occurring across our study area and this uncertainty is influencing the gap-filled Land-sat LST.

# Our study provides important insights into the accuracy of gap-filling techniques for gap-filled products derived from remote sensing data. By assessing the accuracy of the techniques, we can provide the information of reliability and usefulness of remote sensing data products for various applications. In conclusion, our study shows that gap-filling techniques are effective in filling missing values in remote sensing data. However, the accuracy of the gap-filled data varied depending on the test site and the type of gap-filling models used. The accuracy assessment showed that the models performed better in arid regions than in humid regions. Also, the model results are similar in in summer months in all regions. The uncertainty analysis indicates that the variability in the gap-filled data is higher in the arid and semi-arid regions than in humid region. The variability is larger in the cold region than the warm region during winter months.

# Acknowledgments

The authors thank Dr. Heather J. Tollerud and Mr. Roger Auch for reviewing this manuscript. The manuscript contents do not constitute a statement of endorsement, policy, decision, or position on behalf of USGS. The views expressed in this article are those of the author and do not necessarily represent the views or policies of the USGS. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. Hua Shi’s work was performed under Work performed under USGS contract 140G0119C0001.

Disclosure statement

No potential conflict of interest was reported by the author(s).

# References

Bauer, Marvin E. 2020. "Remote Sensing of Environment: History, Philosophy, Approach and Contributions, 1969 –2019."  *Remote Sensing of Environment* 237:111522. doi: <https://doi.org/10.1016/j.rse.2019.111522>.

Chen, J., X. Zhu, J. E. Vogelmann, F. Gao, and S. Jin. 2011. "A simple and effective method for filling gaps in Landsat ETM+ SLC-off images."  *Remote Sensing of Environment* 115 (4):1053-64. doi: 10.1016/j.rse.2010.12.010.

Duan, Si-Bo, Zhao-Liang Li, Jie Cheng, and Pei Leng. 2017. "Cross-satellite comparison of operational land surface temperature products derived from MODIS and ASTER data over bare soil surfaces."  *ISPRS Journal of Photogrammetry and Remote Sensing* 126:1-10. doi: <https://doi.org/10.1016/j.isprsjprs.2017.02.003>.

Foody, Giles M. 2020. "Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification."  *Remote Sensing of Environment* 239:111630. doi: <https://doi.org/10.1016/j.rse.2019.111630>.

Friedl, M. A., F. W. Davis, J. Michaelsen, and M. A. Moritz. 1995. "Scaling and uncertainty in the relationship between the NDVI and land surface biophysical variables: An analysis using a scene simulation model and data from FIFE."  *Remote Sensing of Environment* 54 (3):233-46. doi: <http://dx.doi.org/10.1016/0034-4257(95)00156-5>.

Gao, F., J. Masek, M. Schwaller, and F. Hall. 2006. "On the blending of the landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance."  *IEEE Transactions on Geoscience and Remote Sensing* 44 (8):2207-18. doi: 10.1109/TGRS.2006.872081.

Gotway, C. A., and L. J. Young. 2002. "Combining incompatible spatial data."  *Journal of the American Statistical Association* 97 (458):632-48. doi: 10.1198/016214502760047140.

Hulley, G., and S. Hook. 2021. "MOD21A2 MODIS/Terra Land Surface Temperature/3-Band Emissivity 8-Day L3 Global 1km SIN Grid V061." In, edited by NASA EOSDIS Land Processes DAAC.

Hulley, G., and S. Hook. 2018. "VIIRS/NPP Land Surface Temperature and Emissivity 8-Day L3 Global 1km SIN Grid V001." In, edited by NASA EOSDIS Land Processes DAAC.

Leyk, S., J. H. Uhl, D. Balk, and B. Jones. 2018. "Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States."  *Remote Sensing of Environment* 204:898-917. doi: 10.1016/j.rse.2017.08.035.

Luo, Y., K. Guan, and J. Peng. 2018. "STAIR: A generic and fully-automated method to fuse multiple sources of optical satellite data to generate a high-resolution, daily and cloud-/gap-free surface reflectance product."  *Remote Sensing of Environment* 214:87-99. doi: 10.1016/j.rse.2018.04.042.

Menne, Matthew J., Byron E. Gleason, Jay Lawrimore, Jared Rennie, and Claude N Williams. 2017. "Global Historical Climatology Network - Monthly Temperature, Version 4 BETA " In, edited by NOAA National Centers for Environmental Information. Washington D.C.

Murphy, J. M., D. M. H. Sexton, D. H. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth. 2004. "Quantification of modelling uncertainties in a large ensemble of climate change simulations."  *Nature* 430 (7001):768-72. doi: 10.1038/nature02771.

Niclòs, Raquel, Joan M. Galve, Jose A. Valiente, Maria J. Estrela, and César Coll. 2011. "Accuracy assessment of land surface temperature retrievals from MSG2-SEVIRI data."  *Remote Sensing of Environment* 115 (8):2126-40. doi: <https://doi.org/10.1016/j.rse.2011.04.017>.

NWS. "National Weather Service." NWS, NOAA, Accessed May, 01, 2022. <https://www.weather.gov/>.

Rocchini, D., G. M. Foody, H. Nagendra, C. Ricotta, M. Anand, K. S. He, V. Amici, et al. 2013. "Uncertainty in ecosystem mapping by remote sensing."  *Computers and Geosciences* 50:128-35. doi: 10.1016/j.cageo.2012.05.022.

Rounsevell, M. D. A., A. Arneth, C. Brown, W. W. L. Cheung, O. Gimenez, I. Holman, P. Leadley, et al. 2021. "Identifying uncertainties in scenarios and models of socio-ecological systems in support of decision-making."  *One Earth* 4 (7):967-85. doi: 10.1016/j.oneear.2021.06.003.

