Estimating Daily Average Surface Air Temperature Using Satellite Land Surface Temperature and Top-of-Atmosphere Radiation Products over the Tibetan Plateau

Yuhan Rao[1,2], Shunlin Liang[1,*], Dongdong Wang[1], Yunyue Yu[3], Zhen Song[4], Yuan Zhou[1], Miaogen Shen[5], Baiqing Xu[5]

1. Department of Geographical Sciences, University of Maryland, College Park, Maryland, USA

2. Cooperative Institute for Climate and Satellites-Maryland, University of Maryland, College Park, Maryland, USA

3. Center for Satellite Applications and Research, NESDIS, NOAA, College Park, Maryland, USA

4. State Grid Economic and Technological Research Institute, Beijing, China

5. Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China

Corresponding author: Shunlin Liang, E-mail: sliang@umd.edu

Abstract: The Tibetan Plateau (TP) has experienced rapid warming in recent decades. However, the meteorological stations of the TP are scarce and mostly located at the eastern and southern parts of the TP where the elevation is relatively low, which increases the uncertainty of regional and local climate studies. Recently, the remotely sensed land surface temperature (LST) has been used to estimate the surface air temperature (SAT). However, the thermal infrared based LST is prone to cloud contamination, which limits the availability of the estimated SAT. This study presents a novel all sky model based on the rule-based Cubist regression to estimate all sky daily average SAT using LST, incident solar radiation (ISR), top-of-atmosphere (TOA) albedo and outgoing longwave radiation (OLR). The model is trained using station data of the Chinese Meteorological Administration (CMA) and corresponding satellite products. The output is evaluated using independent station data with the bias of -0.07 °C and RMSE of 1.87°C. Additionally, the 25-fold cross validation shows a stable model performance (RMSE: 1.6-2.8 °C). Moreover, the all sky Cubist model increases the availability of the estimated SAT by nearly three times. We used the all sky Cubist model to estimate the daily average SAT of the TP for 2002-2016 at 0.05°×0.05°. We compared our all sky Cubist model estimated daily average SAT with three reanalysis datasets (i.e., GLDAS, CLDAS, and CMFD). Our model estimation shows similar spatial and temporal dynamics with these existing data but outperforms them with lower bias and RMSE when benchmarked against CMA station data. The estimated SAT data could be very useful for regional and local climate studies over the TP. Although the model is developed for the TP, the framework is generic and may be extended to other regions with proper model training using local data.

Key words: Surface air temperature, land surface temperature, Tibetan Plateau, radiation, rule-based Cubist regression

1. Introduction

The Tibetan Plateau (TP) is the world's highest plateau in central Asia with an average elevation higher than 4000 meters above sea level (ASL) (Figure 1(a)) (Yang et al., 2014). As the world's "Third Pole", it is the origins of major rivers in Asia and regulates regional and global weather patterns (Yao et al., 2018). Many previous studies reported that the Tibetan Plateau, similar with other high mountainous areas, has experienced more rapid surface temperature change comparing to many other parts of the world, especially after early 1950s (Duan and Xiao, 2015; Pepin et al., 2015; Rangwala and Miller, 2012; Yao et al., 2018). The reported warming exists for both mean, minimum, and maximum surface air temperatures (SAT), leading to the decreasing diurnal temperature range of the TP (Duan and Xiao, 2015; Li et al., 2005; Liu et al., 2009; Liu and Chen, 2000; Yang and Ren, 2017). As a consequence of the SAT change, the TP has shown remarkable changes of its cryosphere, hydrological cycles, and ecosystems. For example, Shen et al. (2015a) reported that the snow cover of the TP has reduced by 5.7% during 1997-2012; Yang et al. (2014) demonstrated that the central TP experiences more convective precipitation and more surface runoff while the southern and eastern regions experience reduction in both precipitation and surface run off in recent decades; multiple studies observed that the vegetation activity shows strong response to surface temperature change over the TP (Cong et al., 2017; Shen et al., 2015b; Shen et al., 2016).

---------- Insert Figure 1 here ----------

However, previous studies heavily rely on SAT data measured by unevenly distributed meteorological stations (Figure 1(a)). Figure 1(b) shows the elevation distribution of the CMA stations and the elevation distribution of the radar-based digital elevation model (DEM) of the entire TP. Over 70% of the CMA stations are located at relatively low elevation (< 4,000 meters ASL) and the eastern part of the TP, while almost no operational CMA stations are placed beyond 5,000 meters ASL. The sparse and biased station samples may increase the uncertainty of local and regional climate analysis and corresponding impact studies (Pepin et al., 2015; Rao et al., 2018; Yao et al., 2018). Alternatively, previous studies also use several spatially complete SAT datasets which are produced through either data interpolation or data assimilation, such as, Global Land Data Assimilation System (GLDAS) data, Chinese Meteorological Forcing Data (CMFD), Chinese Land Data Assimilation System (CLDAS), and Climate Research Unit (CRU) high resolution climate dataset. These datasets provide important information for the regions with no station measurements. However, the spatial resolutions of these datasets (except CLDAS) are very coarse (0.25° ~ 0.5°), which may cause large uncertainty in applications, especially for the region of the TP with such complex terrain (An et al., 2018). Additionally, these datasets, developed at global or national scales, have not been comprehensively validated for the TP and usually have large uncertainty associated with the methods or land surface models used during their production. To address the spatial resolution issue for the TP, Ding et al. (2018a) developed a downscaling framework using DEM to produce high resolution (0.01°×0.01°) SAT dataset of the TP (Ding et al., 2018b).

Meanwhile, remotely sensed land surface temperature (LST) has been widely used to study regional and global climate change due to its strong correlation with SAT and its global coverage from multiple satellite missions (Good et al., 2017; Pepin et al., 2016). Using monthly LST data of Moderate Resolution Imaging Spectroradiometer (MODIS), Qin et al. (2009) reported that the warming rate of the TP has shown notable dependency on the elevation during 2000-2006. Despite the strong correlation between LST and SAT, they are two distinct variables with different physical definitions. To overcome this limitation, researchers have developed various methods to estimate SAT using LST of various sensors, such as, MODIS (Huang et al., 2017; Lu et al., 2018; Zhang et al., 2016), Spinning Enhanced Visible and Infrared Imager (SEVIRI) (Good, 2015), and Advanced Very High Resolution Radiometer (AVHRR) (Prince et al., 1998). Most of these methods are based on linear regression using LST and other auxiliary input, such as, land cover, surface roughness, day length, and evapotranspiration (Good, 2015; Huang et al., 2017; Meyer et al., 2016; Noi et al., 2017; Zhang et al., 2016). Recently, studies have also explored different machine learning (ML) models (e.g., support vector machine, artificial neural network, random forest, Cubist regression, etc.) to estimate SAT using LST (Meyer et al., 2016; Noi et al., 2017; Zhang et al., 2016). Generally, the ML models perform better than linear regression models because ML models can better capture the complex relationship between LST and SAT. Besides ML models, spatiotemporal interpolation methods, such as, geographically weighted regression, hierarchical Bayesian model, and kriging regression, have also been used to generate high resolution SAT using LST and other auxiliary inputs (Chen et al., 2014; Li et al., 2018; Lu et al., 2018).

Despite the recent progress, the LST-based SAT estimation still suffers a major limitation caused by the cloud contamination. Since most of current LST data are derived from thermal infrared data, the LST data are unavailable when cloud exists during satellite overpassing time. The cloud contamination has strong impacts on the availability and the quality of the SAT estimated using existing methods. Noi et al. (2017) reported that using four instantaneous MODIS LSTs (i.e., daytime and nighttime LSTs of both Terra MODIS and Aqua MODIS products) can accurately estimate daily average SAT with the root-mean-squared-error (RMSE) less than 2 K. However, the RMSE of the estimated SAT increases (larger than 3 K) when cloud contamination occurs (Noi et al., 2017). To account for cloud contamination, Zhang et al. (2016)'s framework estimate the daily average SAT by dynamically integrating available MODIS LSTs based on their quality. Although Zhang el al. (2016)'s method can increase the availability of the estimated SAT, it still requires at least one high quality clear sky LST and the estimated SAT has different levels of uncertainty due to the changing availability and quality of MODIS LSTs (ranging from 1.5 to 3.5 K) (Zhang et al., 2016). To address the cloud contamination issue, Zhu et al. (2017) developed a parameterization scheme to estimate all sky instantaneous daytime SAT using MODIS atmospheric profile products (i.e., MOD06_L2, MOD07_L2). The parameterization scheme is developed based on clear sky near surface air temperature, surface pressure, and land surface temperature average from MODIS products (Zhu et al., 2017). For cloudy sky conditions, Zhu et al. (2017) extends the linear regression between clear sky SAT and LST to the cloud sky MODIS data. However, the relationship between SAT and LST can be quite different between clear sky and cloudy sky conditions, hence leading to large data uncertainty for cloudy sky condition using Zhu et al. (2017)'s method. Zhang et al. (2008) reported that the annual average cloud coverage of the TP ranges from 40% - 60% during 1971-2004. The frequent cloud contamination can have serious implications on the quality and availability of the estimated SAT using existing methods.

The main objective of this study is to develop a method that can produce daily average SAT of the TP with relatively high resolution (i.e., 0.05°×0.05°) that are not or less prone to frequent cloud contamination. Different from existing studies, we propose a ML model to estimate daily average SAT using all available LSTs and remotely sensed radiation variables at both the surface and top-of-atmosphere (TOA) levels. These radiation variables are available for both clear sky and cloudy sky conditions and contain important information about surface energy exchange. Theoretically, the surface energy exchange regulates SAT and its difference with LST. Thus, including these radiation variables may help capturing the physical process of surface heat exchange thus improving the model performance. In this study, we choose the rule-based Cubist regressing (hereafter referred as Cubist) as our base model since previous studies all reported that the Cubist has the best performance on estimating SAT using LSTs over different regions including the TP (Noi et al., 2017; Zhang et al., 2016). To robustly estimate the all sky SAT, we also compare two different strategies using 1) one generic model for both clear sky and cloudy sky conditions or 2) two separate models for clear sky and cloudy sky conditions separately. To the best of our knowledge, this study is the first study using machine learning models to estimate daily average SAT under all sky condition with remotely sensed products. The estimated all sky SAT dataset can be very important for climate analysis and relevant impact studies for the TP. The structure of this manuscript is organized as follow: section 2 describes the data and necessary data processing used in this study; section 3 summarizes the overall research method, Cubist regression model, and the evaluation strategies of this study; model training and validation results are reported in section 4, while section 5 describe the results of cross comparison with existing datasets; section 6 discusses the advantages and limitations of this study while the conclusion is presented in section 7.

2. Data

The data used in this study include 1) the station measured SAT for model training and evaluation, 2) the SAT of various reanalysis/forcing datasets for cross comparison, and 3) the remotely sensed variables as the model inputs (i.e., elevation, LST, surface variables and radiation variables). Table 1 presents the basic summary of the data used in this study. Each category of the data (i.e., station data, remotely sensed data, and reanalysis/forcing data) is further described in the corresponding subsections with the details that are meaningful to this study. If readers are interested in more details, they should refer to the relevant references listed in Table 1.

---------- Insert Table 1 here ----------

2.1 Station data

In this study, the station measured SAT data are used for both model training and evaluation (Table 1(a)). The main source of the daily average SAT used in this study is 135 meteorological stations of the TP managed by the Chinese Meteorological Administration (CMA). The data between 2002 and 2015 were downloaded from the CMA's National Meteorological Information Center (NMIC) (http://data.cma.cn). Additionally, we also collected daily average SAT of 10 individual experiment stations managed by different research groups of the Institute of Tibetan Plateau Research (ITP) to independently evaluate the Cubist model performance. Different from CMA stations, ITP stations have various length of data records since most of these stations are not operational meteorological stations. Moreover, three out of 10 ITP stations are located in regions with elevation above 5,000 meters ASL, which are used to evaluate Cubist model performance over high elevation regions. The location of the CMA stations and ITP stations is presented in Figure 1(a).

2.2 Remotely sensed data

The remotely sensed data used in this study are listed in Table 1(b). The Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) was downloaded from the United States Geological Survey (USGS, https://topotools.cr.usgs.gov/gmted_viewer/). It is produced by combining multiple high quality DEM datasets from various international institutions. The

GMTED2010 data, with an original resolution of 7.5 arc-seconds, were resampled to 0.05°×0.05° by simple averaging to match with the resolution of other remotely sensed data.

In this study, we use MODIS daily composite LST data in a 0.05°×0.05° grid (i.e., MOD11C1 and MYD11C1), which were downloaded from NASA Land Process Distributed Active Archive Center (i.e., LP DAAC, https://lpdaac.usgs.gov/) (Wan et al., 2015a, b). These products are generated by aggregating MODIS Level 2 LST products (i.e., MOD11_L2 and MYD11_L2) with strict quality control. Each product (MOD11C1 and MYD11C1) contains both daytime and nighttime LSTs from different satellite viewing time. Previous studies have proven that combining all four LST values can improve the accuracy of the estimated SAT (Noi et al., 2017; Zhang et al., 2016).

