https://dx.doi.org/10.1016/j.rse.2017.08.003

Disaggregation of Remotely Sensed Land Surface Temperature: A Simple yet Flexible Index (SIFI) to Assess Method Performances

Lun Gao[a], Wenfeng Zhan[a, b]*, Fan Huang[a], Xiaolin Zhu[c], Ji Zhou[d], Jinling Quan[e], Peijun Du[f], and Manchun Li[f]

a. Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, International Institute for Earth System Science, Nanjing University, Nanjing, Jiangsu 210023, China

b. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, 210023, China

c. Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic, Hong Kong 999077, China

d. School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China

e. Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

f. Department of Geographic Information Science, Nanjing University, Nanjing, Jiangsu 210023, China

Contact information

∗**Corresponding Author**: W. Zhan, Nanjing University at Xianlin Campus, No.163 Xianlin Avenue, Qixia District, Nanjing, Jiangsu 210023, P. R. China; Fax: +86-25-89681030.

E-mail Addresses: gaolun724@foxmail.com (L. Gao), zhanwenfeng@nju.edu.cn (W. Zhan), nju_huangfan@163.com (F. Huang), zhuxiaolin55@gmail.com (X. Zhu), jzhou233@uestc.edu.cn (J. Zhou), quanjinlin_@126.com (J. Quan), dupjrs@126.com (P. Du), and limanchun@nju.edu.cn (M. Li).

Abstract

Disaggregation of land surface temperature (DLST), the aim of which is to generate LSTs with fine resolution, has been attracting increasing attention since the 1980s. The past three decades have been witness to the emergence of DLST methods in large numbers, the accuracies of which were often assessed by comparing the disaggregated with fine spatial resolution LSTs using error indexes such as the root mean square error (RMSE). However, the majority of previous error indexes are, by their nature, insufficient for assessing the performances of DLST methods. This insufficiency is due in part to their lower competence at distinguishing the DLST error from LST retrieval errors and in part to their inability to remove the process controls resulting from different thermal contrasts, temperature units, and resolution ratios among different scenarios in which DLST is conducted. This is also because they are unable to denote the sharpening statuses of the DLST results (e.g., under- or over-sharpening). This *status quo* has made the evaluation of method performances challenging and sometimes unreliable.

To better assess DLST method performances under diversified scenarios, we formulated five protocols, through which a simple yet flexible index (SIFI) was subsequently designed. The establishment of an SIFI includes the following four steps: (1) a detail-based evaluation, which is designed primarily to exclude the impacts of systematic deviations on estimated LSTs; (2) a Gaussian normalization, which is primarily intended to remove the differences in temperature units and thermal contrasts; (3) a triple comparison, with the aim of attenuating the influence of the difference in the resolution ratio in comparisons of method performances; and (4) a piecewise comparison, which is primarily scheduled to distinguish among the three

sharpening statuses, under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening. The evaluation ability of SIFI was compared with those of the RMSE, Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS), and image quality index (Q) using simulation tests and actual thermal data. The results illustrate that SIFI generally outperforms the other indexes; it is able to mitigate the impacts from process errors and controls during evaluation and is able to indicate the sharpening statuses accurately. We believe this new index will likely promote the design of future DLST algorithms and procedures.

Keywords

Thermal remote sensing; land surface temperature; disaggregation; model performance; and accuracy assessment.

1. Introduction

The large-scale monitoring of the thermal status of land surfaces was difficult and even impossible until the advent of satellite thermal infrared remote sensing. Thermal sensors enable the generation of the land surface temperature (LST) products, which are instrumental to research in many disciplines (Anderson et al., 2012; Bisht et al., 2005; Jiménez-Muñoz et al., 2016; Sandholt et al., 2002; Sobrino et al., 2007, 2012; Teggi, 2012). However, spaceborne sensors are subject to a tradeoff between spatial and temporal resolutions (Zhan et al., 2013), and the spatial resolution of thermal spaceborne-derived LST maps is too coarse for many applications. This challenge has encouraged research on the spatial disaggregation of LST (DLST), which is able to generate LST images with high spatial and temporal resolutions.

A quick literature survey shows that DLST has experienced phenomenal growth in the past three decades, and more methods have been proposed, particularly since the 2000s (Zhan et al., 2013). Most of them tried to reconstruct thermal details with the aid of finer resolution data sets (e.g., data in other bands, classification maps, or designed scaling factors), transform these details into thermal ones by statistical inference, and finally add them to the coarse resolution LSTs (Chen et al., 2014). In considering the fast dynamics of LSTs, recent developments in DLST have been focused on the simultaneous disaggregation of the spatial and temporal resolutions (Addesso et al., 2015; Mechri et al., 2014; Moosavi et al., 2015; Weng et al., 2014; Wu et al., 2015; Zhan et al., 2016; Gao et al., 2017). It is anticipated that DLST will continue to be a research focus in the foreseeable future because the resolutions of the current and planned satellite thermal sensors remain far from satisfactory for the relevant applications (Anderson et al., 2012; Lagouarde et al., 2013; Roberts et al., 2012; Teggi & Despini, 2014; Zhou et al., 2013).

As more and more DLST methods are being proposed, an index that is more appropriate for assessing their performances under various scenarios is urgently required. Early studies on DLST often employ conventional indexes that measure the similarity between disaggregated and fine resolution LSTs, such as the root mean square error (RMSE) (Agam et al., 2007) and the mean absolute error (MAE) (Nishii et al., 1996; Stathopoulou & Cartalis, 2009). Subsequent studies also use the Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS), which is able to eliminate the resolution difference between pre- and post-disaggregation LSTs (Gevaert & García-Haro, 2015; Pardo-Igúzquiza et al., 2006). In considering the documented similarity between DLST and optical image fusion (OIF) since 2010, a few researchers have resorted to the indexes that were designed to evaluate the OIF algorithms. These indexes include the universal image quality index (Q) (Mukherjee et al., 2014; Zhou et al., 2016) and the structural similarity index (SSIM) (Rodriguez-Galiano et al., 2012), among others.

Practitioners may also turn to other advanced indexes in OIF that were recently developed for evaluations, such as the four bands multispectral images fusion index (Q4), the Quality with No Reference (QNR) (Vivone et al., 2014), or the combination of various indexes (Despini et al., 2014). However, although it inherited some traits from the OIF, DLST has differed from its counterpart in the following two regards. First, the DLST process strictly requires that the thermal radiance of a single pixel block at the coarse resolution should be equal to the mean thermal radiance of the corresponding disaggregated fine resolution pixels (Liang, 2005) because its applications are primarily quantitative, whereas spectral distortion is occasionally tolerable in OIF (Pohl & Van Genderen, 1998). Second, fine resolution LSTs, which are either obtained by using thermal data from a different sensor or directly produced by the aggregation-and-then-disaggregation strategy, are indispensable for validation (Zhan et al., 2011). Although references from a different sensor can also help in the evaluation of OIF techniques, they are frequently assessed by comparison with the coarse resolution multispectral images in terms of spectral distortion and with the fine resolution panchromatic images with respect to spatial details (Vivone et al., 2014).

Although previous indexes can be used to assess the performances of DLST methods, they are intrinsically flawed in the following three regards. First, their values depend on multiple errors, including those from disaggregation methods but also those due to the preprocessing of LST, which is unrelated to the model performance (e.g., the temperature retrieval error; more clarifications are given in Section 2.1). Any comparison that disregards the errors due to LST preprocessing would no longer be related to the method performances alone. Second, their values depend on multiple controls, including that from the disaggregation method but also those related to the thermal contrast difference and resolution gap. Finally, their values are mostly not indicative of the sharpening statuses including under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening (more clarifications are given in Section 2.3).