Roy, D. P., J. Ju, P. Lewis, C. Schaaf, F. Gao, M. Hansen, and E. Lindquist. 2008. "Multi-temporal MODIS-Landsat data fusion for relative radiometric normalization, gap filling, and prediction of Landsat data."  *Remote Sensing of Environment* 112 (6):3112-30. doi: 10.1016/j.rse.2008.03.009.

Roy, D. P., and L. Yan. 2020. "Robust Landsat-based crop time series modelling."  *Remote Sensing of Environment* 238:110810. doi: <https://doi.org/10.1016/j.rse.2018.06.038>.

Stehman, Stephen V., Cidália C. Fonte, Giles M. Foody, and Linda See. 2018. "Using volunteered geographic information (VGI) in design-based statistical inference for area estimation and accuracy assessment of land cover."  *Remote Sensing of Environment* 212:47-59. doi: <https://doi.org/10.1016/j.rse.2018.04.014>.

Stehman, Stephen V., and Giles M. Foody. 2019. "Key issues in rigorous accuracy assessment of land cover products."  *Remote Sensing of Environment* 231:111199. doi: <https://doi.org/10.1016/j.rse.2019.05.018>.

Thornton, Peter E., Rupesh Shrestha, Michele Thornton, Shih-Chieh Kao, Yaxing Wei, and Bruce E. Wilson. 2021. "Gridded daily weather data for North America with comprehensive uncertainty quantification."  *Scientific Data* 8 (1):190. doi: 10.1038/s41597-021-00973-0.

Wang, Jing, Calvin K. F. Lee, Xiaolin Zhu, Ruyin Cao, Yating Gu, Shengbiao Wu, and Jin Wu. 2022. "A new object-class based gap-filling method for PlanetScope satellite image time series."  *Remote Sensing of Environment* 280:113136. doi: <https://doi.org/10.1016/j.rse.2022.113136>.

Wulder, M. A., J. C. White, J. G. Masek, J. Dwyer, and D. P. Roy. 2011. "Continuity of Landsat observations: Short term considerations."  *Remote Sensing of Environment* 115 (2):747-51. doi: 10.1016/j.rse.2010.11.002.

Wulder, Michael A., David P. Roy, Volker C. Radeloff, Thomas R. Loveland, Martha C. Anderson, David M. Johnson, Sean Healey, et al. 2022. "Fifty years of Landsat science and impacts."  *Remote Sensing of Environment* 280:113195. doi: <https://doi.org/10.1016/j.rse.2022.113195>.

Yan, L., and D. P. Roy. 2018. "Large-area gap filling of landsat reflectance time series by spectral-angle-mapper based spatio-temporal similarity (SAMSTS)."  *Remote Sensing* 10 (4). doi: 10.3390/rs10040609.

Zhang, Xiaoyang, Jianmin Wang, Geoffrey M. Henebry, and Feng Gao. 2020. "Development and evaluation of a new algorithm for detecting 30 m land surface phenology from VIIRS and HLS time series."  *ISPRS Journal of Photogrammetry and Remote Sensing* 161:37-51. doi: <https://doi.org/10.1016/j.isprsjprs.2020.01.012>.

Zhou, Q., G. Xian, and H. Shi. 2020. "Gap fill of land surface temperature and reflectance products in landsat analysis ready data."  *Remote Sensing* 12 (7). doi: 10.3390/rs12071192.

Zhou, Qiang, Zhe Zhu, George Xian, and Congcong Li. 2022. "A novel regression method for harmonic analysis of time series."  *ISPRS Journal of Photogrammetry and Remote Sensing* 185:48-61. doi: <https://doi.org/10.1016/j.isprsjprs.2022.01.006>.

Zhu, X., J. Chen, F. Gao, X. Chen, and J. G. Masek. 2010. "An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions."  *Remote Sensing of Environment* 114 (11):2610-23. doi: 10.1016/j.rse.2010.05.032.

Zhu, X., E. H. Helmer, F. Gao, D. Liu, J. Chen, and M. A. Lefsky. 2016. "A flexible spatiotemporal method for fusing satellite images with different resolutions."  *Remote Sensing of Environment* 172:165-77. doi: 10.1016/j.rse.2015.11.016.

Zhu, Xiaolin, Wenfeng Zhan, Junxiong Zhou, Xuehong Chen, Zicong Liang, Shuai Xu, and Jin Chen. 2022. "A novel framework to assess all-round performances of spatiotemporal fusion models."  *Remote Sensing of Environment* 274:113002. doi: <https://doi.org/10.1016/j.rse.2022.113002>.

Table 1. Main data sources used in the study.

Table 2, NOAA GHCN Station ID (Figure 1), Latitude (Lat.) and Longitude (Lon.), Name, and Land Cover Class (LC, 2020) in three study areas.

Table 3. The correlation (R2) and Root Mean Square Error (RMSE) in selected NOAA GHCN stations for annual accuracy assessment, Atlanta, GA.

Table 4. The correlation (R2) and Root Mean Square Error (RMSE) in selected NOAA GHCN stations for annual accuracy assessment, Sioux Falls, SD.

Table 5. The correlation (R2) and Root Mean Square Error (RMSE) in selected NOAA GHCN stations for annual accuracy assessment, Phoenix, AZ.

Table 6. The number of Landsat clear observations and gap-filled uncertainty (°C) in 2016 and 2020.

Table 7. The correlation (R2) and Root Mean Square Error (RMSE), selected NOAA GHCN stations for comparison analysis among GHCN, gap-filled LST, MODIS LST, VIIRS LST, and ECOSTRESS LST by daily, monthly, and seasonal for three study areas in year 2016 and 2020.

Figure 1. The zoom in 2020 land cover map for three selected study areas within the ARD tiles, GHCN station location are white-purple boxes associate with the ID number in black (the ID, name, land cover class, and detailed information of each station see Table 2). These Land cover map of individual urban centre are not at the same scale.

Figure 2. Workflow of Assessing gap-filled Landsat land surface temperature time series data using different observational datasets.