In this study, we also use three remotely sensed radiation products, including Global LAnd Surface Satellite (GLASS) incident solar radiation (ISR) at the surface, University of Maryland's (UMD) TOA outgoing longwave radiation (OLR), and Beijing Normal University's (BNU) TOA albedo (TOAALB). The GLASS ISR data are derived from multiple satellites' data, including AVHRR, MODIS and available geostationary satellites' data (Zhang et al., 2014). The OLR data are produced using AVHRR and MODIS thermal infrared data based on linear regression models derived from radiative transfer model (RTM) simulations (Zhou et al., submitted). The TOAALB data are also produced using AVHRR and MODIS data with linear models derived from RTM simulations (Song et al., 2018). All radiation products are daily data with the same spatial resolution of 0.05°×0.05° for all sky conditions.

The surface variables used in this study include MODIS surface albedo (SFCALB), Normalized Difference Vegetation Index (NDVI) and Normalized Difference Snow Index

(NDSI). The MODIS surface albedo product (MCD43C1) provides daily surface albedo in a 0.05°×0.05°5grid (Schaaf and Wang, 2015). Although MCD43C1 is a daily product, it estimates daily albedo using MODIS data within a 16-day moving window. Therefore, it might not reflect the real surface information of a specific day especially over regions with rapid surface dynamics. The MODIS NDVI data include MOD13C1 and MYD13C1, which are aggregated 16-day products in the same 0.05°×0.05° grid derived using Terra and Aqua MODIS data respectively (Didan, 2015a, b). Both MCD43C1 and MOD13C1/MYD13C1 data were downloaded from NASA LP DAAC. Furthermore, the daily MODIS NDSI data (i.e., MOD10C1 and MYD10C1) were acquired from National Snow and Ice Data Center (NSIDC, https://nsidc.org/) (Hall and Riggs, 2015a, b) in the same 0.05°×0.05° grid.

Since LST has strong correlation with SAT, we use the all available LSTs (i.e., four instantaneous MODIS) to better capture the diurnal cycle of the surface temperature. Because the difference between LST and SAT is related with surface heat exchange, we propose to include radiation variables (i.e., ISR, OLR, and TOAALB) to reflect the crucial process that may improve the accuracy of estimated SAT. Since surface conditions can also affect the difference between LST and SAT, the surface variables are also used as candidate inputs for the model. However, the radiation variables are available for all sky conditions, while the surface variables are only available for clear sky conditions. Thus, including radiation variables would likely increase the data availability of the estimated SAT.

2.3 Reanalysis/forcing data

In this study, the SAT data of three reanalysis/meteorological forcing datasets (see Table 1(a)) are also used to assess the performance of the Cubist model estimated SAT of the TP.

NASA GLDAS produces reanalysis datasets regularly using multiple land surface models at different spatial resolutions (Rodell et al., 2004). We downloaded the SAT data from the GLDAS NOAH reanalysis dataset via NASA's Goddard Earth Sciences Data and Information Services Center (GES DISC). The spatial and temporal resolution of this dataset is 0.25°×0.25° and 3-hourly respectively with the complete coverage over the global land area except the Antarctic since 2000. In addition, we also downloaded the SAT data of the CMA's CLDAS reanalysis data. The CLDAS data are produced hourly in a 0.0625°×0.0625° grid since 2008 (Shi et al., 2011; Xie et al., 2011). Lastly, the CMFD SAT data were downloaded from the ITP's Third Pole Environment Database (TPE). The CMFD dataset contains 3-hourly SAT in a 0.1°×0.1° grid from 1979 to 2016, which is generated by dynamically adjusting the bias of GLDAS reanalysis SAT data to match CMA station observations via spline interpolation (Chen et al., 2011; Yang et al., 2010).

2.4 Data processing

As mentioned earlier, the GMTED2010 elevation data were aggregated to the 0.05°×0.05° grid which is the same with all other remotely sensed datasets. To train and evaluate the model, we extracted all remotely sensed data for all CMA and ITP stations aforementioned (Table 1(a)) via nearest neighborhood method. These extracted remotely sensed data were then paired with the corresponding station SAT data. The station-satellite data pairs were labeled as either clear sky or cloudy sky observations based on the quality flags of the satellite data. Considering different satellite data may have different cloud masks in their quality flags, we only marked the data as clear sky observations when all satellite data' quality flags were cloud free; otherwise, the data were labeled as cloudy observations.

Since MODIS LSTs are only available under the cloud free condition, the missing values were replaced using a temporal moving window (i.e., ± five days) method. When there is at least one clear sky LST within this 11-day time period, the clear sky LST value which is temporally closest to the target date is used to replace the missing value. The purpose of this step is not to accurately predict LST under the cloudy conditions, but rather to provide a first guess of LST that can be used by the Cubist model to estimate SAT. This step could be replaced by more complex spatio-temporal gap filling of LST data, but it is out of the scope of this study. The sensitivity of the data availability and model performance on the moving window size will be discussed later.

To generate daily NDVI for each grid, 16-day MODIS NDVI data (i.e., MOD13C1 and MYD13C1) were firstly merged into one NDVI time series with corresponding date information for each grid. The merged NDVI time series were then filtered using Savitzkey-Golay method to further remove possible cloud contamination (Chen et al., 2004). The filtering process is to remove the suspiciously low NDVI values caused by unfiltered cloud to increase the confidence of clear sky observations. Finally, the filtered time series were interpolated into daily time series based on the double sigmoid model. For all reanalysis and meteorological forcing datasets, we aggregated their sub-daily SAT values to daily average SAT by averaging all estimations within the same day for each grid but leave them as their native spatial resolutions.

3. Method

The overall design of this study is presented in Figure 2. First, all station-satellite data pairs were extracted and processed as described in section 2 for model training and evaluation. Only part of the CMA station-satellite data (2004-2013) were used for model training, while the rest of the CMA station-satellite data and the ITP station-satellite data were kept for independent model evaluation. The Cubist model was trained using the leave-one-station-out (LOSO) strategy to determine the parameters of the final model. We use LOSO to reduce the risk of overfitting by mimicking the process of estimating SAT for unknown regions with no station data (Meyer et al., 2016). After the model parameters were determined using LOSO, we compared two different modeling strategies to estimate daily average SAT under all sky conditions (i.e., a universal all sky model *v.s.* two separate models for clear and cloudy sky separately). The model strategy with the best accuracy evaluated using station data was chosen as the final model. Lastly, the final model was evaluated by comparing with independent station data, the 25-fold cross validation, and cross comparing with the reanalysis/forcing data. The basis of the Cubist model and the model training/evaluation methods are further descripted in the following subsections.

---------- Insert Figure 2. here ----------

3.1 The Cubist model

The Cubist model is a rule-based regression method developed by (Quinlan, 1993a, 1993b, 1992). The Cubist model does not give one final model like other machine learning methods, but it generates a set of rules and multi-variate predictive models associated with the rules based on the independent variables used. Once the rules and rule-associated models are determined, a specific set of independent variables will correspond to predictive models based on rules that best suits this set of independent variables. It is originally developed as a commercial software with limited documentation comparing to other popular machine learning methods. It has been adapted by researchers using open source statistical language R and become a popular model in different disciplines (Kuhn et al., 2018; Kuhn and Johnson, 2013). Despite the lack of documentation, the Cubist model is summarized herein based on existing research and documentations.

The Cubist model originates from the M5 model tree(Quinlan, 1992), which is an improved version of the simple decision tree model. M5 firstly develops a full tree by recursively partitioning all samples into different nodes, which is called tree growing process. After the tree is produced, a multiple linear regression model is fitted for each node. However, some nodes might have large model errors due to insufficient samples. Then, the M5 creates a smaller set of generalized regression models considering the training error, the numbers of sample and the goodness-of-fit for each node, which is called tree pruning process. After the pruning process, the M5 creates a set of rules and corresponding multi-variate linear regression models (Figure 3). Specifically, a rule is constructed using one or multiple input variables (i.e., variables listed in Table 1(b)). For instance, a rule may be set as “elevation > 3500 meters and OLR < 87 W/m^2”, which creates a subset of data with the elevation higher than 3500 meters ASL and OLR lower than 87 W/m^2. Although the original tree partitioning is recursive, final rule sets after pruning may be overlapping, which means a sample may be assigned into multiple subsets based on different rules. In these cases, predictions from different subsets will be averaged to generate the merged prediction.

---------- Insert Figure 3 here ----------

Next, the Cubist model utilize a boosting-like technique to enhance its prediction performance with a process named committee prediction (Kuhn and Johnson, 2013). Since the M5 model tree does not produce perfect prediction, the Cubist model uses the model error of the initial model to adjust the original dependent variable (i.e., SAT in this study) and creates a new

M5 model tree using same inputs and processes. This process repeats multiple times which is predefined to create the final model and each individual M5 model is considered a committee (Figure 4). The final prediction of the Cubist model is calculated by averaging corresponding predictions of each committee.

---------- Insert Figure 4 here ----------

3.2 Model training and validation

To build and evaluate the proposed model, the station-satellite data pairs were separated into two sets: 1) the training set and 2) the validation set. The training set contains CMA station-satellite data pairs from 2004 to 2013, while the validation set includes all ITP station-satellite data pairs and the ones of CMA of 2002, 2003, 2014, and 2015. As mentioned earlier, we compare two different strategies to estimate SAT of all sky conditions (strategy I: a universal all sky model; strategy II: two models for clear/cloudy sky conditions separately). Table 2 lists all candidate models for both strategies using different combinations of variables listed in Table 1(b), including, elevation, LSTs, radiation variables, and surface variables. Since surface variables are only available for clear sky conditions, only Strategy II's clear sky models include them (i.e., NDVI, NDSI, and SFCALB) as model inputs. In all models listed in Table 2, we use all four instantaneous MODSI LST (or gap-filled LST) within the same day as part of the model input. This is motivated by the assumption that using four instantaneous LST values may better capture the diurnal cycle of surface temperature change (Zhang et al., 2016).

---------- Insert Table 2 here ----------

In the Cubist model, two parameters need to be determined through training, i.e., the number of committees and the number of neighbors. During the training process, we used the

LOSO strategy to select model parameters to avoid the overfitting issue as mentioned earlier (Meyer et al., 2016). Firstly, the training data were grouped by stations. For each iteration, a series of Cubist models were fitted using different combinations of model parameters using data of all stations except one which was randomly chosen. Then, the models were evaluated using the data of the left-out station. After each station has been used as the left-out station to evaluate different model parameters, the final model parameters were selected based on the model performance across all iterations.

To evaluate the final Cubist model, we first used the validation dataset of CMA stations of the year 2002, 2003, 2014 and 2015 to assess the model performance when it is applied to data of different years. Additionally, the model was also evaluated using independent data of 10 ITP stations. Furthermore, we carried out a 25-fold cross validation experiment to examine the robustness of our model. In this experiment, all CMA station-satellite data pairs were randomly divided into 25 folds by station ID. During each iteration, a Cubist model was fitted using 24 folds of data with the same parameters previously determined by the LOSO. This model was then evaluated using the left-out fold of data. This process was repeated 25 times until all 25 folds of data have been used to independently evaluate a Cubist model. This cross validation process is used to examine the sensitivity of the Cubist model on the training datasets.

Lastly, we applied the final Cubist model to the entire TP for the year of 2014. The estimated SAT of the TP was cross compared with three reanalysis/forcing datasets listed in Table 1(a). The main purpose of the cross comparison is to evaluate the spatial and temporal (i.e., seasonal) pattern of the estimated SAT. Additionally, we also compared the accuracy of our Cubist estimation and the existing datasets using CMA station data as the reference since all datasets have their own uncertainty.

4. Evaluation of Cubist Model Performance

4.1 Model strategy comparison

The statistics of all candidate models listed in Table 2 are presented in Table 3. For clear sky models, the full model (i.e., CLR-5) has the best performance with the lowest RMSE and the highest R^2. The clear sky model without surface variables (i.e., CLR-4) also achieves comparable performance with the full model (CLR-5). However, when LSTs are replaced by TOA radiations (i.e., CLR-0 *v.s.* CLR-1, CLR-2 *v.s.* CLR-3), the performance of the models without LSTs are worse than the models with LSTs. The clear sky models indicate that LSTs have strong impacts on the Cubist model performance while radiation variables can be good supplemental variables to improve the model performance. Figure 5 demonstrates the density scatter plots of all Cubist models for the estimated daily average SATs against the CMA station measurements.

-------- Insert Table 3 here --------

For cloudy sky models, the model with the temporally gap-filled LSTs and radiation variables (i.e., CLD-4) has the best performance with lowest RMSE. However, the model performance deteriorates when either gap-filled LSTs or radiation variables are dropped out from the model (i.e., CLD-0, CLD-1). Nonetheless, the model with only gap-filled LSTs (CLD-0) still outperforms the model with only radiation variables (CLD-1), which is similar with clear sky models. Moreover, the best cloudy sky model (CLD-4) performs slightly worse than the corresponding clear sky model (CLR-4) which may be caused by the uncertainty of gap-filled LSTs.