To address these issues, this work designed a new index able to better assess DLST method performances. Followed by the clarifications of background (Section 2) and the five protocols (Section 3) that an index should comply with, Section 4 provides the definition of this index. Sections 5 and 6 exhibit the experiment, the results and discussion, respectively. The conclusions are finally drawn in Section 7.

2. Background

An accurate evaluation of the method performances requires researchers to first identify all the possible errors/controls that may affect the associated evaluation. Generally, the overall errors of disaggregated LSTs (given as err_{overall}) can be expressed as the function of the temperature retrieval errors (given as e_{LST}, including the errors from both the original low-resolution and the reference fine resolution LST images), the image co-registration error (given as e_{cr}), and the DLST error (given as e_{DLST}). In other words, err_{overall} can be expressed as follows:

$$err_{\mathrm{overall}} = q_1(\overbrace{e_{\mathrm{LST}}, e_{\mathrm{cr}}}^{\text{process error}},\ \overset{\text{DLST error}}{e_{\mathrm{DLST}}}) \tag{1}$$

where q_1 is the function between err_{overall} and the three types of errors. Hereafter, we refer to the combination of e_{LST} and e_{cr} as the ‘process errors’ because they primarily stem from the pre-processes that are performed before DLST is conducted (more clarifications are given in Section 2.1).

Nevertheless, it remains unsuitable to use e_{DLST} to represent the performances of the DLST methods because e_{DLST} is also dependent on several other controls in addition to the performance control (given as c_{pm}). These controls are involved in scenarios under which the method performance can be distorted; they include scenarios with different thermal contrasts (given as c_{tc}), temperature units (given as c_{tu}), and resolution ratios (given as c_{rr}). Therefore, the DLST error can be given by the following:

$$e_{\mathrm{DLST}} = q_2(\overbrace{c_{\mathrm{tc}}, c_{\mathrm{tu}}, c_{\mathrm{rr}}}^{\text{process control}},\ \overset{\text{performance}}{c_{\mathrm{pm}}}) \tag{2}$$

where q_2 is the function between the DLST error and the associated controls. Hereafter we refer to the combination of c_{tc}, c_{tu}, and c_{rr} as the 'process controls' (refer to Section 2.2 for more details).

The above analysis indicates that the evaluation of method performances should be conducted by c_{pm} rather than by $err_{overall}$. In other words, the impacts from the process errors and controls should be excluded before the precise evaluation of method performances. In addition, the performance evaluations would be further improved, once the sharpening statuses, including the under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening, is determined. Elaborate interpretations of this issue are presented in Section 2.3.

2.1. Process errors

As indicated by Eq. (1) and graphically represented in Fig. 1, the overall errors for disaggregated LSTs include both the DLST and process errors. The process errors can be divided into temperature retrieval and image registration errors.

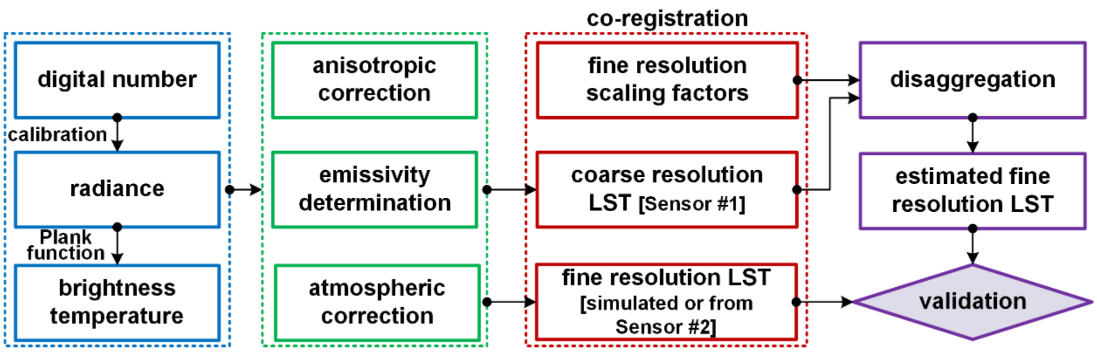

Fig. 1. Graphical representation of the combined temperature retrieval and DLST processes, in which the process errors (including the temperature retrieval and coregistration errors) and DLST errors are blended.

The remote retrieval of the surface temperature is a complex process (Fig. 1). Temperature retrieval errors may be directly due to noise-equivalent temperature differences (NEΔT) (Gillespie et al., 1998), or due to inaccuracies/differences in the conversion from the digital number (DN) to the thermal radiance (i.e., the radiometric calibration process) and then to the brightness temperature (BT) (see Fig. 1). These types of errors depend directly on the calibration coefficients that are estimated from calibration fields (Chander et al., 2009). For LSTs estimated from mature spaceborne thermal sensors that have been carefully calibrated (e.g., MODerate-resolution Imaging Spectroradiometer, MODIS), this type of error is insignificant. However, this error is not trivial for some other spaceborne thermal sensors (e.g., early Landsat series) (Barsi et al., 2007) or when using fine resolution LSTs (e.g., at a meter level) as validation data, which are usually obtained from airborne thermal missions when the calibration may not be adequately accurate (Sobrino et al., 2004).

Temperature retrieval errors may also stem from the uncertainties in the surface thermal anisotropy, determination of emissivity, and atmospheric corrections (see Fig. 1), such as the approximate parameterizations in the mono-window or single-channel algorithms (Qin et al., 2001; Jiménez-Muñoz & Sobrino, 2003). In this regard, these errors reflect the accuracy of the estimation of the true LST from the thermal radiance. Being subject to the inaccurate parameterization of atmospheric thermal radiation, the problems in estimating the emissivity, and the surface thermal anisotropy, the errors are usually lower over relatively homogeneous surfaces (approximately 0.5 to 2.0 K) but considerably higher over heterogeneous terrains (Li et al., 2013). These errors may be much greater over urban areas (reaching 5.0 K or more) due to significant urban thermal anisotropy (Lagouarde et al., 2010).

Errors due to inaccurate temperature retrieval errors are primarily linear and systematic, i.e., the retrieved LSTs compared with the ground truth are systematically higher or lower, with a small portion of errors being nonlinear and random. For example, systematic deviations may occur between the coarse LSTs to be disaggregated and the reference finer resolution LSTs come from other sources (Merlin et al., 2010; Bechtel et al., 2012; Zakšek et al., 2012; Zhou et al., 2015). Errors due to inaccurate calibration are linear because the calibration process itself is often conducted through a linear function (Chander et al., 2009). Errors caused by surface thermal anisotropy are typically systematic for neighboring pixels once the accompanying land cover types are similar, but they may become random for nearby pixels with different land cover types (Lagouarde et al., 2010). Errors caused by inaccurate atmospheric thermal parameterizations are also primarily systematic, because the reflected downward and/or upward atmospheric thermal radiance can be under or over corrected (Li et al., 2013). By comparison, errors that are attributable to inaccurate emissivity depend on the associated estimation process, and these errors can be either systematic or random.