Figure 3. Gap-filled Landsat LST (top) and Original Landsat (5, 7, and 8) LST (bottom) in Atlanta, GA in 1991, 2000, 2016 and 2020 from left to right.

Figure 4. Gap-filled Landsat LST (top) and Original Landsat (5, 7, and 8) LST (bottom) in Sioux Falls, SD in 1991, 2000, 2016 and 2020 from left to right.

Figure 5. Gap-filled Landsat LST (top) and Original Landsat (5, 7, and 8) LST (bottom) in Phoenix, AZ in 1991, 2000, 2016 and 2020 from left to right.

Figure 6. Time series of land surface temperature and air temperature at 12 stations in Atlanta, GA. The blue boxes are valid Landsat observations, and red dots are gap-filled values at each Landsat acquisition date. The black triangles are station air temperature that match all Landsat acquisition dates within an ARD tile in Atlanta, GA (<https://www.ncdc.noaa.gov/ghcn-daily-description>). Information of GHCN stations see the Figure 1 and Table 2.

Figure 7. Time series of land surface temperature and air temperature at 12 stations in Sioux Falls, SD. The blue boxes are valid Landsat observations, and red dots are gap-filled values at each Landsat acquisition date. The black triangles are station air temperature that match all Landsat acquisition dates within an ARD tile in Atlanta, GA (<https://www.ncdc.noaa.gov/ghcn-daily-description>). Information of GHCN stations see the Figure 1 and Table 2.

Figure 8. Time series of land surface temperature and air temperature at 12 stations in Phoenix AZ. The blue boxes are valid Landsat observations, and red dots are gap-filled values at each Landsat acquisition date. The black triangles are station air temperature that match all Landsat acquisition dates within an ARD tile in Atlanta, GA (<https://www.ncdc.noaa.gov/ghcn-daily-description>). Information of GHCN stations see the Figure 1 and Table 2.

Figure 9. Annual clear observation (top) and annual mean of gap-filled uncertainty (bottom) for three study areas in year 2016 and 2020.

Figure 10. Uncertainty of gap-filling models in the three study areas: Atlanta (A), Sioux Falls (B), and Phoenix (C) by selected NOAA GHCN validation stations in 2020 (top) and 2016 (bottom). X axis is dates of year and Y axis is uncertainty values from the gap-filling model. The information of GHCN stations see Figure 1 and Table 2.

Table 1. Main data sources used in the study.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Type** | **Resolution** | **Temporal** | **Spectral accuracy** | **Source** |
| Landsat ARD |  |  |  | ~0.5 kelvin (vary |  |
| LST Collection 1 | LST | 30 m | 7 days | by pixel) | USGS |
| GHCN | Air Temp. | points | Daily | - | NOAA |
| MODIS LST | LST | 1000 m | Weekly | 1.5~2.5 kelvin | NASA |
| VIIRS LST | LST | 1000 m | Weekly | 1.5~2.5 kelvin | NASA |
| ECOSTRESS | LST | 70 m | Daily | 1~2 kelvin | NASA |
| DAYMET | LST | 1000 m | monthly | - | ORNL |

Table 2, NOAA GHCN Station ID (Figure 1), Latitude (Lat.) and Longitude (Lon.), Name, and Land Cover Class (LC, 2020) in three study areas.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atlanta, GA | | | | | Sioux Falls, SD | | | | | Phoenix, AZ | | | | |
| ID | Name | Lat. | Lon. | LC\*\* | ID | Station Name | Lat. | Lon. | LC | ID | Station Name | Lat. | Lon. | LC |
| 1 | USC00091640 | 33.60 | -85.08 | 90 | 22 | USC00211263 | 43.50 | -96.70 | 81 | 50 | USC00020632 | 33.81 | -111.65 | 52 |
| 2 | USC00092485 | 34.00 | -84.75 | 21 | 23 | USC00216565 | 43.46 | -96.73 | 81 | 50 | USW00093139 | 33.82 | -111.90 | 52 |
| 3 | USC00099486 | 34.30 | -83.86 | 81 | 25 | USC00390128 | 43.48 | -96.76 | 21 | 51 | USC00021282 | 33.60 | -111.71 | 52 |
| 4 | USC00098740 | 33.33 | -83.70 | 22 | 26 | USC00391032 | 43.52 | -96.67 | 22 | 52 | USC00023190 | 33.55 | -111.44 | 22 |
| 5 | USW00013874 | 33.63 | -84.44 | 24 | 27 | USC00391076 | 43.54 | -96.68 | 81 | 53 | USC00025700 | 33.56 | -111.54 | 52 |
| 6 | USC00091665 | 34.20 | -84.79 | 21 | 28 | USC00392302 | 43.53 | -96.81 | 23 | 54 | USC00028214 | 33.43 | -111.92 | 90 |
| 7 | USC00092180 | 34.26 | -83.49 | 22 | 33 | USC00392984 | 43.54 | -96.84 | 21 | 55 | USC00028499 | 33.60 | -112.30 | 21 |
| 8 | USC00096335 | 33.40 | -84.91 | 71 | 35 | USC00394037 | 43.58 | -96.80 | 21 | 56 | USC00029634 | 33.43 | -112.00 | 24 |
| 9 | USC00092006 | 34.17 | -84.73 | 43 | 36 | USW00014944 | 43.52 | -96.75 | 24 | 57 | USW00023183 | 33.34 | -112.15 | 22 |
| 10 | USC00098950 | 33.87 | -83.54 | 81 | 37 | USC00395090 | 43.50 | -96.79 | 81 | 58 | USC00027281 | 33.46 | -111.48 | 52 |
| 11 | USC00092318 | 33.60 | -83.84 | 21 | 40 | USC00399042 | 43.53 | -96.75 | 21 | 59 | USC00020288 | 33.50 | -112.36 | 21 |
| 12 | USC00094700 | 33.53 | -84.35 | 24 | 41 | USC00390422 | 43.50 | -96.67 | 71 | 60 | USC00025521 | 33.11 | -112.03 | 22 |
| 13 | USW00053819 | 33.36 | -84.57 | 21 | 44 | USW00094950 | 43.55 | -96.66 | 82 | 61 | USC00025270 | 33.07 | -111.77 | 71 |
| 14 | USC00099466 | 33.93 | -83.73 | 81 | 45 | USC00390281 | 43.59 | -96.73 | 21 | 62 | USC00027370 | 33.21 | -111.68 | 24 |
| 15 | USW00003888 | 33.78 | -84.52 | 21 | 45 | USC00397666 | 43.51 | -96.67 | 81 | 63 | USC00021514 | 33.38 | -112.07 | 21 |
| 16 | USW00053838 | 34.27 | -83.83 | 22 | 46 | USC00395671 | 43.55 | -96.63 | 95 | 64 | USC00028112 | 33.69 | -112.08 | 22 |
| 17 | USW00053863 | 33.88 | -84.30 | 23 |  |  |  |  |  | 65 | USW00003184 | 33.78 | -112.52 | 21 |
| 18 | USC00093271 | 33.26 | -84.28 | 23 |  |  |  |  |  | 66 | USC00029464 | 33.62 | -111.91 | 21 |
| 19 | USW00053873 | 34.12 | -84.85 | 23 |  |  |  |  |  | 67 | USW00003192 | 33.48 | -111.93 | 24 |
| 20 | USC00092283 | 34.23 | -84.13 | 23 |  |  |  |  |  |  |  |  |  |  |