-------- Insert Figure 5 here --------

The best all sky model is the one with both gap-filled LSTs and radiation variables (ALL-4), followed by the model with only gap-filled LSTs (ALL-0) and the model with only radiation variables (ALL-1). Moreover, the best all sky model (ALL-4) has better overall performance than the best cloudy-sky model (CLD-4) but underperforms the best clear sky model (CLR-5). However, the all sky model, ALL-4, can estimate daily average SAT for much more observations instead of the clear sky only model (CLR-5) (i.e., number of data points: 371,395 *v.s.* 102,457) with comparable overall accuracy. In practice, this advantage can largely increase the data availability without notably sacrificing the data quality.

Table 4 summarizes the validation results for all 12 Cubist models using temporally independent CMA station data (of the year 2002, 2003, 2014, and 2015). For all sky models, we further calculated the statistics for clear sky and cloudy sky observations separately to directly compare with the results of clear/cloudy sky models. For clear sky condition, all models with LSTs as inputs show comparable performance with each other but outperforms the models without LSTs. This further confirms that LSTs have major contribution to accurately estimate daily average SAT. The best model is still the full model with all variables as inputs (CLR-5). For cloudy and all sky models, the models with both gap-filled LSTs and radiation variables are the best model of its category (i.e., CLD-4 and ALL-4). Figure 6 shows the density scatter plots of the best models within each category (i.e., CLR-5, CLD-4, and ALL-4). For the model ALL-4, the density scatter plots of clear sky and cloudy sky observations are also presented separately to directly compare with the results of CLD-4 and CLR-5 (Figure 6 (c-d)).

-------- Insert Table 4 here --------

For all sky models, when the samples are separated into clear/cloudy conditions, the estimation of all sky models can achieve similar or even better accuracy with the estimation of the corresponding clear/cloudy sky models. For example, the statistics of ALL-4 for clear sky and cloudy sky observations are similar with the statistics of the corresponding clear sky model (CLR-4; RMSE: 1.634 ℃5v.s. 1.643 ℃; Bias: -0.069 ℃5v.s. -0.116 ℃; R^2: 0.967 v.s. 0.967) and cloudy sky model (CLD-4; RMSE: 1.922 ℃5v.s. 1.917 ℃; Bias: -0.058 ℃5v.s. -0.025 ℃; R^2: 0.954 v.s. 0.955). In general, the all sky model with gap-filled LSTs and radiation variables (i.e., ALL-4) shows the best overall performance with satisfactory accuracy and the capability of overcoming cloud contamination issue. Therefore, we only evaluate the ALL-4 model in the remaining part of this study.

-------- Insert Figure 6 here --------

4.2 Independent validation with ITP station data

To further validate the all sky model independently, the data of 10 ITP stations of varying time periods were used in this study. Out of these 10 stations, three of them are located at elevation higher than 5,000 ASL. Figure 7 presents the validation results using these ITP station data. The estimated daily average SAT show good agreement with the station measurements with nearly zero bias. However, the RMSE is slightly larger than the ones of model training and validation results using the CMA station data (RMSE: 2.18°C *v.s.* 1.84°C). Furthermore, the accuracy of estimated SAT for stations with elevation higher than 5,000 meters ASL is slightly worse than other ITP stations (RMSE: 2.29°C *v.s.* 2.05°C). This result is possible considering that the training data do not contain any stations above 5,000 meters ASL. This lack of representation could increase the uncertainty of the estimated SATs over high elevation regions (Zhang et al., 2016).

---------- Insert Figure 7 here ----------

4.3 Model sensitivity analysis

Figure 8 shows the RMSE and R^2 of the final LOSO result of each individual CMA stations during the model training process. In general, the final Cubist model performs well for most stations with RMSE lower than 2 °C and R^2 higher than 0.95 (Figure 8 (a, d)). However, there are some stations with relatively large uncertainty. These stations appear to be located at the regions with relatively complex terrain. Figure 8(b) shows that stations above 4,000 meters ASL may show larger RMSE during this LOSO analysis. However, Figure 8(c) demonstrates that the Cubist model estimation of nearly 80% of the CMA stations has RMSE less than 2.1°C. This result shows comparable or slightly better performance than previous studies' clear sky only models. Overall, the LOSO result suggests that the Cubist model trained with limited amount of station data may be applied to other regions of the Plateau with acceptable accuracy for all sky conditions. During the LOSO analysis, the results of CMA stations located at complex terrain show relatively large RMSE (Figure 8a). This is likely caused by the spatial scale difference between satellite data (0.05°×0.05°) and station point measurements. The microclimate regime of complex terrain could increase the difference between station SAT (point measurement) and estimated SAT (area average estimation).

---------- Insert Figure 8 here ----------

Similar with the LOSO analysis, the 25-fold cross validation experiment was also designed to show the robustness of the proposed Cubist model. Figure 9 exhibits the RMSE of all 25 models trained with slightly different set of training data. In general, most models in the 25-fold cross validation experiment has low RMSE (1.6°C – 2.2°C). However, some models (i.e., model No. 2, 9, 19) show relatively large RMSEs (>2.5°C). The potential cause of the poor performance for these models is described in the discussion section.

---------- Insert Figure 9 here ----------

5. Cross comparison with reanalysis/forcing data

In addition to evaluate the proposed model with station measurements, we also compared the Cubist estimation of the TP with three reanalysis/meteorological forcing datasets listed in Table 1(a). Figure 10 compares the spatial pattern of the monthly mean SAT of our Cubist model estimation (Figure 10 (a-d)) with GLDAS (Figure 10 (e-h)), CLDAS (Figure 10 (i-l)), and CMFD (Figure 10 (m-p)) for January, April, July, and October 2014. Overall, all four datasets show very similar spatial and temporal SAT gradients across the entire TP. Generally, the SAT is higher at the regions with low altitudes (i.e., the northern and southeast parts of the TP) while the high elevation regions (e.g., the western and central areas of the TP) have lower temperature. Additionally, all datasets show the same seasonal SAT dynamics.

Despite the consistency, there are still notable differences among these datasets. Even though GLDAS may capture the overall spatial pattern of the SAT, it does not have the same level of spatial details as the Cubist estimation because GLDAS's resolution is very coarse (0.25° *v.s.* 0.05°). The lack of the spatial details can be troublesome because parts of the TP have very complex terrain. It is not suitable to use such coarse resolution to represent the climate and ecosystem processes of those regions. Additionally, the CLDAS SAT appears to have larger spatial gradients , especially for April and July 2014.

Considering all these datasets have their own uncertainties, we use validation years' CMA station measurements as a reference to compare the accuracy of these four datasets. Both CMFD and GLDAS show substantial underestimation (Bias: -2.47°C *v.s.* -3.11°C) when compared to the reference CMA station data. The CLDAS data notably overestimates the surface temperature with a bias of 1.07°C while the Cubist model estimation shows nearly zero bias (-0.07°C). Furthermore, the Cubist model estimation shows smaller uncertainty than other three datasets (RMSE: 1.84°C (Cubist) *v.s.* 4.82°C (GLDAS) *v.s.* 4.20 (CMFD) *v.s.* 3.31°C (CLDAS)). In summary, the Cubist model estimated SAT of the TP has better accuracy than existing reanalysis and forcing datasets and can capture the SAT's spatial and temporal dynamics of the TP.

-------- Insert Figure 10 here --------

6. Discussions

Overall, in the proposed all sky Cubist model, four MODIS LSTs are used to capture the spatial pattern and diurnal cycle of surface temperature while radiation variables are useful to characterize the energy exchange which may regulate the difference between LST and SAT. Additionally, the radiation variables can also be good indicators of the presence of cloud since their values may be affected by cloud coverage. Therefore, the Cubist model may create rules based on these radiation variables to automatically separate data into different subsets with corresponding models. This may be one of the reasons that the all sky model can get similar or even better results than separate clear/cloudy sky models. By introducing the radiation variables into the Cubist model, our proposed all sky model can estimate daily average SAT of the entire TP for all sky conditions with satisfactory accuracy.

In the Cubist model, we used the temporally gap-filled MODIS LSTs, within an 11-day moving window, to replace the missing value of LSTs caused by cloud. This process improved the data availability by more than three times (see Table 3 & 4). However, the impact of the temporal window size on the performance of the Cubist model estimation needs to be examined. Table 5 presents the model performance and the data availability for both model training and validation against different window sizes (ranging from 0 to ±5 days). The model inputs are fixed as LSTs, radiation variables, geolocations, elevations, day of year (i.e., ALL-4). The model using temporal window size of 0 day is equivalent to the clear sky model CLR-4. The increasing temporal window size significantly increases the data availability, while the RMSE only increases slightly (~0.2°C) with the increasing moving window size. The temporal moving window of ±5 days yields almost 100% data availability for these CMA stations with the satisfactory accuracy and precision, which notably outperforms existing forcing data. However, we are aware that this sensitivity test was only conducted for the limited regions with CMA stations. The results may not be the same for the entire TP. It is very challenging to repeat the same analysis for the entire TP because there are no true SAT data of the entire TP to help us understand the impact of the moving window size on the performance of our model. In the future, this moving window process can be replaced with other more advanced spatio-temporal interpolation methods or substituted with microwave based all sky LST products to provide the better first guess of LSTs for cloudy sky conditions (e.g., Wang et al., 2019; Zhang et al., 2019; Zhou et al., 2017).

In addition, we define clear sky condition very conservatively – all four instantaneous MODIS LST and daily satellite radiation data need to be cloud free. This is because the purpose of this study is to estimate daily average SAT. This may lead to larger variations of the model performance for cloudy sky conditions. As concluded by Zhang et al. (2016), the number of high-quality cloud free instantaneous MODIS LST within one day can have notable impact on the final model uncertainty.

-------- Insert Table 5 here --------

During the 25-fold cross validation, some Cubist models trained using a random subset of the CMA station data show large uncertainty when they were applied to independent station data. To better understand potential error sources, we examined four Cubist models with largest RMSEs in the 25-fold cross validation experiment (i.e., model 2, 9, 19, 25). Figure 11 presents the density scatter plots of validation results, the comparison of data distributions of the training and validation data, and the quantile-quantile (Q-Q) plots between the training and validation datasets. For most of these cases (except model 2), the data distributions of training and validation data are notably different. For example, Fold-19 and Fold-25 all show that the distributions of their validation data are shifted rightwards from the distributions of their training data; Fold-9 exhibits a double-peak distribution of its validation data which is different from the near normal distribution of it training data. The Q-Q plots also confirm these differences among training/validation data's distribution. This common characteristic of these three cases underpins the assumption of machine learning models. Machine learning models are designed to predict unknown situations by learning from existing data/observations. The underlying assumption of most machine learning models is that the training data should represent the overall data distribution reasonably well. If this assumption is invalid, like in our case, the performance of the prediction/estimation can be notably affected. Therefore, it is very important to ensure the representativeness of the data during model training process. Nevertheless, when we examine the distributions of the training data for all four cases, it is quite assuring to see that they share almost the same distribution despite the difference of their training data. This implies that when the amount of training data is large enough the sample distribution may be very close to the real data distribution. However, it is always the best practice to examine and increase, if possible, the representativeness of the training data to ensure the trained machine learning model is not biased from the beginning.

---------- Insert Figure 11 here ----------

7. Conclusion

This study demonstrates that combining LSTs and radiation variables at both the surface and TOA levels can produce daily average SAT data under all sky conditions with high accuracy. With a reasonably defined temporal moving window to fill the gap of missing LST caused by cloud contamination, the all sky Cubist model can largely increase the data availability of estimated daily average SAT over the Tibetan Plateau. The model has been validated using spatially and temporally independent station data and cross validation with nearly zero bias and reasonable RMSEs (1.8-2.2 °C). When cross compared with the existing reanalysis/forcing datasets, the Cubist model estimated SAT can represent the spatial and temporal dynamics of the surface temperature of the TP and retain important spatial details. When all datasets were benchmarked against the CMA station data, the Cubist model show better performance with no notable bias and much smaller RMSEs. However, the 25-fold cross-validation practice suggests that the representativeness of the training dataset is of great importance to produce a high-quality machine learning model with no built-in bias.

With the all sky Cubist model, we generated a $0.05° \times 0.05°$ daily average surface air temperature dataset for the entire Tibetan Plateau for 2002-2016. The resulting dataset is of great value to study recent climate warming and corresponding impacts over the entire Tibetan Plateau. However, as mentioned earlier, the training data are only from the finite CMA stations. Therefore, users should be aware of the potential larger uncertainty for the regions of which the weather/climate patterns might not be represented by the CMA station data, such as, the regions with very high elevations (e.g., above 6,000 meters ASL) or with complex topography. Although the model is developed for the Tibetan Plateau, the framework of this model could be extended to other regions since the underlining mechanism should be similar. However, when the framework is applied to other areas, the model should be properly retrained using the data of the target area to ensure that the model are built correctly with representative training data.