We should note in particular that co-registration errors might also be incorporated into the evaluation of the DLST methods. These errors are derived from the mismatch among the coarse resolution LSTs, the fine resolution scaling factors, and the fine resolution LSTs used for validation (see Fig. 1). Errors due to this type of mismatch (i.e., inaccurate coregistration) depend on the practitioner and are highly nonlinear and random. These errors are difficult to quantify and even more difficult to suppress before the evaluation of the method performances. More discussions on this issue are further provided in Section 6.4.

2.2. Process controls

As indicated by Eq. (2), process controls can also distort the performance evaluation, and they usually include the thermal contrast control (c_{tc}), temperature unit control (c_{tu}), and resolution ratio control (c_{rr}).

First, a DLST may be performed over different areas with a great variety of thermal contrast. For example, an RMSE of 1.5 K for a method that is applied over urban areas with high thermal contrast vs. a value of 1.0 K for another method over vegetated areas does not necessarily indicate that the 1.5 K RMSE method performs worse than the other one. Any index that disregards c_{tc} would no longer be indicative of the method performance. Second, a DLST may be just as well implemented at three levels with different units, including the digital number (DN, no unit) level (Liu and Moore, 1998; Zhukov et al., 1999), the radiance (unit: $\mathrm{W{\cdot}m^{-2}{\cdot}\mu m^{-1}{\cdot}sr^{-1}}$) level (Liu and Zhu, 2012), and the temperature level with centigrade (°C), Fahrenheit (°F), or Kelvin (K) degrees as its units (Zhan et al., 2013). The index values should be comparable when DLST is performed at all these levels. Third, DLST may be further conducted with different resolution ratios between pre- and post-disaggregation LSTs. The resolution ratio usually ranges from several times (e.g., from ~100 to 30 m for the downscaling of Landsat thermal images) (Gao et al., 2017) to as large as several tens of times (e.g., from ~5,000 to 100 m for the downscaling of geostationary thermal images) (Bechtel et al., 2012). Given the identical RMSEs for these two cases, it is understandable that the method performance for the former case will be worse than the latter. Any index that disregards c_{rr} will be uninformative regarding the method performance.

2.3. Sharpening statuses

In general, DLST is used to try to generate fine resolution LST; it can also be perceived as a process that adds thermal details to the background low-resolution LSTs (Chen et al., 2014). Through a DLST method, the added thermal details may be less than or more than needed. In addition, the added thermal details may be ‘much more’ than needed, making the disaggregated LSTs even further away from the reference fine resolution LSTs than the background low-resolution ones. This scenario may sometimes be acceptable for image fusion that aims for target detection, but it is unacceptable for DLST because the application of LSTs is primarily quantitative (e.g., surface flux estimation).

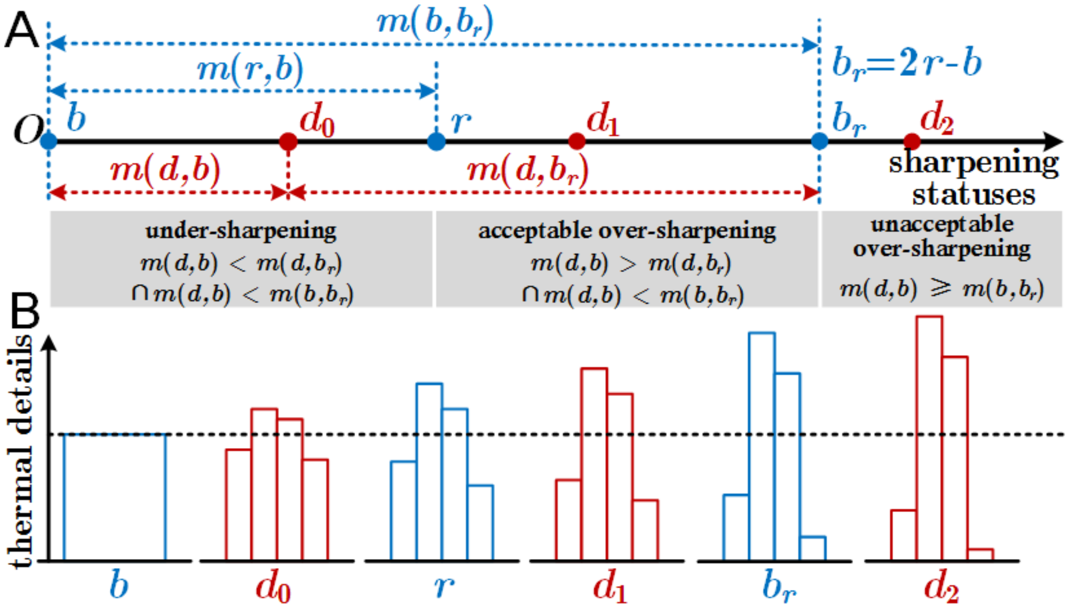

Fig. 2. Conceptual description of the three sharpening statuses (A). The coarse resolution, disaggregated, and fine resolution LST images are denoted by the three dots at b, d, and r, respectively. In the coordinate axis that starts at O, b and r remain constant for a single DLST process while d can be located at any point on this axis depending on the specific DLST method. d_0, d_1, and d_2 represent the

under-sharpening, *acceptable over-sharpening*, and *unacceptable over-sharpening* cases, respectively. b_r is the mirror image of b when using r as the center of symmetry, and it can be estimated by finding $2r - b$. $m(\cdot)$ is a distance metric between two LST images, and it corresponds to the RMSE when the Euclidean distance is used . A simple graphical illustration of the sharpening statuses is further provided in (B), where it is assumed that a single LST pixel is divided into four pixels with different LST values.

We therefore provide the three possible sharpening statuses for the DLST (see Fig. 2), including the *under-sharpening* (corresponding to d_0), *acceptable over-sharpening* (corresponding to d_1) and *unacceptable over-sharpening* (corresponding to d_2). Note that the sharpening statuses in Fig. 2 is displayed in a single dimension of thermal details. Please refer to Appendix A for the description of the sharpening statuses at higher dimensions.

(1) *Under-sharpening*: This term signifies that generally less thermal details are added to the coarse resolution LSTs than needed. In this case, the distance between d (i.e., the disaggregated LSTs) and b (i.e., the background low-resolution LSTs) is shorter than that between d and b_r (i.e., the mirror image of b) and shorter than that between b and b_r, i.e., $m(d, b) < m(d, b_r)$ & $m(d, b) < m(b, b_r)$ (see Fig. 2A), where $m(\cdot)$ is the distance between the two associated LST images.

(2) *Acceptable over-sharpening*: This term implies that more thermal details are added than needed, but these redundant details remain tolerable. In this case, the distance between d and b is greater than that between d and b_r, whereas they are

still lower than that between b and b_r: $m(d, b) > m(d, b_r)$ & $m(d, b) < m(b, b_r)$ (see Fig. 2A).

(3) *Unacceptable over-sharpening*: This status suggests that DLST fails, not just because there are more added thermal details than necessary, but because they also lead to a situation in which the disaggregated LSTs are further away from the fine resolution ones when compared with the original background low-resolution LSTs. In other words, the post-disaggregation results are even worse than the pre-disaggregation ones, a consequence that is intolerable for the quantitative applications of DLST. In this case, $m(d, b) > m(b, b_r)$ (see Fig. 2A).

Note that under-sharpening and acceptable over-sharpening are divided by the reference fine resolution LST image (r), while acceptable and unacceptable over-sharpening are separated by the mirror image of the background LSTs (i.e., b_r). Herein, b_r is defined as the mirror image of b with r as the center of symmetry. In other words, the addition of b and b_r is equal to $2r$ ($b + b_r = 2r$). Physically, the sharpening is no longer tolerable once the disaggregated LSTs possess more thermal details than b_r does because in this case the disaggregated LSTs are less quantitatively accurate even when compared with b (see Fig. 2).