Table 3. The correlation (R2) and Root Mean Square Error (RMSE) in selected NOAA GHCN stations for annual accuracy assessment, Atlanta, GA.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Station name** | **Year** | **LC**\* | | | **No. Obs.** # | | | R2 | | RMSE | | |
|  |  |  | | | **G**+ **(O**^**)** | G | O vs. V‡ | | G vs. V | | O vs. V | G vs. V |
| USC00092283 |  | 22 | | | 23 (19) | 31 | 0.70 | | 0.79 | | 6.31 | 6.41 |
| USC00099466 | 1991 | 81 | | | 22 (18) | 31 | 0.84 | | 0.86 | | 4.57 | 3.92 |
| USC00098740 |  | 21 | | | 14 (12) | 31 | 0.77 | | 0.78 | | 6.44 | 6.96 |
| USC00098950 |  | 81 | | | 30 (28) | 79 | 0.77 | | 0.84 | | 7.57 | 5.98 |
| USC00092180 | 2000 | | 22 | 38 (38) | | 79 | 0.63 | | 0.76 | | 7.77 | 6.13 |
| SW00053863 |  | | 24 | 28 (23) | | 79 | 0.83 | | 0.81 | | 9.33 | 11.01 |
| USC00099486 |  | | 81 | 24 (26) | | 86 | 0.74 | | 0.81 | | 5.61 | 2.88 |
| USW00053838 | 2016 | | 23 | 45 (45) | | 86 | 0.88 | | 0.82 | | 5.64 | 5.68 |
| USC00094700 |  | | 22 | 24 (24) | | 86 | 0.70 | | 0.85 | | 6.16 | 8.56 |
| USC00092006 |  | | 41 | 60 (38) | | 83 | 0.48 | | 0.72 | | 7.81 | 3.49 |
| USC00092283 | 2020 | | 23 | 56 (29) | | 83 | 0.44 | | 0.80 | | 9.84 | 5.81 |
| USC00091965 |  | | 23 | 62 (38) | | 83 | 0.36 | | 0.82 | | 12.47 | 6.09 |

Table 4. The correlation (R2) and Root Mean Square Error (RMSE) in selected NOAA GHCN stations for annual accuracy assessment, Sioux Falls, SD.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Station name** | **Year** | **LC** | | **No. Obs.** | | | R2 | | RMSE | | |
|  |  | |  | **G (O)** | G | O vs. V | | G vs. V | | O vs. V | G vs. V |
| USW00014944 |  | | 24 | 12 (9) | 38 | 0.94 | | 0.79 | | 4.34 | 7.19 |
| USC00395090 | 1991 | | 81 | 9 (9) | 31 | 0.88 | | 0.75 | | 6.04 | 9.74 |
| USC00390128 |  | | 21 | 17 (17) | 38 | 0.98 | | 0.86 | | 3.26 | 6.24 |
| USC00216565 |  | | 81 | 43 (32) | 96 | 0.89 | | 0.58 | | 6.14 | 10.44 |
| USC00390281 | 2000 | | 21 | 34 (25) | 96 | 0.88 | | 0.74 | | 5.09 | 5.29 |
| USC00392984 |  | | 23 | 18 (12) | 96 | 0.84 | | 0.56 | | 6.37 | 10.74 |
| USW00014944 |  | | 24 | 25 (14) | 117 | 0.87 | | 0.76 | | 6.43 | 8.14 |
| USC00211263 | 2016 | | 81 | 66 (22) | 117 | 0.73 | | 0.59 | | 5.48 | 10.31 |
| USC00391076 |  | | 23 | 66 (16) | 117 | 0.73 | | 0.78 | | 4.27 | 8.37 |
| USC00391032 |  | | 22 | 58 (25) | 125 | 0.59 | | 0.74 | | 11.44 | 9.85 |
| USC00395090 | 2020 | | 81 | 26 (13) | 125 | 0.57 | | 0.68 | | 11.48 | 8.48 |
| USW00014944 |  | | 24 | 30 (13) | 125 | 0.77 | | 0.70 | | 9.23 | 8.31 |