Despite the improved accuracy and data availability of the daily average SAT dataset, there are still uncertainties require further investigations to improve the resulted data. First, all input satellite data have different level of uncertainties which can be propagated into this empirical-based estimation. Therefore, it will be beneficial to understand the sensitivity of the estimated SAT regarding to the uncertainty of each individual input variables. Secondly, due to the complex terrain of the TP, there are grids with large elevation gradients. It can be very challenging to assess the accuracy of these regions. Although we have conducted independent validation using 10 non-CMA stations (including three stations above 5000 meters), it is still necessary to use more independent data of the regions with very high elevations and complex topography, if available, to comprehensively evaluate the quality of the estimated SAT data. The independent evaluation would further provide better characterization of the data quality for the Tibetan Plateau, which is essential for users who need this dataset. Moreover, there are growing demands of grid-level uncertainty assessment to improve the confidence of local and regional applications using various climate datasets. Thus, it is of the best interest to provide the uncertainty value associated with each grid for the estimated SAT data using advanced statistical methods, such as, Markov Chain Monte Carlo (MCMC), bootstrapping etc. The estimation uncertainty should account for both input data uncertainty and model uncertainty. The measure of estimation uncertainty will allow users to better understand the strength and weakness of the estimation for their own applications. Lastly, we are planning to extend this model using AVHRR data to generate long term SAT climate data records of the TP (since 1982) to enable climate applications for the last four decades.

Acknowledgements

Yuhan Rao was supported by the China Scholarship Council. This work was also partially supported by the NOAA grant NA14NES4320003 (Cooperative Institute for Climate and Satellites‐CICS) at the University of Maryland/ESSIC. The meteorological forcing dataset used in this study was developed by Data Assimilation and Modeling Center for Tibetan Multi-spheres spheres, Institute of Tibetan Plateau Research, Chinese Academy of Sciences. The Cubist regression model is implemented using “cubist” R package developed by Max Kuhn. We would like to acknowledge Ying Kong for assisting processing CMA station data of the Tibetan Plateau, Lirong Ding and Xiaodong Zhang for assisting the access of CLDAS data, Drs. Liping Zhu, Junbo Wang, Yaoming Ma, Weiqiang Ma, and Ying Xie for providing station measurements managed by ITP. We also want to acknowledge NASA and CMA for making their land data assimilation system products available for research purpose.

Reference

An, S., Zhu, X., Shen, M., Wang, Y., Cao, R., Chen, X., Yang, W., Chen, J., Tang, Y., 2018. Mismatch in elevational shifts between satellite observed vegetation greenness and temperature isolines during 2000–2016 on the Tibetan Plateau. Global Change Biology 24, 5411–5425. https://doi.org/10.1111/gcb.14432

Chen, F., Liu, Y., Liu, Q., Qin, F., 2014. A statistical method based on remote sensing for the estimation of air temperature in China. International Journal of Climatology 35, 2131–2143. https://doi.org/10.1002/joc.4113

Chen, J., Jönsson, Per., Tamura, M., Gu, Z., Matsushita, B., Eklundh, L., 2004. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. Remote Sensing of Environment 91, 332–344. https://doi.org/10.1016/j.rse.2004.03.014

Chen, Y., Yang, K., He, J., Qin, J., Shi, J., Du, J., He, Q., 2011. Improving land surface temperature modeling for dry land of China. Journal of Geophysical Research: Atmospheres 116. https://doi.org/10.1029/2011JD015921

Cong, N., Shen, M., Piao, S., 2017. Spatial variations in responses of vegetation autumn phenology to climate change on the Tibetan Plateau. J Plant Ecol 10, 744–752. https://doi.org/10.1093/jpe/rtw084

Didan, K., 2015a. MOD13C1 MODIS/Terra Vegetation Indices 16-Day L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MOD13C1.006

Didan, K., 2015b. MYD13C1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MYD13C1.006

Ding, L., Zhou, J., Zhang, X., Liu, S., Cao, R., 2018a. Downscaling of surface air temperature over the Tibetan Plateau based on DEM. International Journal of Applied Earth Observation and Geoinformation 73, 136–147. https://doi.org/10.1016/j.jag.2018.05.017

Ding, L., Zhou, J., Zhang, X., Liu, S., Cao, R., 2018b. A long-term 0.01° surface air temperature dataset of Tibetan Plateau. Data in Brief 20, 748–752. https://doi.org/10.1016/j.dib.2018.08.107

Duan, A., Xiao, Z., 2015. Does the climate warming hiatus exist over the Tibetan Plateau? Scientific Reports 5, 13711. https://doi.org/10.1038/srep13711

Good, E., 2015. Daily minimum and maximum surface air temperatures from geostationary satellite data. J. Geophys. Res. Atmos. 120, 2014JD022438. https://doi.org/10.1002/2014JD022438

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

Hall, D.K., Riggs, G.A., 2015a. MODIS/Terra Snow Cover Daily L3 Global 0.05Deg CMG, Version 6. https://doi.org/10.5067/MODIS/MOD10C1.006

Hall, D.K., Riggs, G.A., 2015b. MODIS/Aqua Snow Cover Daily L3 Global 0.05Deg CMG, Version 6. https://doi.org/10.5067/MODIS/MYD10C1.006

Huang, F., Ma, W., Wang, B., Hu, Z., Ma, Y., Sun, G., Xie, Z., Lin, Y., 2017. Air temperature estimation with MODIS data over the Northern Tibetan Plateau. Adv. Atmos. Sci. 34, 650–662. https://doi.org/10.1007/s00376-016-6152-5

Kuhn, M., Johnson, K., 2013. Regression trees and rule-based models, in: Applied Predictive Modeling. Springer, pp. 173–220.

Kuhn, M., Weston, S., Keefer, C., Coulter, N., Quinlan, J.R., 2018. Cubist: Rule- And Instance-Based Regression Modeling.

Li, X., Cheng, G., Lu, L., 2005. Spatial Analysis of Air Temperature in the Qinghai-Tibet Plateau. Arctic, Antarctic, and Alpine Research 37, 246–252. https://doi.org/10.1657/1523-0430(2005)037[0246:SAOATI]2.0.CO;2

Li, X., Zhou, Y., Asrar, G.R., Zhu, Z., 2018. Developing a 1 km resolution daily air temperature dataset for urban and surrounding areas in the conterminous United States. Remote Sensing of Environment 215, 74–84. https://doi.org/10.1016/j.rse.2018.05.034

Liu, X., Chen, B., 2000. Climatic warming in the Tibetan Plateau during recent decades. International Journal of Climatology 20, 1729–1742. https://doi.org/10.1002/1097-0088(20001130)20:14<1729::AID-JOC556>3.0.CO;2-Y

Liu, X., Cheng, Z., Yan, L., Yin, Z.-Y., 2009. Elevation dependency of recent and future minimum surface air temperature trends in the Tibetan Plateau and its surroundings. Global and Planetary Change 68, 164–174. https://doi.org/10.1016/j.gloplacha.2009.03.017

Lu, N., Liang, S., Huang, G., Qin, J., Yao, L., Wang, D., Yang, K., 2018. Hierarchical Bayesian space-time estimation of monthly maximum and minimum surface air temperature. Remote Sensing of Environment 211, 48–58. https://doi.org/10.1016/j.rse.2018.04.006

Meyer, H., Katurji, M., Appelhans, T., Müller, M.U., Nauss, T., Roudier, P., Zawar-Reza, P., 2016. Mapping Daily Air Temperature for Antarctica Based on MODIS LST. Remote Sensing 8, 732. https://doi.org/10.3390/rs8090732

Noi, P.T., Degener, J., Kappas, M., 2017. Comparison of Multiple Linear Regression, Cubist Regression, and Random Forest Algorithms to Estimate Daily Air Surface Temperature from Dynamic Combinations of MODIS LST Data. Remote Sensing 9, 398. https://doi.org/10.3390/rs9050398

Pepin, N., Bradley, R.S., Diaz, H.F., Baraer, M., Caceres, E.B., Forsythe, N., Fowler, H., Greenwood, G., Hashmi, M.Z., Liu, X.D., Miller, J.R., Ning, L., Ohmura, A., Palazzi, E., Rangwala, I., Schöner, W., Severskiy, I., Shahgedanova, M., Wang, M.B., Williamson, S.N., Yang, D.Q., 2015. Elevation-dependent warming in mountain regions of the world. Nature Climate Change 5, 424.

Pepin, N.C., Maeda, E.E., Williams, R., 2016. Use of remotely sensed land surface temperature as a proxy for air temperatures at high elevations: Findings from a 5000 m elevational transect across Kilimanjaro. J. Geophys. Res. Atmos. 121, 2016JD025497. https://doi.org/10.1002/2016JD025497

Prince, S.D., Goetz, S.J., Dubayah, R.O., Czajkowski, K.P., Thawley, M., 1998. Inference of surface and air temperature, atmospheric precipitable water and vapor pressure deficit using Advanced Very High-Resolution Radiometer satellite observations: comparison with field observations. Journal of Hydrology 212–213, 230–249. https://doi.org/10.1016/S0022-1694(98)00210-8

Qin, J., Yang, K., Liang, S., Guo, X., 2009. The altitudinal dependence of recent rapid warming over the Tibetan Plateau. Climatic Change 97, 321. https://doi.org/10.1007/s10584-009-9733-9

Quinlan, J.R., 1993a. Constructing Decision Trees, in: C4.5: Programs for Machine Learning. Elsevier, pp. 17–26.

Quinlan, J.R., 1993b. Combining instance-based and model-based learning, in: Proceedings of the Tenth International Conference on Machine Learning. pp. 236–243.

Quinlan, J.R., 1992. Learning with continuous classes, in: 5th Australian Joint Conference on Artificial Intelligence. World Scientific, pp. 343–348.

Rangwala, I., Miller, J.R., 2012. Climate change in mountains: a review of elevation-dependent warming and its possible causes. Climatic Change 114, 527–547. https://doi.org/10.1007/s10584-012-0419-3

Rao, Y., Liang, S., Yu, Y., 2018. Land Surface Air Temperature Data Are Considerably Different Among BEST-LAND, CRU-TEM4v, NASA-GISS, and NOAA-NCEI. Journal of Geophysical Research: Atmospheres 123, 5881–5900. https://doi.org/10.1029/2018JD028355

Rodell, M., Houser, P.R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin*, J.K., Walker, J.P., Lohmann, D., Toll, D., 2004. The Global Land Data Assimilation System. Bull. Amer. Meteor. Soc. 85, 381–394. https://doi.org/10.1175/BAMS-85-3-381

Schaaf, C., Wang, Z., 2015. MCD43C1 MODIS/Terra+Aqua BRDF/AlbedoModel Parameters Daily L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MCD43C1.006

Shen, M., Piao, S., Chen, X., An, S., Fu, Y.H., Wang, S., Cong, N., Janssens, I.A., 2016. Strong impacts of daily minimum temperature on the green-up date and summer greenness of the

Tibetan Plateau. Global Change Biology 22, 3057–3066. https://doi.org/10.1111/gcb.13301

Shen, M., Piao, S., Dorji, T., Liu, Q., Cong, N., Chen, X., An, S., Wang, S., Wang, T., Zhang, G., 2015. Plant phenological responses to climate change on the Tibetan Plateau: research status and challenges. Natl Sci Rev 2, 454–467. https://doi.org/10.1093/nsr/nwv058

Shen, S.S.P., Yao, R., Ngo, J., Basist, A.M., Thomas, N., Yao, T., 2015. Characteristics of the Tibetan Plateau snow cover variations based on daily data during 1997–2011. Theor Appl Climatol 120, 445–453. https://doi.org/10.1007/s00704-014-1185-0

Shi, C., Xie, Z., Qian, H., Liang, M., Yang, X., 2011. China land soil moisture EnKF data assimilation based on satellite remote sensing data. Sci. China Earth Sci. 54, 1430–1440. https://doi.org/10.1007/s11430-010-4160-3

Song, Z., Liang, S., Wang, D., Zhou, Y., Jia, A., 2018. Long-term record of top-of-atmosphere albedo over land generated from AVHRR data. Remote Sensing of Environment 211, 71–88. https://doi.org/10.1016/j.rse.2018.03.044

Wan, Z., Hook, S., Hulley, G., 2015a. MOD11C1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MOD11C1.006

Wan, Z., Hook, S., Hulley, G., 2015b. MYD11C1 MODIS/Aqua Land Surface Temperature/Emissivity Daily L3 Global 0.05Deg CMG V006. https://doi.org/10.5067/MODIS/MYD11C1.006

Wang, T., Shi, J., Ma, Y., Husi, L., Comyn□Platt, E., Ji, D., Zhao, T., Xiong, C., 2019. Recovering Land Surface Temperature Under Cloudy Skies Considering the Solar-

Cloud-Satellite Geometry: Application to MODIS and Landsat-8 Data. Journal of Geophysical Research: Atmospheres 0. https://doi.org/10.1029/2018JD028976

Xie, Y., Koch, S., McGinley, J., Albers, S., Bieringer, P.E., Wolfson, M., Chan, M., 2011. A Space–Time Multiscale Analysis System: A Sequential Variational Analysis Approach. Mon. Wea. Rev. 139, 1224–1240. https://doi.org/10.1175/2010MWR3338.1

Yang, K., He, J., Tang, W., Qin, J., Cheng, C.C.K., 2010. On downward shortwave and longwave radiations over high altitude regions: Observation and modeling in the Tibetan Plateau. Agricultural and Forest Meteorology 150, 38–46. https://doi.org/10.1016/j.agrformet.2009.08.004