3. Protocols and clarifications for designing a DLST index

A majority of the previous DLST studies used error indexes (e.g., the RMSE), which is commonly directly estimated on the basis of the disaggregated LSTs and the reference fine resolution LSTs to evaluate the performances of the proposed DLST methods (Agam et al., 2007; Zhou et al., 2016). However, the aforementioned analysis shows that, without carefully differentiating the process errors, process controls, and sharpening statuses, the evaluation of the method performances by error indexes such as the RMSE may be inaccurate or even misleading. According to the analysis in Section 2, we propose that a suitable intercomparison index should comply with the following protocols:

PTL #1*: The index should simultaneously measure how much the disaggregated LSTs are similar to the fine resolution LSTs as well as how much the disaggregated LSTs are different from the original (i.e., coarse) LSTs.*

PTL #1 is adopted with adaptations from Wald et al. (1997) in which '*any synthetic image should be as identical as possible to the multispectral set of images that the corresponding sensor would observe with the high resolution*'. In PTL #1, the similarity between the disaggregated and reference fine resolution LSTs (i.e., a pairwise comparison) can be evaluated by distance measures (e.g., the RMSE and MAE). Nevertheless, it remains insufficient to only measure this similarity between the disaggregated and reference LSTs using indexes such as the RMSE because the RMSE can be low, indicating that a high accuracy is achieved, even if the actual DLST procedure fails. For instance, the $m(d_0, r)$ may remain small, even if $d_0 = b$ (i.e., DLST fails or no DLST has been conducted), because the $m(b, r)$ could already be small (see Fig. 2). In other words, the RMSE between the disaggregated and fine resolution LSTs may remain small when no or very few (i.e., under-sharpened) thermal details are added to coarse LSTs because in many cases, the RMSE between the coarse and fine resolution LSTs is already small (e.g., less than 2.0 K). Therefore, the dissimilarity between the disaggregated and coarse resolution LSTs also requires special consideration.

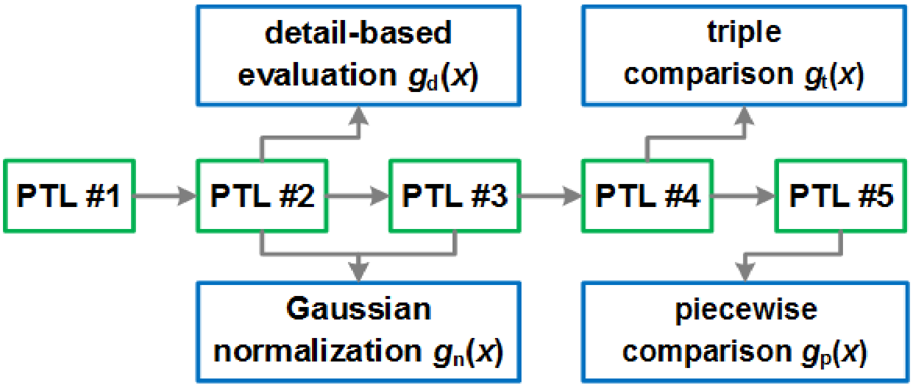

Fig. 3. The five protocols (PTLs) and the associated strategies used to design a suitable index for assessing the DLST results. $g_d(x)$, $g_n(x)$, $g_t(x)$, and $g_p(x)$ represent four functions (or procedures) that characterize the detail-based evaluation, Gaussian normalization, triple comparison, and piecewise comparison required by PTLs #2 to #5.

***PTL #2**: The index should be independent of the temperature retrieval errors.*

PTL #2 addresses temperature retrieval errors and it demands a detail-based evaluation as well as Gaussian normalization for the index design, which are given by the following equations, respectively:

$$g_{\mathrm{d}}(x) = x - \overline{x} \tag{3}$$

$$g_{\mathrm{n}}(x) = (x - \mu_x) \cdot \sigma_b^{-1} \tag{4}$$

where $g_d(x)$ and $g_n(x)$ are the two functions denoting the detail-based evaluation and Gaussian normalization, respectively; x and $\overline{x}$ are the fine resolution LST images and their aggregated ones at the coarse resolution; and μ_x and σ_b are the mean and standard deviation of LST images. It is notable that σ_b, rather than the standard deviations of d and r, is used uniformly for these three LST images because the latter two standard deviations are more sensitive to outliers.

As analyzed in Section 2.1, most of the temperature retrieval errors are linear and systematic. On one hand, the detail-retrieval process expressed by Eq. (3) is able to remove most of the systematic errors due to imprecise atmospheric thermal correction and locally systematic errors due to surface thermal anisotropy as well as a part of the errors due to inaccurate emissivity determinations. It is reasonable that temperature retrieval errors caused by these factors can be suppressed by subtracting the corresponding aggregated LSTs at coarse resolution because LST estimations due to such factors are systematically higher or lower for adjacent pixels. On the other hand, the Gaussian normalization given by Eq. (4) is able to eliminate the linear errors due to inaccurate calibration because this type of normalization is invariant in response to linear transformations. The strict proof is provided in Appendix B.

PTL #3*: The index should be comparable when DLST is performed among areas with different LST contrasts, and it should be comparable when DLST is performed using different types of temperature units.*

PTL #3 addresses the thermal contrast control (c_{tc}) and temperature unit control (c_{tu}). In addition to being able to remove a part of the temperature retrieval errors, the Gaussian normalization $g_{\mathrm{n}}(x)$ given by Eq. (4) is expected to be capable of suppressing c_{tu} because the temperature unit conversion (e.g., from K to °C and to °F) is linear. The Gaussian normalization is also able to suppress c_{tc} once the standard deviation of an image is used to represent its thermal contrast (Wang & Bovik, 2002). Note that ERGAS, with its expression given in Section 5.3, also possesses a normalization factor (i.e., the mean of each band). However, it is unable to address the fully linear unit conversion with both the gain and offset (e.g., from K to °C) as well as the dissimilar scenarios with LST contrast differences, e.g., between the highly heterogeneous urban surfaces and relatively homogeneous cultivated lands.

***PTL #4**: The index should be comparable when DLST is performed with different resolution ratios between pre- and post-disaggregation LSTs.*

PTL #4 addresses the resolution ratio control (c_{rr}). A triple comparison function among the coarse (b), disaggregated (d), and reference fine resolution (r) LSTs can suppress c_{rr}, which is given as follows:

$$g_{\mathrm{t}}(b,d,r) = m(d,r)/m(d,b) \tag{5}$$

where $g_{\mathrm{t}}(\cdot)$ denotes the triple comparison function; and $m(\cdot)$ is a distance metric between two LST images, and it corresponds to the RMSE when the Euclidean distance is used. In two disaggregation cases when the $m(d, r)$ remain the same but the in-between resolution gaps are different, it is reasonable that the case with a larger resolution gap indicates a better model performance. Eq. (5) is efficient at suppressing c_{rr} in such cases. This is because the specific case with a larger resolution gap also likely suggests a higher $m(d, b)$, and with the division given by Eq. (5), the resulting $g_{\mathrm{t}}(\cdot)$ will decrease, consequently indicating a better performance. Note that $g_{\mathrm{t}}(\cdot)$ physically measures the similarity between d and r as well as the dissimilarity between b and r and it is thus related to PTL #1.