Table 5. The correlation (R2) and Root Mean Square Error (RMSE) in selected NOAA GHCN stations for annual accuracy assessment, Phoenix, AZ.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Station name** | **Year** | **LC** | | **No. Obs.** | | | R2 | | RMSE | | |
|  |  | |  | **G (O)** | G | O vs. V | | G vs. V | | O vs. V | G vs. V |
| USW00023183 |  | | 22 | 16 (8) | 31 | 0.85 | | 0.71 | | 9.07 | 9.37 |
| USW00093139 | 1991 | | 52 | 27 (13) | 31 | 0.41 | | 0.81 | | 9.28 | 7.52 |
| USC00025270 |  | | 71 | 26 (16) | 31 | 0.77 | | 0.72 | | 10.22 | 9.85 |
| USC00021282 |  | | 52 | 34 (33) | 83 | 0.87 | | 0.88 | | 10.43 | 11.38 |
| USC00023190 | 2000 | | 22 | 61 (60) | 78 | 0.84 | | 0.85 | | 5.86 | 5.82 |
| USC00025512 |  | | 23 | 33 (31) | 83 | 0.80 | | 0.79 | | 8.70 | 8.79 |
| USW00093139 |  | | 52 | 66 (54) | 89 | 0.73 | | 0.79 | | 5.32 | 5.89 |
| USW00023183 | 2016 | | 22 | 66 (30) | 89 | 0.73 | | 0.79 | | 7.15 | 6.13 |
| USW00003192 |  | | 24 | 46 (41) | 89 | 0.81 | | 0.85 | | 12.15 | 12.52 |
| USC00020288 |  | | 24 | 64 (53) | 88 | 0.91 | | 0.89 | | 6.94 | 9.28 |
| USC00027281 | 2020 | | 52 | 61 (37) | 88 | 0.59 | | 0.89 | | 13.38 | 7.59 |
| USC00028499 |  | | 24 | 74 (60) | 88 | 0.59 | | 0.85 | | 9.67 | 7.75 |

Table 6. The number of Landsat clear observations and gap-filled uncertainty (°C) in 2016 and 2020.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Atlanta, GA (H24V13)** | | **Sioux Falls, SD (H16V06)** | | **Phoenix, AZ (H07V13)** | |
| **Year** | **Class\*** | **Clear Obs.** | **Uncertainty** | **Clear Obs.** | **Uncertainty** | **Clear Obs.** | **Uncertainty** |
|  | H | 1 | 2.25 | 1 | 4.65 | 4 | 2.34 |
| 2016 | M | 30 | 0.65 | 25 | 0.94 | 41 | 0.76 |
|  | L | 61 | 0.3 | 46 | 0.3 | 70 | 0.03 |
|  | H | 1 | 4.95 | 1 | 4.71 | 1 | 3.68 |
| 2020 | M | 21 | 0.97 | 22 | 1.08 | 45 | 0.68 |
|  | L | 48 | 0.3 | 40 | 0.3 | 79 | 0.3 |

Table 7. The correlation (R2) and Root Mean Square Error (RMSE), selected NOAA GHCN stations for comparison analysis among GHCN, gap-filled LST, MODIS LST, VIIRS LST, and ECOSTRESS LST by daily, monthly, and seasonal for three study areas in year 2016 and 2020.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Station** | **Frequency** | | **No.**  **Sample** | | **GHCN vs.**  **Gap-filled** | | | **GHCN vs.**  **VIIRS (MODIS)** | | | **GHCN vs.**  **ECOSTRESS** | |
|  |  | |  | RMSE | | R2 | RMSE | | R2 | RMSE | | R2 |
| USC00094648 | Daily | 84 | | 5.77 | | 0.76 | 6.35 | | 0.64 | 10.34 | | 0.42 |
| Atlanta, GA | Monthly | 12 | | 7.0 | | 0.97 | 5.31 | | 0.92 | 8.72 | | 0.61 |
| 2020 (Urban) | Seasonal | 4 | | 6.05 | | 0.98 | 4.89 | | 0.96 | 8.49 | | 0.74 |
| USC00097827 | Daily | 84 | | 5.27 | | 0.77 | 4.88 | | 0.66 | 4.51 | | 0.30 |
| Atlanta, GA | Monthly | 12 | | 4.69 | | 0.98 | 2.25 | | 0.94 | 4.87 | | 0.49 |
| 2020 (Forest) | Seasonal | 4 | | 3.68 | | 0.99 | 1.91 | | 0.94 | 5.05 | | 0.46 |
| USC00098740 | Daily | 87 | | 6.35 | | 0.88 | 7.51 | | 0.87 | - | | - |
| Atlanta, GA | Monthly | 12 | | 7.60 | | 0.92 | 7.68 | | 0.91 | - | | - |
| 2016 (Urban) | Season | 4 | | 7.79 | | 0.94 | 10.79 | | 0.95 | - | | - |
| USW00014944 | Daily | 125 | | 8.45 | | 0.69 | 7.57 | | 0.87 | 10.96 | | 0.57 |
| Sioux Falls, SD | Monthly | 12 | | 6.07 | | 0.87 | 7.16 | | 0.94 | 6.45 | | 0.80 |
| 2020 (Urban) | Season | 4 | | 4.02 | | 0.99 | 6.66 | | 0.97 | 13.53 | | 0.87 |
| USC00391076 | Daily | 125 | | 9.25 | | 0.67 | 8.05 | | 0.78 | 10.75 | | 0.54 |
| Sioux Falls, SD | Monthly | 12 | | 9.25 | | 0.85 | 8.05 | | 0.90 | 10.75 | | 0.77 |
| 2020 (Agri.) | Season | 4 | | 5.8 | | 0.93 | 6.71 | | 0.93 | 12.55 | | 0.87 |
| USW00094950 | Daily | 117 | | 6.7 | | 0.80 | 5.53 | | 0.84 | - | | - |
| Sioux Falls, SD | Monthly | 12 | | 4.93 | | 0.95 | 2.94 | | 0.96 | - | | - |
| 2016 (Urban) | Seasonal | 4 | | 2.70 | | 0.99 | 1.03 | | 0.98 | - | | - |
| USC00023190 | Daily | 89 | | 4.93 | | 0.87 | 11.38 | | 0.92 | 11.05 | | 0.67 |
| Phoenix, AZ | Monthly | 12 | | 4.43 | | 0.98 | 11.16 | | 0.94 | 13.60 | | 0.43 |
| 2020 (Urban) | Season | 4 | | 2.50 | | 0.98 | 6.31 | | 0.96 | 13.70 | | 0.18 |
| USC00027281 | Daily | 89 | | 7.04 | | 0.90 | 9.31 | | 0.89 | 9.93 | | 0.66 |
| Phoenix, AZ | Monthly | 12 | | 7.62 | | 0.99 | 8.63 | | 0.96 | 10.02 | | 0.88 |
| 2020 (Shrub) | Season | 4 | | 3.97 | | 0.99 | 4.93 | | 0.98 | 9.59 | | 0.82 |
| USW00003192 | Daily | 89 | | 6.14 | | 0.89 | 8.92 | | 0.82 | - | | - |
| Phoenix, AZ | Monthly | 12 | | 5.85 | | 0.97 | 7.87 | | 0.95 | - | | - |
| 2016 (Urban) | Season | 4 | | 2.86 | | 0.88 | 3.34 | | 0.87 | - | | - |