Yang, K., Wu, H., Qin, J., Lin, C., Tang, W., Chen, Y., 2014. Recent climate changes over the Tibetan Plateau and their impacts on energy and water cycle: A review. Global and Planetary Change 112, 79–91. https://doi.org/10.1016/j.gloplacha.2013.12.001

Yang, Y., Ren, R., 2017. On the contrasting decadal changes of diurnal surface temperature range between the Tibetan Plateau and southeastern China during the 1980s–2000s. Adv. Atmos. Sci. 34, 181–198. https://doi.org/10.1007/s00376-016-6077-z

Yao, T., Thompson, L.G., Mosbrugger, V., Zhang, F., Ma, Y., Luo, T., Xu, B., Yang, X., Joswiak, D.R., Wang, W., Joswiak, M.E., Devkota, L.P., Tayal, S., Jilani, R., Fayziev, R., 2012. Third Pole Environment (TPE). Environmental Development 3, 52–64. https://doi.org/10.1016/j.envdev.2012.04.002

Yao, T., Xue, Y., Chen, D., Chen, Fahu, Thompson, L., Cui, P., Koike, T., Lau, W.K.-M., Lettenmaier, D., Mosbrugger, V., Zhang, R., Xu, B., Dozier, J., Gillespie, T., Gu, Y., Kang, S., Piao, S., Sugimoto, S., Ueno, K., Wang, L., Wang, W., Zhang, F., Sheng, Y., Guo, W., Ailikun, Yang, X., Ma, Y., Shen, S.S.P., Su, Z., Chen, Fei, Liang, S., Liu, Y., Singh, V.P., Yang, K., Yang, D., Zhao, X., Qian, Y., Zhang, Y., Li, Q., 2018. Recent Third Pole's rapid warming accompanies cryospheric melt and water cycle intensification and interactions between monsoon and environment: multi-disciplinary approach with observation, modeling and analysis. Bull. Amer. Meteor. Soc. https://doi.org/10.1175/BAMS-D-17-0057.1

Zhang, H., Zhang, F., Ye, M., Che, T., Zhang, G., 2016. Estimating daily air temperatures over the Tibetan Plateau by dynamically integrating MODIS LST data. J. Geophys. Res. Atmos. 121, 2016JD025154. https://doi.org/10.1002/2016JD025154

Zhang, X., Liang, S., Zhou, G., Wu, H., Zhao, X., 2014. Generating Global LAnd Surface Satellite incident shortwave radiation and photosynthetically active radiation products from multiple satellite data. Remote Sensing of Environment 152, 318–332. https://doi.org/10.1016/j.rse.2014.07.003

Zhang, X., Peng, L., Zheng, D., Tao, J., 2008. Cloudiness variations over the Qinghai-Tibet Plateau during 1971–2004. J. Geogr. Sci. 18, 142–154. https://doi.org/10.1007/s11442-008-0142-1

Zhang, X., Zhou, J., Göttsche, F., Zhan, W., Liu, S., Cao, R., 2019. A Method Based on Temporal Component Decomposition for Estimating 1-km All-Weather Land Surface Temperature by Merging Satellite Thermal Infrared and Passive Microwave Observations. IEEE Transactions on Geoscience and Remote Sensing 1–22. https://doi.org/10.1109/TGRS.2019.2892417

Zhou, W., Peng, B., Shi, J., 2017. Reconstructing spatial–temporal continuous MODIS land surface temperature using the DINEOF method. JARS 11, 046016. https://doi.org/10.1117/1.JRS.11.046016

Zhu, W., Lű, A., Jia, S., Yan, J., Mahmood, R., 2017. Retrievals of all-weather daytime air temperature from MODIS products. Remote Sensing of Environment 189, 152–163. https://doi.org/10.1016/j.rse.2016.11.011

Zhou, Y., Liang, S., Wang, D., Song, Z., Yu, Y., (submitted). Generating 35-year high-resolution outgoing longwave radiation record from AVHRR data. Remote Sensing of Environment.

Tables

Table 1(a). The summary of observational and reanalysis/forcing surface air temperature data used in this study.

Dataset	Data Type	Resolution (Spatial/temporal)	Data Source	Reference
CMA	Station data	- / Daily	NMIC	-
ITP	Station data	- / Daily	TPE	Yao et al. (2012)
CMFD	Reanalysis data	0.10° / 3-hourly	TPE	Chen et al. (2011)
CLDAS	Reanalysis data	0.0625°/ hourly	NMIC	Shi et al. (2011)
GLDAS	Reanalysis data	0.25° / 3-hourly	NASA GES DISC	Rodell et al. (2004)

Table 1(b). The summary of the remotely sensed data used in this study.

Variable	Dataset(s)	Variable Category	Resolution (Spatial/temporal)	Data Source	References
Elevation	GMTED2010	Geo-location	7. ″ / Static	USGS	Danielson & Gesch (2011)
Land surface temperature (LST)	MOD11C1, MYD11C1	Clear sky only	0.05°/ Daily	NASA LP DAAC	Wan et al. (2015a, b)
Incident solar radiation (ISR)	GLASS05B01	All sky	0.05°/ Daily	UMD	Zhang et al. (2014)
Outgoing longwave radiation (OLR)	AVHOLR	All sky	0.05°/ Daily	UMD	Zhou et al. (Submitted)
Top-of-atmosphere albedo (TOAALB)	AVHALB	All sky	0.05°/ Daily	BNU	Song et al. (2018)
Land surface albedo (SFCALB)	MCD43C1	Clear sky only	0.05°/ Daily	NASA LP DAAC	Schaaf & Wang (2015)
Normalized Difference Vegetation Index (NDVI)	MOD13C1, MYD13C1	Clear sky only	0.05°/ 16-day	NASA LP DAAC	Didan (2015a, b)
Normalized Difference Snow Index (NDSI)	MOD10C1, MYD10C1	Clear sky only	0.05°/ Daily	NSIDC	Hall & Riggs (2015a, b)

Table 2. The summary of variables used in different cubist models.

Model Type	#	Geolocation & Elevation	Day of Year	LST	Radiation Variables	Surface Variables
Clear Sky (CLR)	0	Yes	Yes	Yes	-	-
	1	Yes	Yes	-	Yes	-
	2	Yes	Yes	Yes	-	Yes
	3	Yes	Yes	-	Yes	Yes
	4	Yes	Yes	Yes	Yes	-
	5	Yes	Yes	Yes	Yes	Yes
Cloudy Sky (CLD)	0	Yes	Yes	Gap-filled	-	-
	1	Yes	Yes	-	Yes	-
	4	Yes	Yes	Gap-filled	Yes	-
All Sky (ALL)	0	Yes	Yes	Gap-filled	-	-
	1	Yes	Yes	-	Yes	-
	4	Yes	Yes	Gap-filled	Yes	-

Table 3. The comparison of cubist model training statistics for different models listed in Table 2.

Model Type	Model Number	Bias (℃)	RMSE (℃)	R^2	Data Count
Clear Sky Model (CLR)	0	-0.134	1.373	0.978	102,457
	1	-0.027	1.937	0.955	102,457
	2	-0.133	1.344	0.979	102,457
	3	-0.039	1.864	0.958	102,457
	4	-0.111	1.291	0.980	102,457
	5	-0.108	1.265	0.981	102,457
Cloudy Sky Model (CLD)	0	-0.091	1.618	0.969	268,938
	1	-0.018	2.058	0.949	268,938
	4	-0.096	1.484	0.974	268,938
All Sky Model (ALL)	0	-0.104	1.549	0.972	371,395
	1	-0.021	2.048	0.951	371,395
	4	-0.106	1.434	0.976	371,395

Table 4. The comparison of statistics for validation results for different cubist models listed in Table 2. In this table, the validation for all sky models is further separated for clear sky and cloudy sky data.

Model Type	Model Number	Data Type	Bias (℃)	RMSE (℃)	R^2	Data Count
Clear Sky Model (CLR)	0	-	-0.141	1.638	0.967	40,528
	1	-	-0.085	2.373	0.931	40,528
	2	-	-0.144	1.637	0.967	40,528
	3	-	-0.096	2.372	0.932	40,528
	4	-	-0.116	1.643	0.967	40,528
	5	-	-0.113	1.631	0.967	40,528
Cloudy Sky Model (CLD)	0	-	-0.027	1.983	0.951	109,713
	1	-	0.021	2.489	0.924	109,713
	4	-	-0.025	1.917	0.955	109,713
All Sky Model (ALL)	0	All	-0.067	1.884	0.957	150,241
		Clear	0.028	1.647	0.967	40,528
		Cloudy	-0.106	1.986	0.952	109,713
	1	All	-0.024	2.460	0.927	150,241
		Clear	0.097	2.362	0.932	40,528
		Cloudy	-0.080	2.496	0.924	109,713
	4	All	-0.059	1.837	0.959	150,241
		Clear	-0.069	1.634	0.967	40,528
		Cloudy	-0.058	1.922	0.954	109,713

Table 5. The statistics of all-sky Cubist model training/validation using different window size for temporal gap-filling for MODIS LSTs. The model with the window size of 0 day is the full clear-sky model (i.e., CLR-5 in Table 2).

Window Size	Training Samples			Validation Samples		
	Data Availability	Bias (℃)	RMSE (℃)	Data Availability	Bias (℃)	RMSE (℃)
0 day	27.55%	-0.108	1.265	26.88%	-0.113	1.631
±1 days	78.13%	-0.089	1.387	77.04%	-0.096	1.759
±2 days	87.64%	-0.102	1.401	86.41%	-0.024	1.734
±3 days	95.62%	-0.091	1.395	94.40%	-0.067	1.801
±4 days	97.48%	-0.104	1.423	96.38%	-0.027	1.782
± days	99.87%	-0.106	1.434	99.65%	-0.059	1.837

List of figures

Figure 1. (a) The elevation map of the Tibetan Plateau and the location of the China Meteorological Administration (CMA) stations (black triangles) and the Institute of Tibetan Plateau Research (ITP) stations (red pentagram) within the Tibetan Plateau; (b) the elevation distributions of the CMA stations (blue line) and the GMTED DEM for the entire TP (red line).

Figure 2. The overall flowchart of the model training and evaluation strategies of this study.

Figure 3. The schematic of the non-committee Cubist regression model.

Figure 4. The structure of the rule based Cubist regression model with committees.

Figure 5. The density scatter plots of all 12 Cubist models listed in Table 2 for training results: (a) CLR-0, (b) CLR-1, (c) CLR-2, (d) CLR-3, (e) CLR-4, (f) CLR-5, (g) CLD-0, (h) CLD-1, (i) CLD4, (j) ALL-0, (k) ALL-1, (l)ALL-4.

Figure 6. The density scatter plots of the validation results for the best model in each category (a) CLR-5, (b) CLD-4, (c) clear sky observations of ALL-4, (d) cloudy sky observations of ALL-4, and (e) all sky observations of ALL-4.

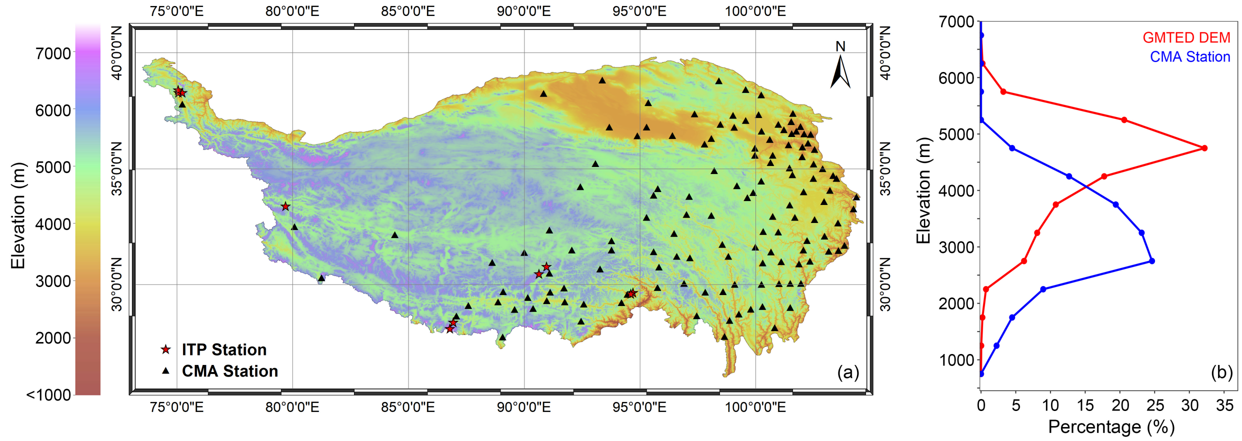
Figure 7. The density scatter plot of the independent validation for the final Cubist model using data of 10 Institute of Tibetan Plateau Research (ITP) stations.

Figure 8. The results of the leave-one-station-out (LOSO) experiment: (a) the spatial distribution of the RMSE of all CMA stations; (b) the scatter plot between the RMSE and the elevation of all CMA stations; (c) the histogram of the RMSE of all CMA stations, where the solid blue vertical line indicates the median value and the two dashed vertical lines refer to the 25% and 75% quantiles respectively; (d) the spatial distribution of the R^2 of all CMA stations. The background colors of (a) and (d) are the elevation of the GMTED2010 data.