***PTL #5**: The index should be indicative of the sharpening status.*

PTL #5 addresses the sharpening statuses. As illustrated in Section 2.3, the differentiation among the three sharpening statuses requires a piecewise function that considers the position of d on the axis shown in Fig. 2A, with d_0, d_1, and d_2 denoting the under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening, respectively. In combining Eq. (5) as required by PTL #4, we provide the following piecewise function to satisfy PTL #5:

$$g_{\mathrm{p}}(b,d,r)=\begin{cases} m(d,r)/m(d,b), & d=d_0,\ m(d,b)\le m(d,b_r)\cap m(d,b)<m(b,b_r) \\ -m(d,r)/m(d,b_r), & d=d_1,\ m(d,b)>m(d,b_r)\cap m(d,b)<m(b,b_r) \\ \mathrm{NaN}, & d=d_2,\ m(d,b)\ge m(b,b_r) \end{cases} \quad (6)$$

where $g_{\mathrm{p}}(\cdot)$ is the piecewise function; NaN indicates that the disaggregated LSTs are unacceptable for quantitative applications. Note that (1) for the acceptable over-sharpening, $m(d, b_r)$ rather than $m(d, b)$ is used as the division factor, aiming at weighting d_0 and d_1 equally once they have the same distance away from r; and (2) the minus symbol when $d = d_1$ is used to differentiate the acceptable over-sharpening from the under-sharpening.

4. Definition

4.1. Standard definition

Using guidance from the proposed protocols, we were able to design a simple yet flexible index (known as the SIFI hereafter) to assess the performances of the DLST methods. Its standard definition is given as follows:

$$\text{SIFI}=\begin{cases} m(D,R)/m(D,B), & m(D,B)\le m(D,B_R)\cap m(D,B)<m(B,B_R) \\ -m(D,R)/m(D,B_R), & m(D,B)>m(D,B_R)\cap m(D,B)<m(B,B_R) \\ \text{NaN}, & m(D,B)\ge m(B,B_R) \end{cases} \tag{6}$$

where B, D, and R are the three variables that correspond to the background coarse resolution (b), disaggregated (d), and reference fine resolution (r) LSTs, respectively; and B_R is the mirror image of B that use R as the center of symmetry, i.e., $B_R = 2R - B$. They are obtained by using following equations:

$$\begin{gathered} X = g_{\mathrm{n}}\left(g_{\mathrm{d}}(x)\right) \\ \begin{cases} g_{\mathrm{d}}(x) = x - \overline{x}; \\ g_{\mathrm{n}}(x) = (x-\mu_x)\cdot\sigma_b^{-1} \end{cases} \end{gathered} \tag{7}$$

where X denotes B, D, or R, while x denotes b, d, or r; and $g_{\mathrm{d}}(x)$ and $g_{\mathrm{n}}(x)$ and their associated variables are well explained in the text subsequent to Eqs. (3) and (4). The piecewise functions given by Eq. (6) represent the under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening, respectively.

From Eq. (6), one can infer that the SIFI ranges from negative to positive infinity. SIFI approximates to zero once the disaggregated LSTs are close to the fine resolution LSTs (i.e., $m(D, R)\rightarrow 0$). The SIFI becomes very large when few thermal details are added − disaggregated LSTs are relatively close to (but not completely equivalent to) the coarse LST (i.e., $m(D, B)\rightarrow 0$). The SIFI then becomes negatively large when more thermal details than needed have been added − disaggregated LSTs are close to the mirror of the coarse LSTs (i.e., $m(D, B_R) \rightarrow 0$). The SIFI is assigned as 'NaN' when redundant details (that would make the DLST fail) have been added (i.e., $m(D, B) \geq m(B, B_R)$). In general, the variation in SIFI is continuous for the under-sharpening and acceptable over-sharpening (refer to Fig. 5 for more visual illustrations); it becomes discontinuous for unacceptable over-sharpening by setting its value as NaN. From Eq. (7), one can also infer that for the under-sharpening, the smaller the SIFI values, the better the DLST results, while this phenomenon is reversed for the acceptable over-sharpening. Note that $m(D, B)$ will always be greater than zero because at least '*some*' details may be added by the DLST process. In addition, this study calculates the distance metric $m(\cdot)$ between two LST images in a global fashion, i.e., only a single distance value is estimated for an entire image. More discussions on the moving window based calculation $m(\cdot)$ for two images are provided in Section 6.4.

4.2. Simplifications under specific conditions

The following simplifications can only be justified under particular conditions, and researchers should use Eqs. (6) and (7) to calculate the SIFI when the following conditions/assumptions are not satisfied. First, in considering that the statistical downscaling method is regularly used for DLST (Zhan et al., 2013) and the relationships between the coarse resolution LSTs and scaling factors are usually less represented by the statistical downscaling methods, under-sharpening (i.e., the added thermal details are insufficient) appears more frequently than the other two statuses. Second, the Euclidean distance (i.e., the RMSE) is usually considered the most frequently used similarity metric (More discussions on the use of distance metrics other than the RMSE are provided in Section 6.4). When the over-sharpening does not occur and the Euclidean distance is employed, the SIFI given by Eq. (6) can be simplified into the following equation:

$$\mathrm{SIFI} = \mathrm{RMSE}(D, R) \cdot [\mathrm{RMSE}(D, B)]^{-1} \tag{8}$$

Once the fine and coarse resolution LSTs are also coming from an identical source, i.e., the disaggregated LSTs are validated by the aggregation-and-then-disaggregation strategy, then the SIFI given by Eq. (8) can be further deduced into the following:

$$\mathrm{SIFI} = \frac{\mathrm{RMSE}(d, r)}{\mathrm{RMSE}(b, d)} = \frac{\sqrt{E[(d-r)^2]}}{\sqrt{E[(b-d)^2]}} \tag{9}$$

where $E(\cdot)$ denotes the expectation. Please refer to Appendix C for the proof of the simplification from Eq. (8) to (9). Note that the aggregation-and-then-disaggregation strategy may be feasible for the development of new algorithms, but the aim of the DLST is to generate finer-resolution LSTs. For the validation of disaggregated results by fine resolution LSTs from another source, researchers should use the complete form, i.e., Eqs. (6) and (7), rather than Eqs. (8) or (9), to calculate the SIFI.

5. Experiments

5.1. Datasets and utilities

Two study areas with different surface landcover types were selected (Fig. 4). The first test site, which is labeled BJM, is situated in the northwestern part of the Beijing Metropolis (39°3’6”N – 40°4’33”N; 115°5’24”E – 116°3’40”E). The BJM consists of a mixture of urban, rural, and mountainous surfaces. This site was chosen mostly because of its high heterogeneity, which makes it appropriate for testing model performances. The other site, which is labeled HNP, is located in the Henan Province (34°0’44”N – 34°7’60”N, 114°0’17”E – 115°0’22”E), and it corresponds to an area covered by fallow field, wheat paddock, and small towns, with a substantially flat terrain. We chose the HNP because the DLST over rural areas is one of its most important applications (Agam et al., 2007; Bindhu et al., 2013).