Map

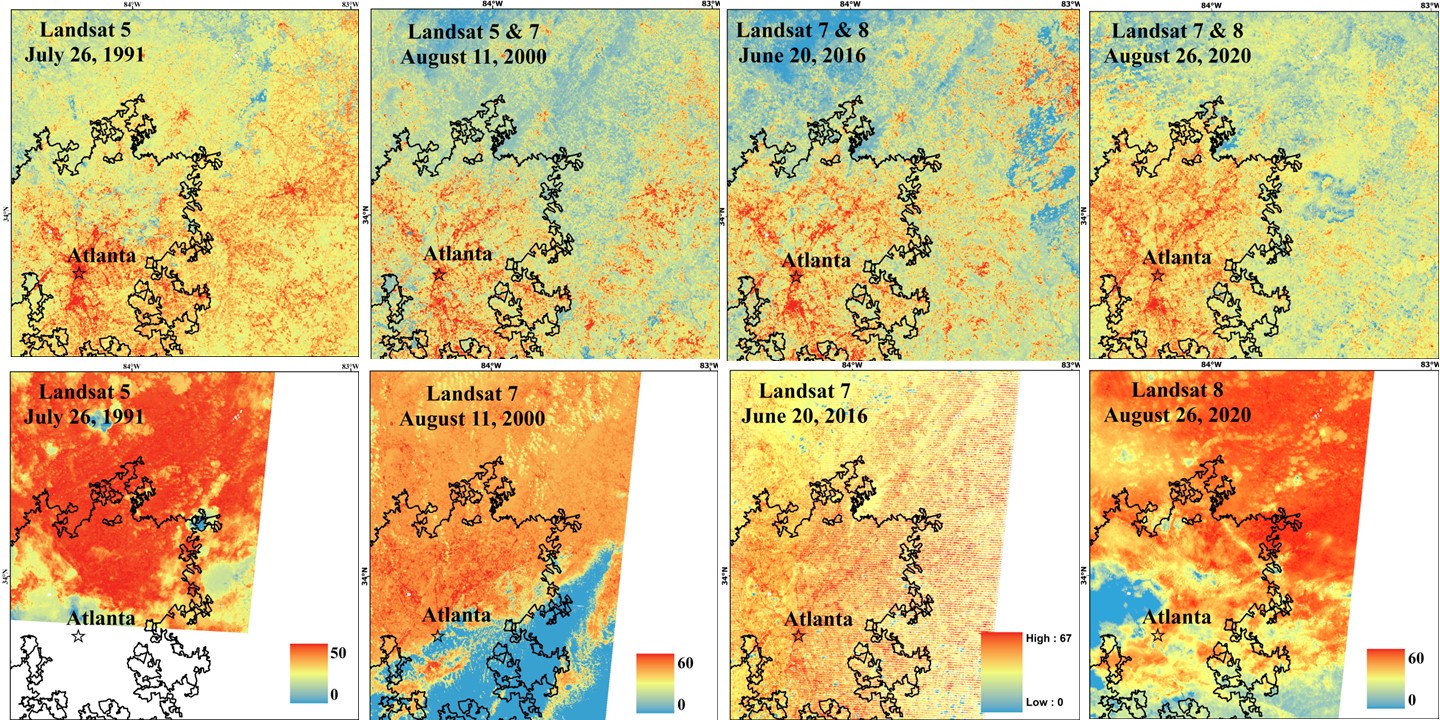
Description automatically generated

Figure 1. The zoom in 2020 land cover map for three selected study areas within the ARD tiles, GHCN station location are white-purple boxes associate with the ID number in black (the ID, name, land cover class, and detailed information of each station see Table 2). These Land cover map of individual urban centre are not at the same scale.

Diagram

Description automatically generated

Figure 2. Workflow of assessing gap-filled Landsat land surface temperature time series data using different observational datasets. \*Models details see (Zhou, Xian, and Shi 2020)

Figure 3. Gap-filled Landsat LST (top) and Original Landsat (5, 7, and 8) LST (bottom) in Atlanta, GA on the selected dates of 1991, 2000, 2016 and 2020 from left to right.

Map

Description automatically generatedFigure 4. Gap-filled Landsat LST (top) and Original Landsat (5, 7, and 8) LST (bottom) in Sioux Falls, SD on the selected dates of 1991, 2000, 2016 and 2020 from left to right.