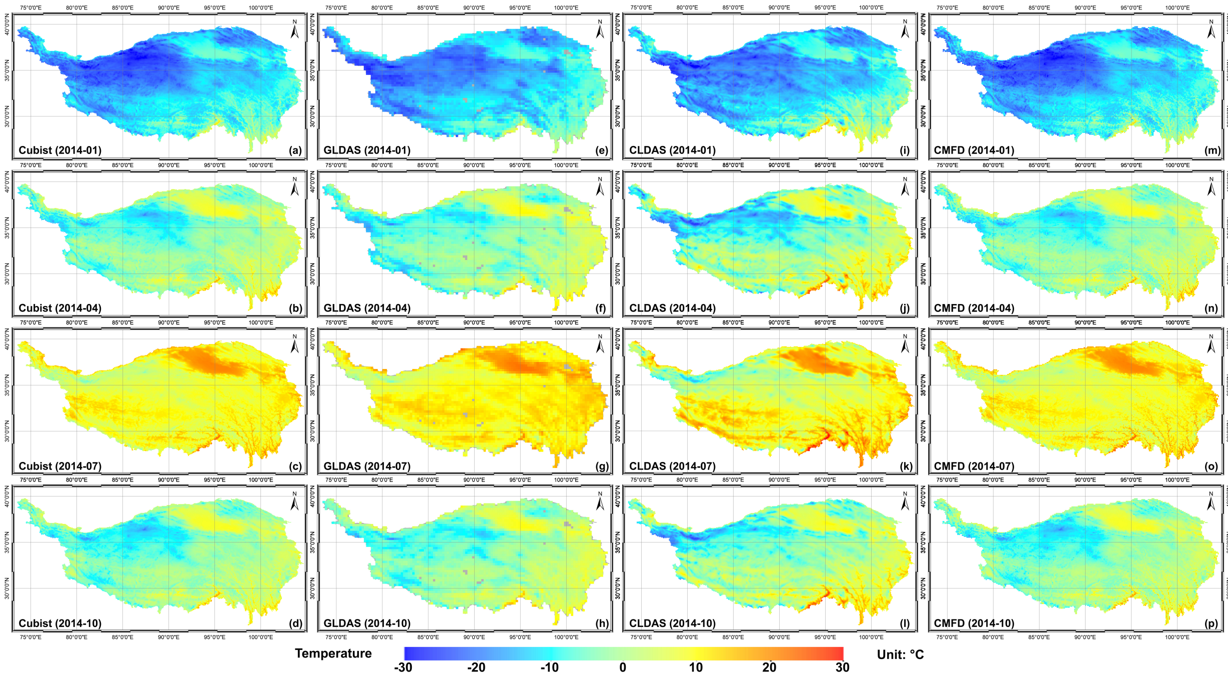
Figure 9. The RMSE of the 25-fold cross validation experiment using all CMA station data.

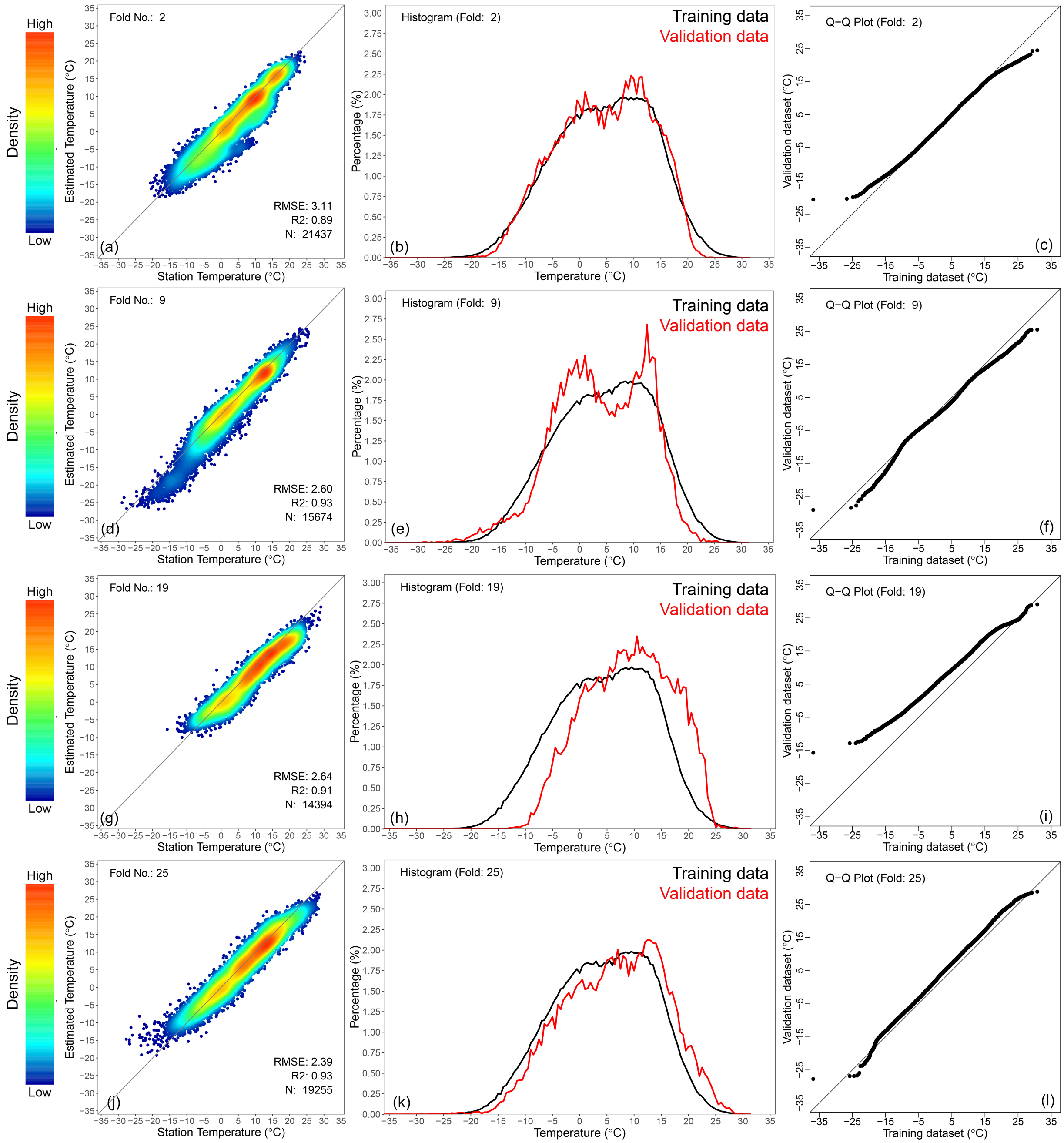
Figure 10. The spatial and temporal patterns of (a-d) the Cubist model estimated surface air temperature (SAT), (e-h) the GLDAS SAT, (i-l) the CLDAS SAT, and (m-p) the CMFD SAT data of January, April, July and October 2014.

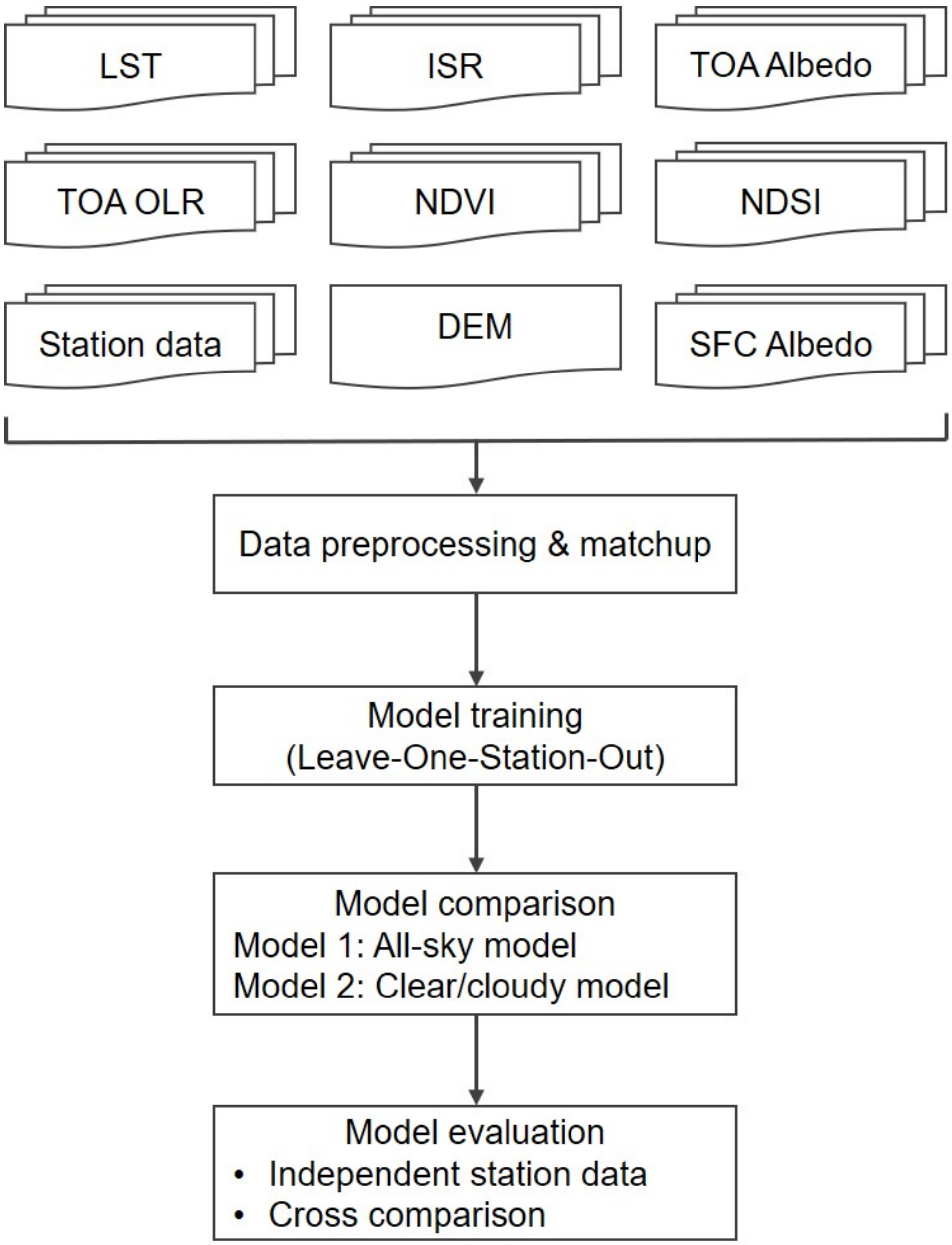
Figure 11. First column (a, d, g, j): the density scatter plots of four Cubist models with largest RMSEs in Figure 9 (i.e., No. 2, 9, 19, 25); second column (b, e, h, k): the comparisons of the data distributions of SATs from the training and validation datasets of each model; third column (c, f, i, l): the quantile-quantile (Q-Q) plots of SATs between the training and validation datasets of each model.