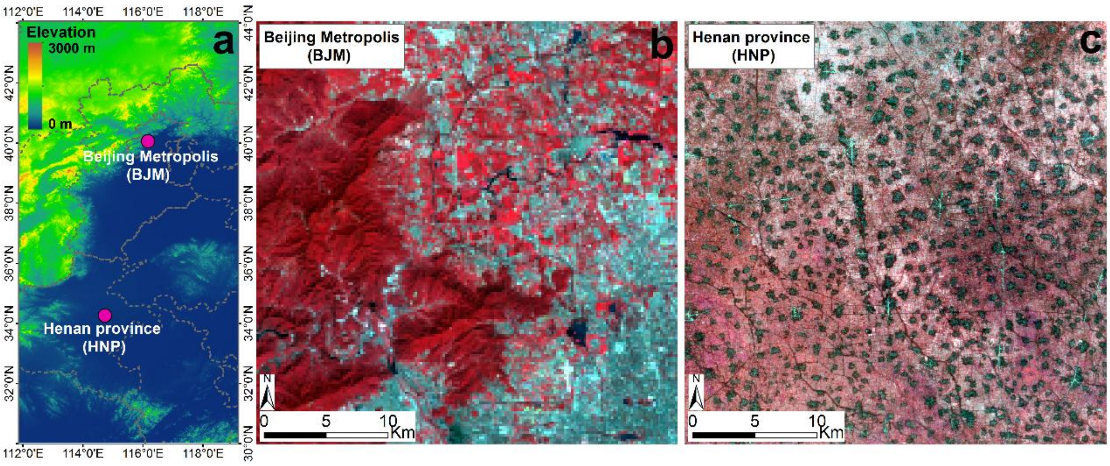

Fig. 4. Geographical location of the two test areas. (a) shows the region over North and East China; (b) demonstrates the northwestern section of the Beijing Metropolis (BJM), as represented by indicating the ASTER bands 3, 2, and 1 as the red, green, and blue channels, respectively; and (c) describes a typical area in central Henan Province (HNP), as provided by indicating TM bands 4, 3, and 2 as the associated

channels. The spatial resolutions of (b) and (c) were both aggregated to 200 m from their original resolutions.

To validate SIFI in its ability to attenuate the impacts, process errors and controls on the evaluation of the method performances, three datasets captured from three satellite sensors were prepared. The first dataset was acquired by Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) at the BJM on August 31, 2004. This dataset includes the spectral reflectance and the associated LST product (AST08, with the spatial resolution of 90 m). The original LSTs (90 m) were aggregated into coarse resolutions, on grids of 100, 200, 400, 800, and 1000 m. The second dataset was obtained by MODIS. It was acquired simultaneously with the ASTER data, and it primarily includes the MODIS LST product (MOD11A1, with a resolution of 1000 m). The third dataset was acquired by the Thematic Mapper (Landsat-5) at the HNP on September 22, 2009. The associated LSTs (with a resolution of 120 m) were retrieved by the mono-window algorithm (Qin et al., 2001) and were further aggregated into coarse resolution datasets, on grids of 200, 400, 800, and 1000 m.

To validate the SIFI's ability to remove the temperature retrieval errors (for PTL #2), the upscaled ASTER LSTs at a resolution of 1000 m were systematically shifted by a constant value (including 1.0, 2.0, and 3.0 K) and were then disaggregated into 200 m, which was then compared with the upscaled ASTER LSTs at 200 m. In addition, the MODIS LSTs at 1000 m were also disaggregated into 200 m and were referenced to the upscaled ASTER LSTs at 200 m. To illustrate the SIFI's independence from the thermal contrast control (for PTL #3), the upscaled ASTER and TM LSTs at 1000 m over both the BJM and HNP, where the thermal contrasts differ, were disaggregated into 200 m and compared with the corresponding reference LSTs at 200 m. To show the SIFI's competency at excluding the temperature unit control (for PTL #3), the upscaled 1000-m TM (band 6) thermal radiance (unit: $W \cdot m^{-2} \cdot \mu m^{-1} \cdot sr^{-1}$) and LSTs (unit: K) were both disaggregated and compared with the reference 200-m radiance and LSTs, respectively. To show the SIFI's ability to attenuate the resolution ratio control (for PTL #4), the upscaled ASTER LSTs at 1000, 800, and 400 m were disaggregated into 200 and 100 m and compared with the reference fine resolution LSTs. To show the SIFI's ability to interpret the sharpening statuses (for PTL #5), the upscaled ASTER LSTs at 1000 m were also disaggregated into 200 m using various DSLT methods.

5.2. Generation of a series of DLST methods

The performances of the DLST methods primarily depend on the chosen scaling factors and regression tool as well as the window size used for regression (Zhan et al., 2013). This study employed a series of scaling factors and moving window sizes to generate a large number of DLST methods with different performances, while the regression tool was kept unchanged during the evaluation process, and it was designated the quadratic function (Kustas et al., 2003). We acknowledge that advanced regression tools, such as the support vector machine, are usually able to produce better disaggregation results than simple polynomial functions (Keramitsoglou et al., 2013; Ghosh & Joshi, 2014). Nonetheless, the aim of this article is to evaluate method performances rather than to develop high-accuracy methods.

The following scaling factors or their combinations were used: the normalized difference water index (NDWI) (McFeeters, 1996), the panchromatic band (Liu &

Moore, 1998), the normalized difference vegetation index (NDVI) (Kustas et al., 2003), the vegetation fraction (f_v) (Agam et al., 2007), the emissivity (ε) (Nichol, 2009), the product $f_\mathrm{v}\cdot\varepsilon$ (Stathopoulou & Cartalis, 2009), the albedo (Dominguez et al., 2011), the normalized multi-band drought index (NMDI) (Liu & Zhu, 2012), the normalized difference built-up index (NDBI) (Wang et al., 2014), and all multi-spectral bands of ASTER or TM (Ghosh & Joshi, 2014). With the local regression strategy (Gao et al., 2017), the moving-window sizes ranging from 3×3 to 21×21 pixels were employed.

5.3. Validation strategy

Three indexes commonly used for assessments were employed for comparison. They were the RMSE, ERGAS, and Q, with ERGAS and Q being calculated using the following equations:

$$\begin{cases} \mathrm{ERGAS} = 100 \cdot (L_\mathrm{r}/L_\mathrm{b}) \cdot \sqrt{\dfrac{\mathrm{RMSE}(d,r)}{\mu_\mathrm{r}}} \\ Q = 4\sigma_{dr}\mu_\mathrm{d}\mu_\mathrm{r}(\sigma_d^2+\sigma_r^2)^{-1}\cdot(\mu_\mathrm{d}{}^2+\mu_\mathrm{r}{}^2)^{-1} \end{cases} \tag{10}$$

where L_b and L_r are the spatial resolutions of the background coarse resolution LSTs (i.e., b) and the reference fine resolution LSTs (i.e., r); σ_{dr} is the covariance between the disaggregated LSTs (i.e., d) and r; μ_d and μ_r are the associated means; and σ_d and σ_r are the associated standard deviations. For the RMSE and ERGAS, their values range from zero to positive infinity, in theory. Their values are zero once the best results have been achieved, while their values become greater with poorer results. For Q, its values change from -1.0 to 1.0, with 1.0 indicating the best obtained results (Wang & Bovik, 2002). This study did not consider the mean absolute error (MAE) and the structural similarity index measure (SSIM) that were used for DLST

(Rodriguez-Galiano et al., 2012) because these two factors have a parallel performance with the RMSE and Q, respectively.