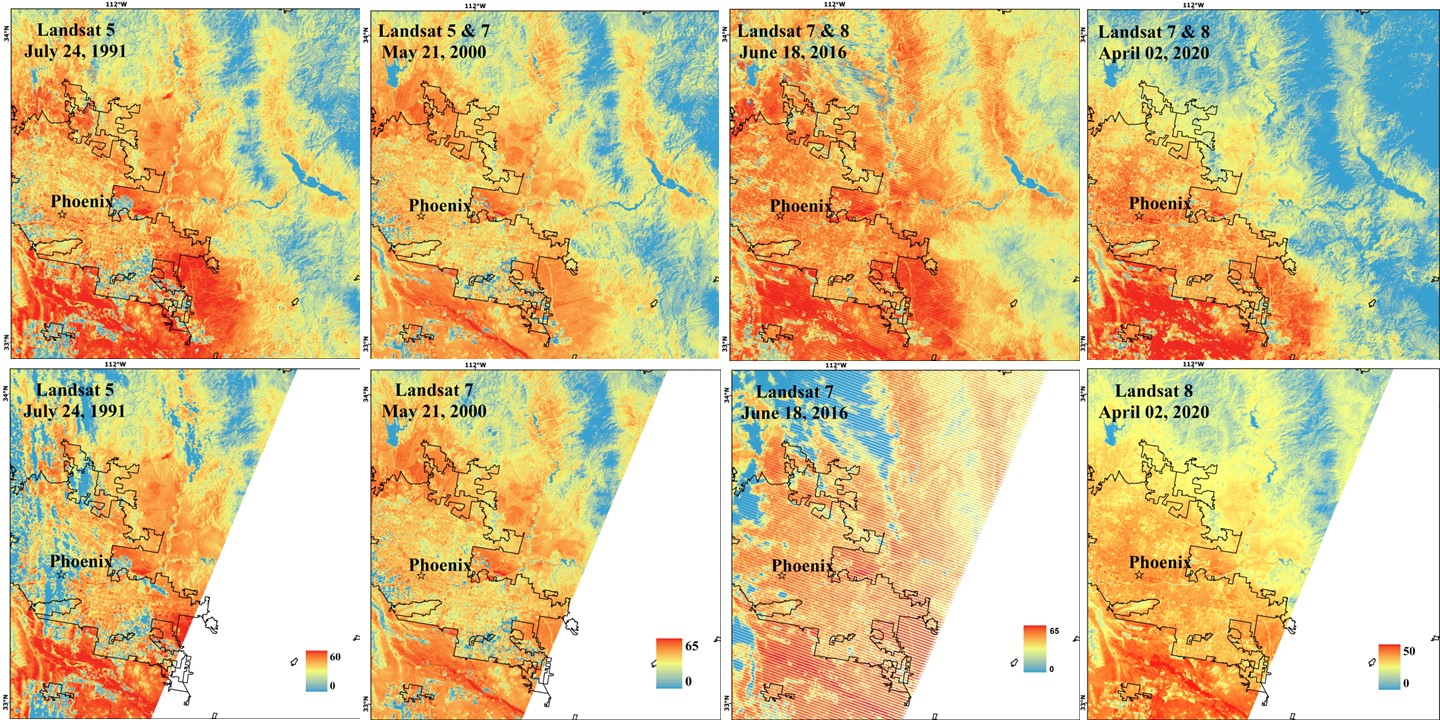
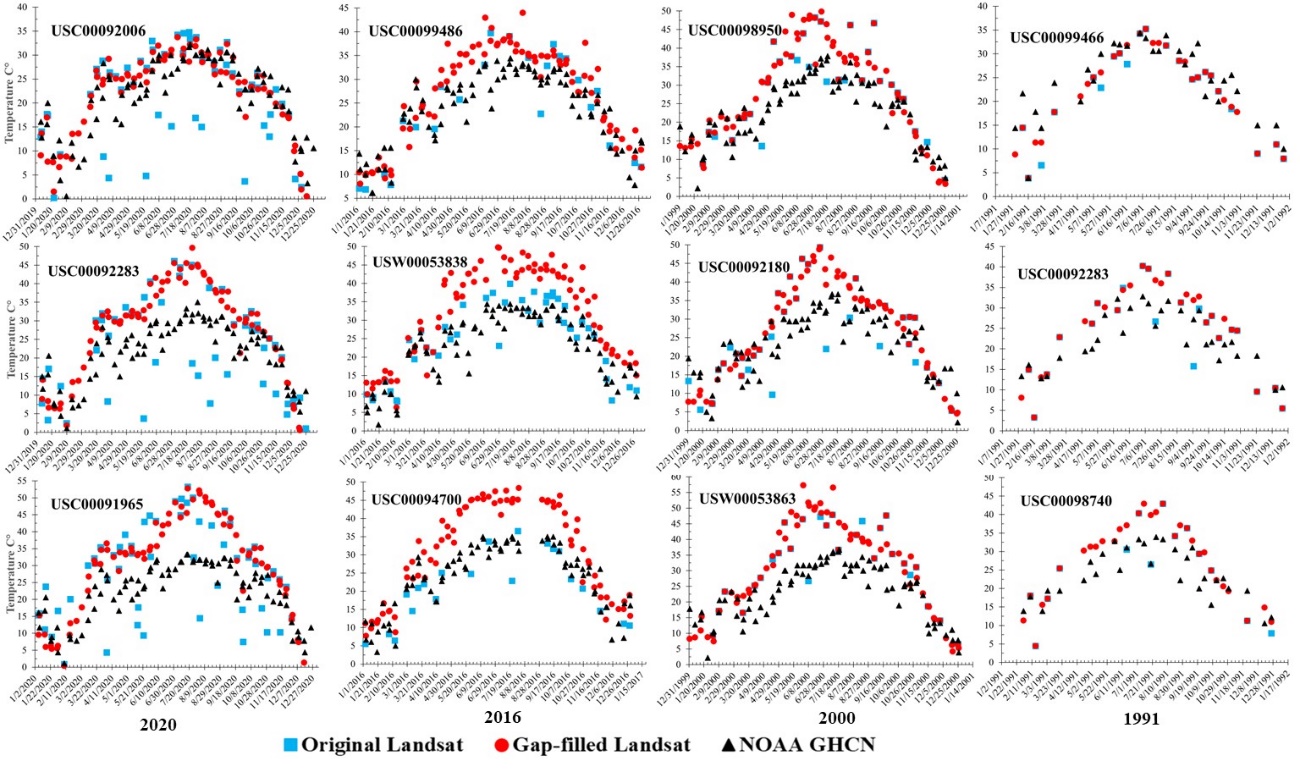


Figure 5. Gap-filled Landsat LST (top) and Original Landsat (5, 7, and 8) LST (bottom) in Phoenix, AZ on the selected dates of 1991, 2000, 2016 and 2020 from left to right.