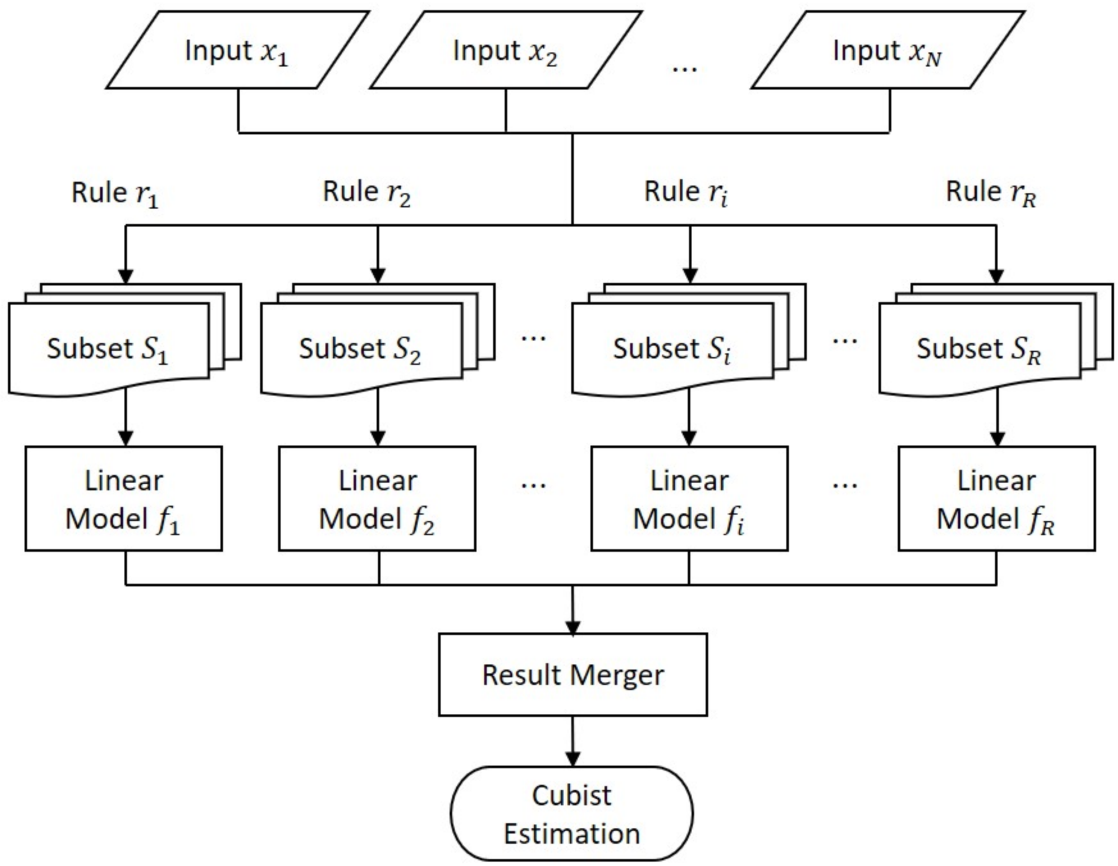


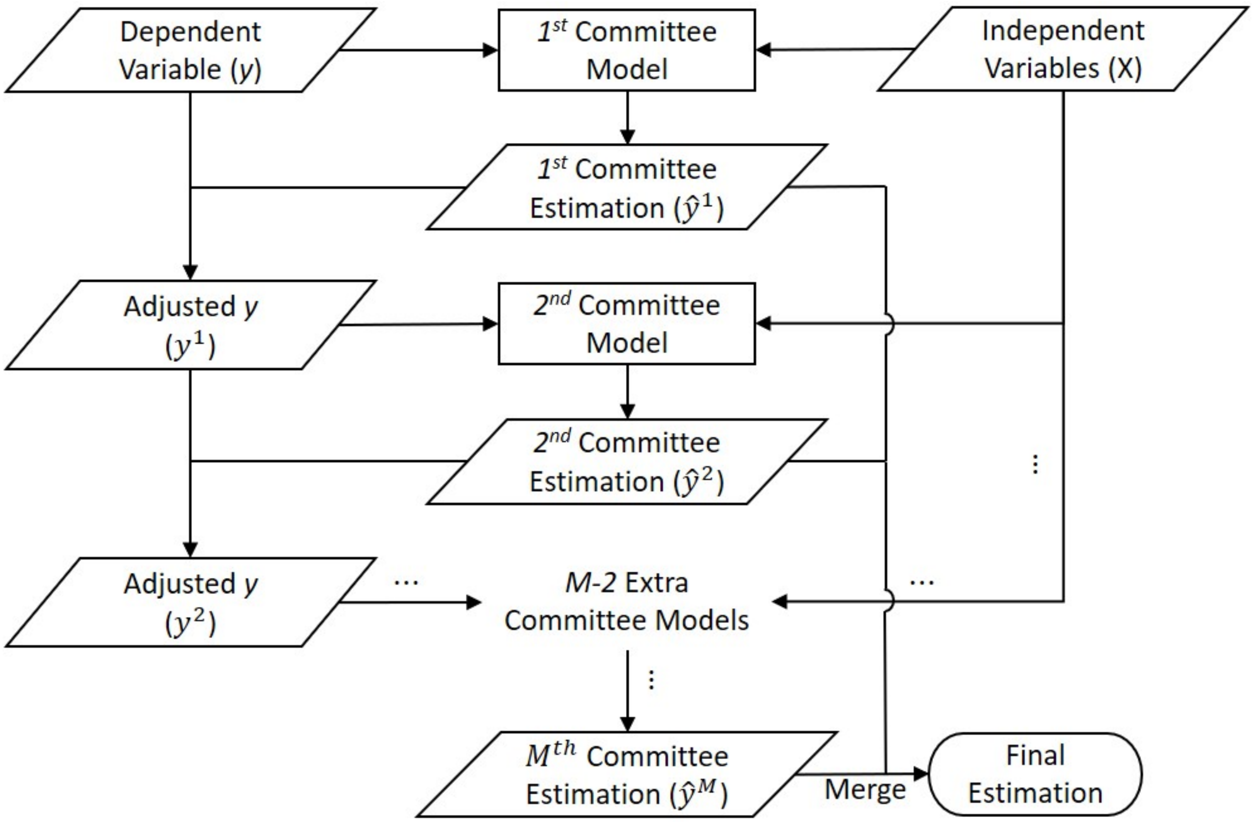
75°0'0"E
80°0'0"E
85°0'0"E
90°0'0"E
95°0'0"E
100°0'0"E
40°0'0"N
35°0'0"N
30°0'0"N
Elevation (m)
7000
6000
5000
4000
3000
2000
<1000
N
ITP Station
CMA Station
(a)
GMTED DEM
CMA Station
Elevation (m)
1000
0
5
10
15
20
25
30
35
Percentage (%)
(b)

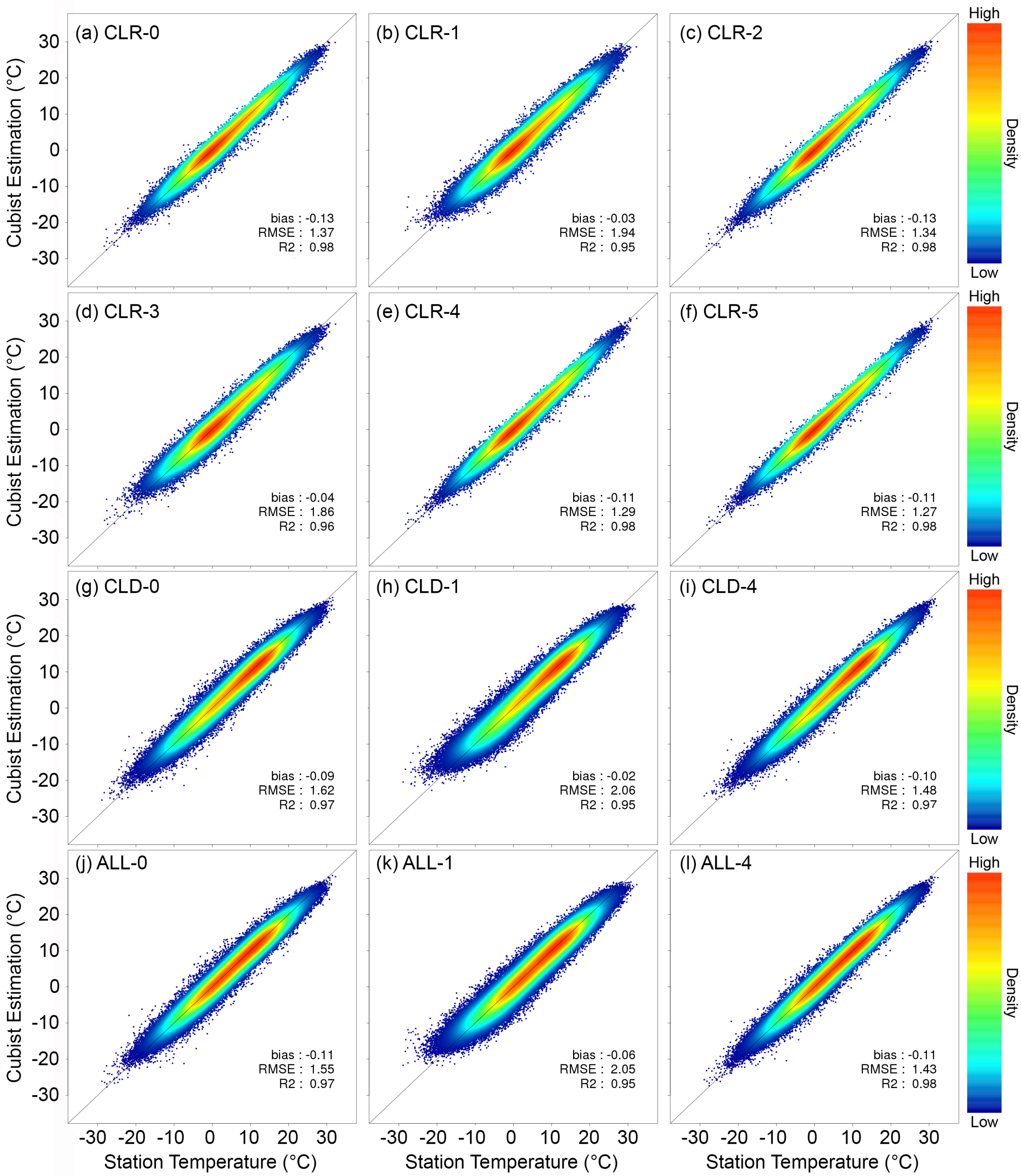

Cubist (2014-01)
(a)
Cubist (2014-04)
(b)
Cubist (2014-07)
(c)
Cubist (2014-10)
(d)
GLDAS (2014-01)
(e)
GLDAS (2014-04)
(f)
GLDAS (2014-07)
(g)
GLDAS (2014-10)
(h)
CLDAS (2014-01)
(i)
CLDAS (2014-04)
(j)
CLDAS (2014-07)
(k)
CLDAS (2014-10)
(l)
CMFD (2014-01)
(m)
CMFD (2014-04)
(n)
CMFD (2014-07)
(o)
CMFD (2014-10)
(p)
Temperature
-30
-20
-10
0
10
20
30
Unit: °C


High
Density
Low
Fold No.: 2
RMSE: 3.11
R2: 0.89
N: 21437
Estimated Temperature (°C)
Station Temperature (°C)
(a)
Histogram (Fold: 2)
Training data
Validation data
Percentage (%)
Temperature (°C)
(b)
Q−Q Plot (Fold: 2)
Validation dataset (°C)
Training dataset (°C)
(c)
Fold No.: 9
RMSE: 2.60
R2: 0.93
N: 15674
(d)
Histogram (Fold: 9)
(e)
Q−Q Plot (Fold: 9)
(f)
Fold No.: 19
RMSE: 2.64
R2: 0.91
N: 14394
(g)
Histogram (Fold: 19)
(h)
Q−Q Plot (Fold: 19)
(i)
Fold No.: 25
RMSE: 2.39
R2: 0.93
N: 19255
(j)
Histogram (Fold: 25)
(k)
Q−Q Plot (Fold: 25)
(l)

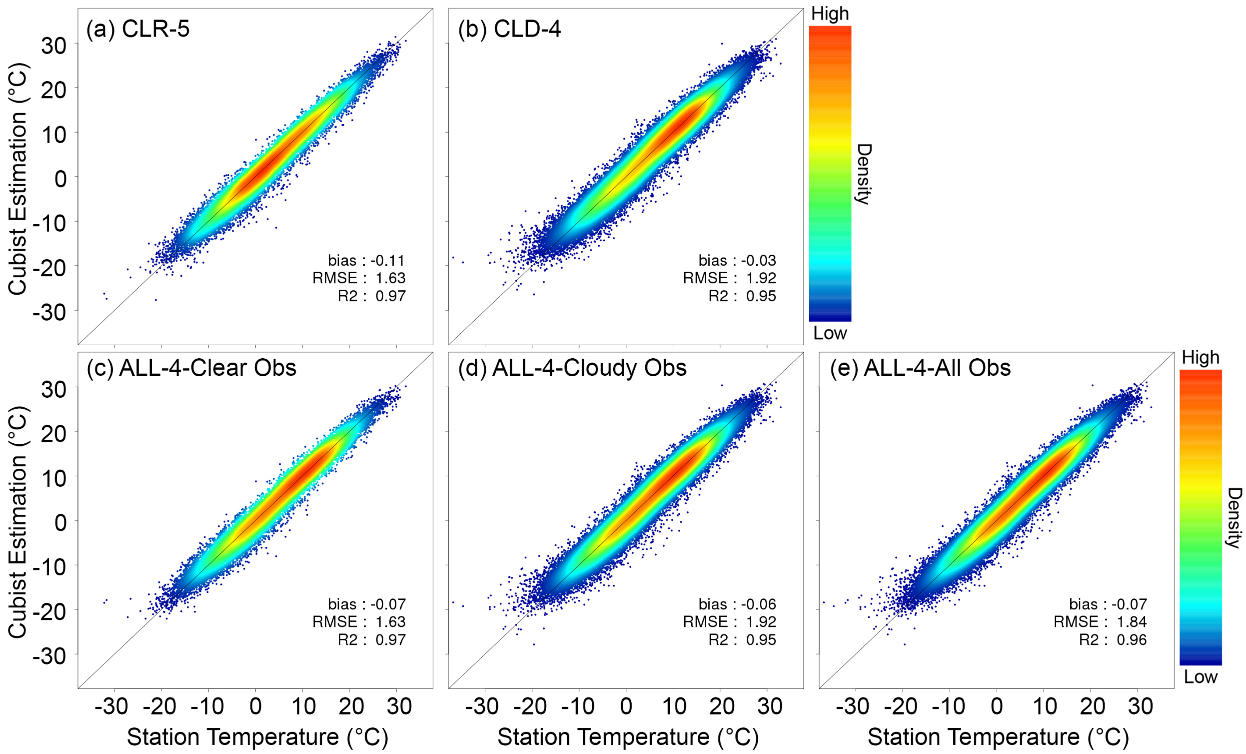

LST
ISR
TOA Albedo
TOA OLR
NDVI
NDSI
Station data
DEM
SFC Albedo
Data preprocessing & matchup
Model training
(Leave-One-Station-Out)
Model comparison
Model 1: All-sky model
Model 2: Clear/cloudy model
Model evaluation
Independent station data
Cross comparison