We used three strategies to validate the feasibility of the SIFI. The first was through simple mathematical simulation tests that only include a small number of pixels (refer to Section 6.1); the second was by using actual thermal data (refer to Section 6.2); and the third was through conceptual comparisons of the functionality and design philosophy among different indexes (refer to Section 6.3). Validations based on real thermal data can be further divided into two relatively separated parts. In the first part, different indexes were compared under scenarios that correspond to the proposed protocols (refer to Sections 6.2.1 to 6.2.4). The second part (refer to Section 6.2.5) compared the different indexes through human visual interpretations (HVIs). The HVI has been demonstrated to be plausible and is widely recognized to obtain a relatively accurate image quality from the human visual perspective (Wang & Bovik, 2002). Twenty-two graduate students majoring in remote sensing were recruited and subsequently asked to assign ranks independently for a group of disaggregated LST images (the image with a better quality has a higher score). The quality of a specific LST image was then calculated by averaging all 22 ranks for this specific image. According to the calculated image qualities, the HVI ranks of a series of LST images were finally designated using positive integers, with the higher rank indicating the better disaggregation result.

6. Results and discussion

6.1. Comparisons based on simple simulation tests

To explain the differences among the RMSE, ERGAS, Q, and SIFI, we present two simple simulation tests here. In the first test, let us consider a single pixel to be disaggregated into half its original resolution, i.e., this pixel is disaggregated into four subpixels (Fig. 5). Let us further assume that the coarse pixel has an LST value of 302 K, while the values of the four subpixels, from left to right and from above to below, are 302 − α, 302 + α, 302 + α, and 302 − α (unit: K; $\alpha > 0$, and it reflects the added thermal detail), with 301, 303, 303, and 301 K being the actual values. The variations in the RMSE, ERGAS, Q, and SIFI as a function of α are provided in Fig. 5.

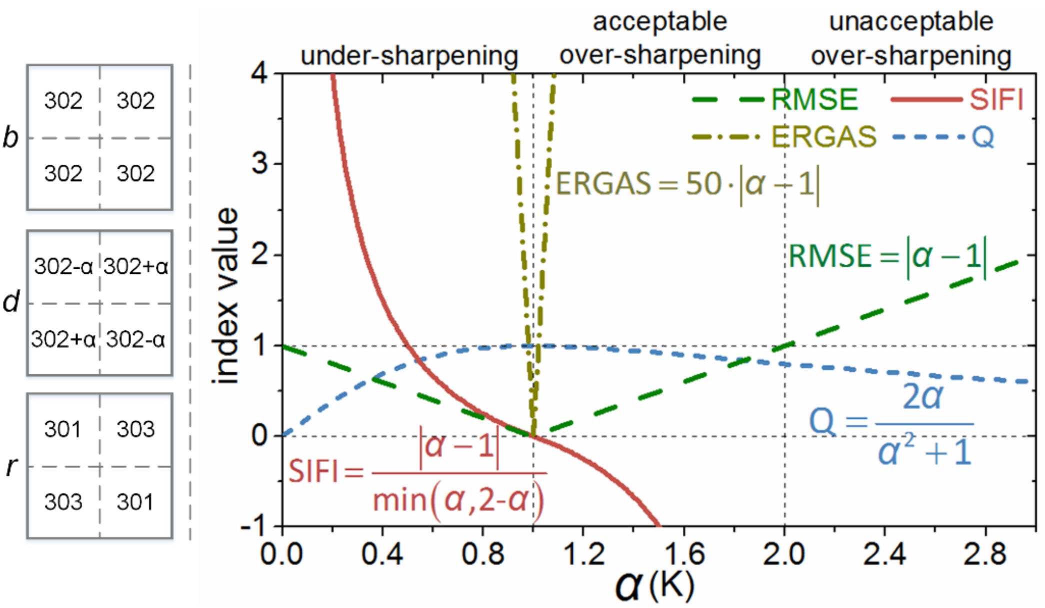

Fig. 5. Variations in the RMSE, ERGAS, Q, and SIFI as a function of the added thermal detail (quantified by α). *b*, *d*, and *r* represent the original coarse resolution, disaggregated, and reference fine resolution LSTs (unit: K), respectively; together,

they represent a simple DLST process in which a single pixel is disaggregated into four subpixels.

This simulation test shows that the natures of the RMSE and ERGAS are similar. Their values both decrease when $0.0 < \alpha \leq 1.0$ and increase when $1.0 < \alpha < +\infty$, and they are both axisymmetric with regard to $\alpha = 1.0$. Q increases from zero to 1.0 when $0.0 < \alpha \leq 1.0$, while it decreases, but in a smoother way, from 1.0 to zero when $1.0 < \alpha < +\infty$, indicating its asymmetry with regard to $\alpha = 1.0$. By comparison, the SIFI changes from positive infinity to zero ($0.0 < \alpha \leq 1.0$) and then to negative infinity ($1.0 < \alpha < 2.0$). SIFI is centrosymmetric when $0.0 < \alpha < 2.0$ but is set as NaN when $\alpha \geq 2.0$. When compared with the RMSE, ERGAS, and Q, the SIFI differ in the following three regards: first, when compared with Q, the symmetry of SIFI, RMSE, and ERGAS shows that SIFI assigns more importance to the quantitative differences between LST images. Second, the values of the three commonly used indexes are unable to indicate the sharpening statuses; however, the calculated SIFI is capable of this type of indication. The background LSTs have been under-sharpened when $0.0 < \alpha < 1.0$ (SIFI > 0.0), acceptably over sharpened when $1.0 \leq \alpha < 2.0$ (SIFI < 0.0), and unacceptably over-sharpened when $\alpha \geq 2.0$ (SIFI = NaN). Third, SIFI is especially sensitive to the case in which DLST is poorly performed because its value rapidly increases when α is close to zero or two, indicating that SIFI would be more suitable for differentiating poor disaggregation results from minor in-between differences.

In the second test, let us consider two pixels to be disaggregated into half of their original resolutions. Two coarse resolution LSTs are disaggregated into eight fine resolution LSTs. The pixel values of the original coarse resolution, disaggregated, and reference fine resolution LSTs are shown in Fig. 6, where δ is a variable that reflects the thermal contrast between these adjacent coarse resolution pixels.

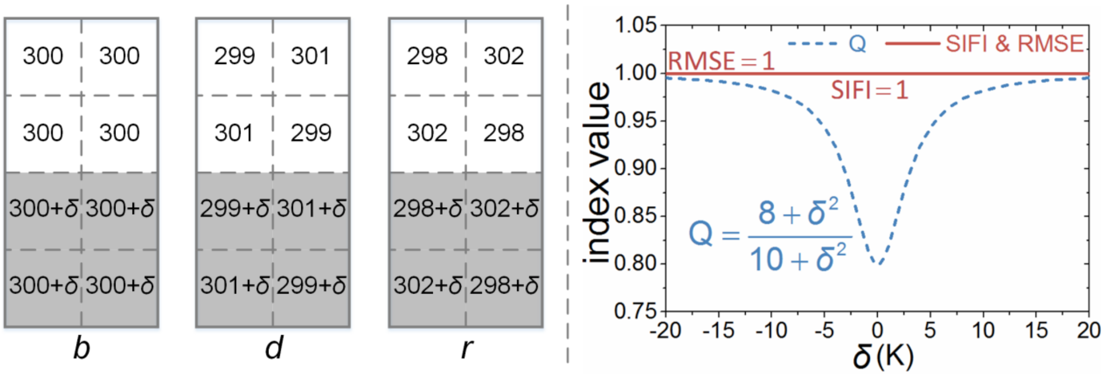

Fig. 6. Variations in the RMSE, Q and SIFI as a function of the thermal contrast between two adjacent pixels, which is represented by δ (unit: K). *b*, *d*, and *r* are the original coarse resolution, disaggregated, and reference fine resolution LSTs, respectively.