Figure 6. Time series of land surface temperature and air temperature at 12 stations in Atlanta, GA. The blue boxes are valid Landsat observations, and red dots are gap-filled values at each Landsat acquisition date. The black triangles are station air temperature that match all Landsat acquisition dates within an ARD tile in Atlanta, GA (<https://www.ncdc.noaa.gov/ghcn-daily-description>). Information of GHCN stations see the Figure 1 and Table 2.

Chart

Description automatically generatedFigure 7. Time series of land surface temperature and air temperature at 12 stations in Sioux Falls, SD. The blue boxes are valid Landsat observations, and red dots are gap-filled values at each Landsat acquisition date. The black triangles are station air temperature that match all Landsat acquisition dates within an ARD tile in Sioux Falls, SD (<https://www.ncdc.noaa.gov/ghcn-daily-description>). Information of GHCN stations see the Figure 1 and Table 2.

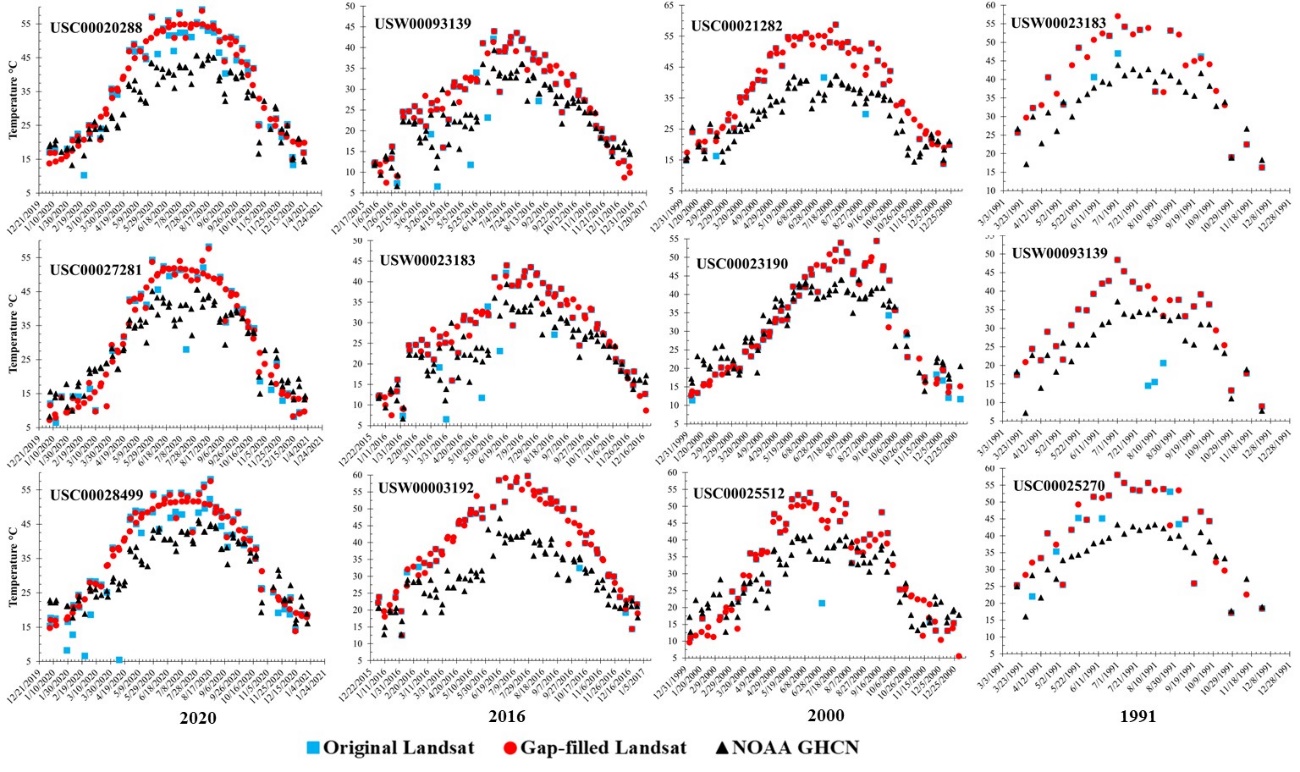


Figure 8. Time series of land surface temperature and air temperature at 12 stations in Phoenix AZ. The blue boxes are valid Landsat observations, and red dots are gap-filled values at each Landsat acquisition date. The black triangles are station air temperature that match all Landsat acquisition dates within an ARD tile in Phoenix, AZ (<https://www.ncdc.noaa.gov/ghcn-daily-description>). Information of GHCN stations see the Figure 1 and Table 2.

Timeline, map

Description automatically generated

Figure 9. Annual clear observation (top) and annual mean of gap-filled uncertainty (bottom) for three study areas in year 2016 and 2020.

Diagram, engineering drawing

Description automatically generatedFigure 10. Uncertainty of gap-filling models in the three study areas: Atlanta (A), Sioux Falls (B), and Phoenix (C) by selected NOAA GHCN validation stations in 2020 (top) and 2016 (bottom). X axis is dates of year and Y axis is uncertainty values from the gap-filling model. The information of GHCN stations see Figure 1 and Table 2.