Input x_1
Input x_2
...
Input x_N
Rule r_1
Rule r_2
Rule r_i
Rule r_R
Subset S_1
Subset S_2
...
Subset S_i
...
Subset S_R
Linear Model f_1
Linear Model f_2
...
Linear Model f_i
...
Linear Model f_R
Result Merger
Cubist Estimation

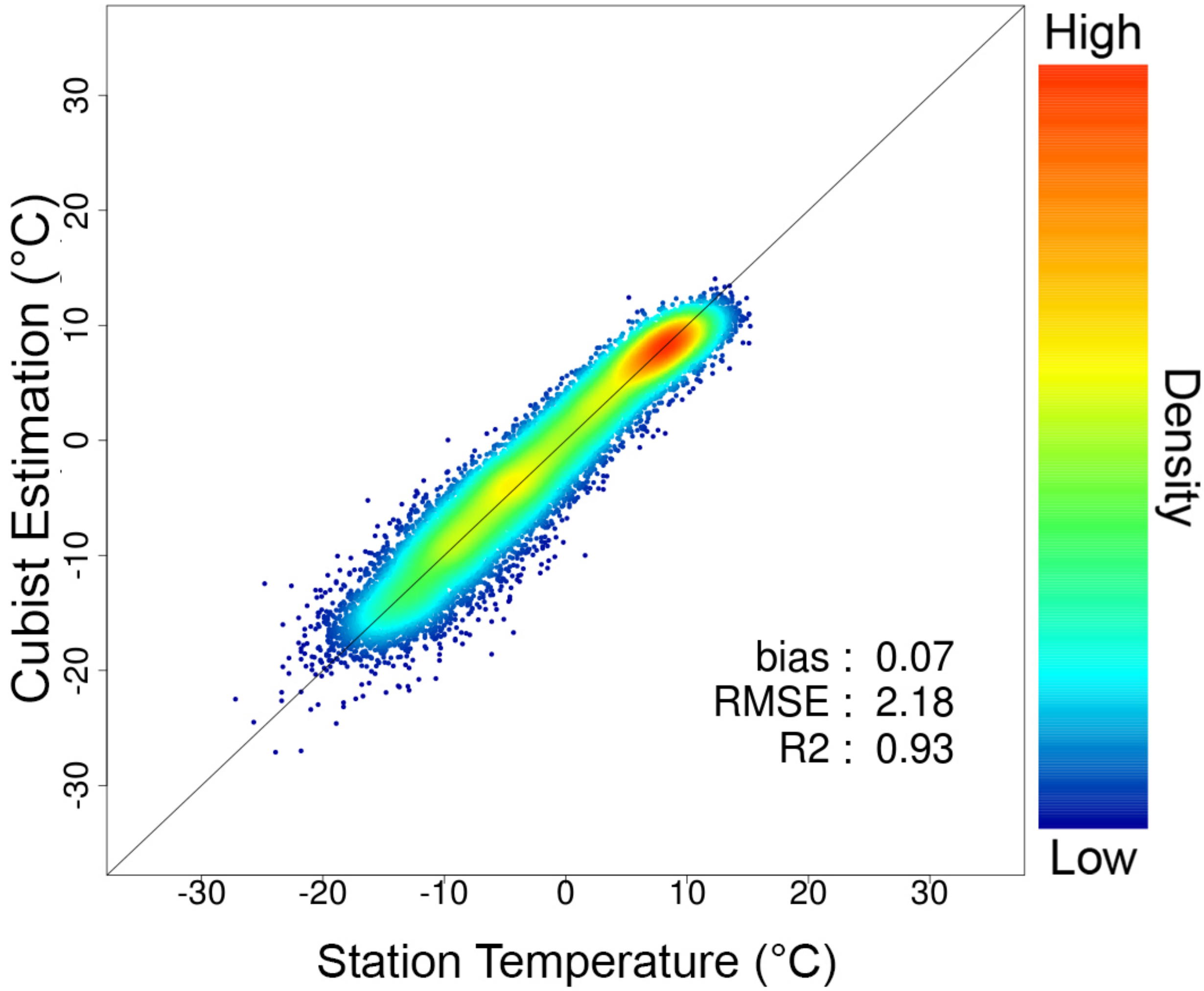

Dependent Variable (y)
1^{st} Committee Model
Independent Variables (X)
1^{st} Committee Estimation ($\hat{y}^1$)
Adjusted y (y^1)
2^{nd} Committee Model
2^{nd} Committee Estimation ($\hat{y}^2$)
Adjusted y (y^2)
M-2 Extra Committee Models
M^{th} Committee Estimation ($\hat{y}^M$)
Merge
Final Estimation

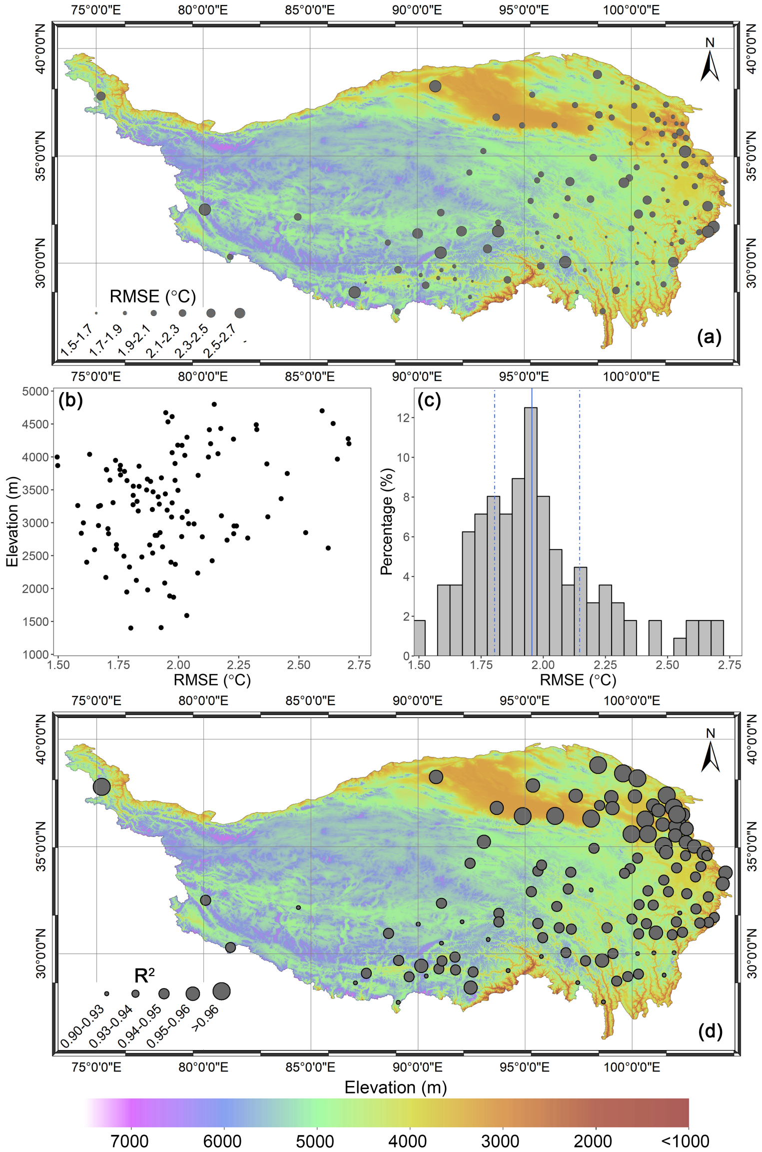


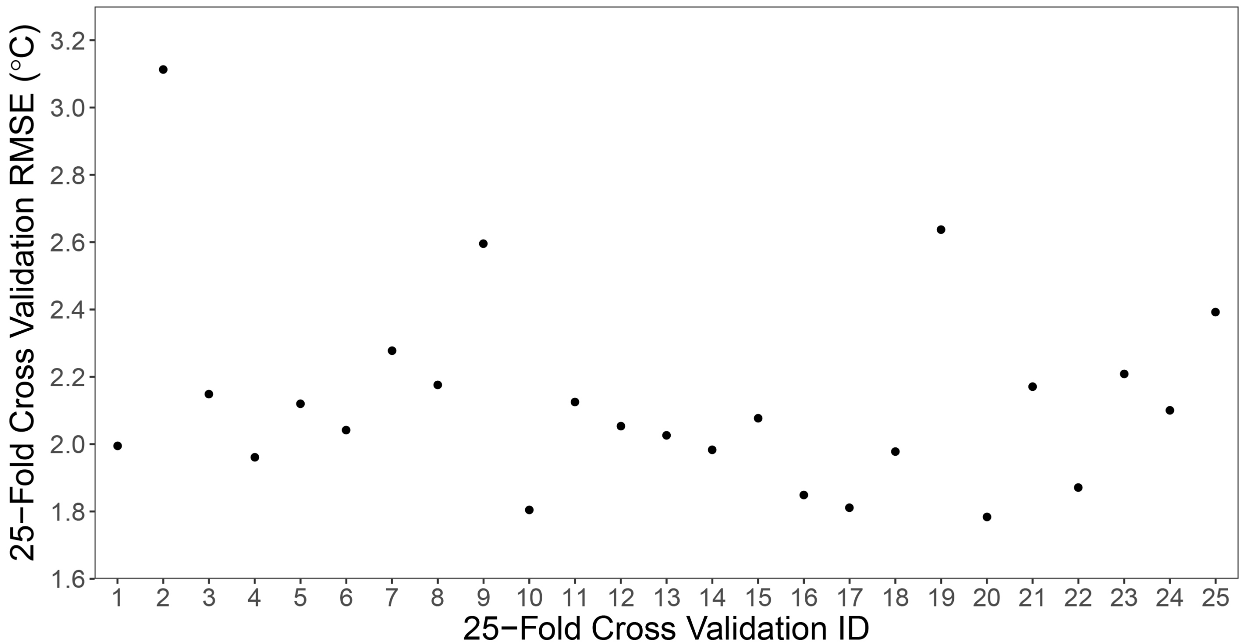
(a) CLR-0
bias : -0.13
RMSE : 1.37
R2 : 0.98
(b) CLR-1
bias : -0.03
RMSE : 1.94
R2 : 0.95
(c) CLR-2
bias : -0.13
RMSE : 1.34
R2 : 0.98
(d) CLR-3
bias : -0.04
RMSE : 1.86
R2 : 0.96
(e) CLR-4
bias : -0.11
RMSE : 1.29
R2 : 0.98
(f) CLR-5
bias : -0.11
RMSE : 1.27
R2 : 0.98
(g) CLD-0
bias : -0.09
RMSE : 1.62
R2 : 0.97
(h) CLD-1
bias : -0.02
RMSE : 2.06
R2 : 0.95
(i) CLD-4
bias : -0.10
RMSE : 1.48
R2 : 0.97
(j) ALL-0
bias : -0.11
RMSE : 1.55
R2 : 0.97
(k) ALL-1
bias : -0.06
RMSE : 2.05
R2 : 0.95
(l) ALL-4
bias : -0.11
RMSE : 1.43
R2 : 0.98
Cubist Estimation (°C)
Station Temperature (°C)
High
Density
Low



(a) CLR-5
bias : -0.11
RMSE : 1.63
R2 : 0.97
(b) CLD-4
bias : -0.03
RMSE : 1.92
R2 : 0.95
High
Density
Low
(c) ALL-4-Clear Obs
bias : -0.07
RMSE : 1.63
R2 : 0.97
(d) ALL-4-Cloudy Obs
bias : -0.06
RMSE : 1.92
R2 : 0.95
(e) ALL-4-All Obs
bias : -0.07
RMSE : 1.84
R2 : 0.96
High
Density
Low
Cubist Estimation (°C)
Station Temperature (°C)
30
20
10
0
-10
-20
-30


High
Density
Low
bias : 0.07
RMSE : 2.18
R2 : 0.93
Cubist Estimation (°C)
Station Temperature (°C)
-30
-20
-10
0
10
20
30


RMSE (°C)
1.5-1.7
1.7-1.9
1.9-2.1
2.1-2.3
2.3-2.5
2.5-2.7
(a)
(b)
Elevation (m)
RMSE (°C)
(c)
Percentage (%)
RMSE (°C)
R²
0.90-0.93
0.93-0.94
0.94-0.95
0.95-0.96
>0.96
(d)
Elevation (m)
7000
6000
5000
4000
3000
2000
<1000


25−Fold Cross Validation RMSE (°C)
3.2
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
25−Fold Cross Validation ID

A Cubist model is developed to estimate daily average SAT for all-sky conditions;

TOA radiation products improve the accuracy of estimated temperature;

A $0.05° \times 0.05°$ daily temperature data (2002-2016) is created over the Tibetan Plateau;