The simulation results in Fig. 6 illustrate that Q is a function of δ, while RMSE and SIFI are identically equal to 1.0. Here, the ERGAS is not included due to having similar properties to the RMSE in this case. These simulations again indicate that SIFI and RMSE are more highly related to the absolute quantitative differences between images, while Q varies with the thermal contrast δ. For the assessment of images quality that is specifically perceived through human visualization, the quantitation property (the absolute difference between images) is sometimes unimportant because it will not affect the human interpretation of images (Wang & Bovik, 2002). However, the DLST as commonly shown is used to assist the detailed analysis of the associated quantitative applications such as the upscaling of *in situ* data to the pixel level, urban thermal environment mapping (Zhou et al., 2013), or evapotranspiration estimation (Anderson et al., 2012). These applications require that an index should remain consistent even if the thermal contrast among adjacent pixels varies. From this viewpoint, an index is applicably better when it is invariant with the thermal contrast, and therefore the RMSE and SIFI in this case are more suitable than the Q for quantitative applications of DLST.

6.2. Comparisons based on real thermal data

6.2.1. *Scenario 1 corresponding to PTL #2*

Under this scenario, the capabilities of RMSE, ERGAS, Q, and SIFI are compared when there are temperature retrieval errors (corresponding to PTL #2). Table 1 offers the index values for the cases with various systematic LST retrieval errors. The results show that RMSE and ERGAS vary according to the added systematic error (Δ), and specifically, the RMSE increases from 2.17 to 3.73 K for Cases #1 to #4, while the Q and SIFI remain unchanged for these four cases. This finding demonstrates that the RMSE and ERGAS highly depend on systematic temperature retrieval errors, while Q and SIFI are insensitive to such an error. These results reveal that Q and SIFI are better for evaluating model performances than RMSE and ERGAS because model performances should have been unrelated to the Δ (i.e., temperature retrieval error). The less feasibility by RMSE and ERGAS is also evident by comparing Cases #3 and #5, wherein RMSE and ERGAS are consistent. In theory, the performance should have been worse in Case #5 than that of Case #3 because the scaling factor used for Case #5 is from ASTER, which has co-registration errors with MODIS LST, making the DLST method for Case #5 not well.

Table 1. Comparisons of index values when biases in temperature retrieval occur.

Cases*	TOD**	RMSE (K)	ERGAS	Q	SIFI
#1	AST + 0	2.17	0.14	0.88	1.06
#2	AST + 1	2.41	0.16	0.88	1.06
#3	AST + 2	2.98	0.19	0.88	1.06
#4	AST + 3	3.73	0.24	0.88	1.06
#5	MOD	2.98	0.19	0.82	2.31

* Cases #1 to #5 all used a completely consistent DLST approach with the NMDI as the scaling factor and 7×7 as the moving window size. LSTs were all disaggregated from 1000 to 200 m over the BJM.

** TOD stands for 'type of data'. For 'AST + Δ', the 1000-m LSTs were upscaled from the 200 m ASTER/LSTs, while the validation data were the combination of the 200-m ASTER/LSTs and a systematic error of Δ (unit: K). For 'MOD', the 1000-m LSTs were the MODIS/LSTs, while the validation data were the 200 m ASTER/LSTs.

6.2.2. *Scenario 2 corresponding to PTL #3*

Under this scenario, the indexes are compared when the thermal contrast and temperature units differ (corresponding to PTL #3). The index values over areas with different thermal contrasts are provided in Table 2. The results include three pairs (Cases #1 and #2, Cases #3 and #4, and Cases #5 and #6), each having an identical RMSE by using different combinations of scaling factors and moving window sizes. For each pair, the RMSE and ERGAS have almost identical values, whereas the Q or SIFI show a different behavior. The Q and SIFI values indicate that the disaggregation over the region with a higher thermal contrast (i.e., the BJM, and its thermal contrast defined by standard deviation is 4.4 K) achieves a better result. This interpretation is reasonable because the specific DLST method should possess a better performance over the regions with higher thermal contrasts once the associated RMSE remains unchanged, e.g., the RMSE is 1.38 K for Cases #1 and #2. In other words, when using absolute distances between disaggregated and fine resolution LSTs, RMSE and ERGAS tend to overestimate the model performance over relatively homogeneous regions with a lower thermal contrast, while they underestimate the performance for heterogeneous regions.

The temperature levels considered here include the at-sensor radiance (unit: $W \cdot m^{-2} \cdot \mu m^{-1} \cdot sr^{-1}$) and the LST in Kelvin units. The DN and at-sensor brightness temperature levels were excluded because they have a very significant linear relationship with the radiance (Barsi et al., 2007), and the LST values in Celsius and Fahrenheit were excluded due to their linear relationship with the LST in Kelvin. The results in Table 3 illustrate that the RMSE and ERGAS vary greatly with the temperature level (unit), whereas the Q and SIFI values show small differences. Despite behaving differently at different levels, in practice, the performance of a certain DLST method should not be significantly altered among these levels. These results suggest that the RMSE and ERGAS are inappropriate for this type of assessment, while the latter two indexes are a better option.

Table 2. Comparisons of index values over areas with different thermal contrasts.

Cases*	region	contrast ** (K)	RMSE (K)	ERGAS	Q	SIFI
#1	BJM	4.4	1.38	0.09	0.95	0.69
#2	HNP	1.5	1.38	0.09	0.56	1.30
#3	BJM	4.4	1.44	0.09	0.95	0.71
#4	HNP	1.5	1.44	0.10	0.47	1.40
#5	BJM	4.4	1.58	0.10	0.94	0.79

#6	HNP	1.5	1.58	0.10	0.60	0.93

***** Cases #1 to #6 used six different DLST methods considering ASTER band 9, NDVI, NDVI, $f_v \cdot \varepsilon$, ASTER band 5, and TM band 4, respectively, as the scaling factor and considering 11×11, 9×9, 21×21, 7×7, 9×9, and 7×7 as the moving window size, respectively. LSTs were all disaggregated from 1000 to 200 m.

** The thermal contrasts for the BJM and HNP, as represented by the standard deviation (σ), are 4.4 and 1.5 K, respectively.

Table 3. Comparisons of index values with different temperature units.

Cases*	unit	RMSE	ERGAS	Q	SIFI
#1	$W \cdot m^{-2} \cdot \mu m^{-1} \cdot sr^{-1}$	0.13	0.34	0.62	1.98
#2	K	1.15	0.08	0.64	1.85
#3	$W \cdot m^{-2} \cdot \mu m^{-1} \cdot sr^{-1}$	0.14	0.34	0.57	2.60
#4	K	1.19	0.08	0.56	2.40

***** Cases #1 and #2 used a completely consistent DLST approach with NDVI as the scaling factor and 21×21 as the moving window size, while the DLST approach for Cases #3 and #4 was using vegetation fraction as the scaling factor and 21×21 as the moving window size. The disaggregation was performed for the same LST image over the HNP; and LSTs were all disaggregated from 1000 to 200 m.

6.2.3. *Scenario 3 corresponding to PTL #4*

Under this scenario, the indexes are compared when the resolution gap between the background coarse resolution and reference fine resolution LSTs changes (corresponding to PTL #4). A comparison of the index values when the DLST methods are performed with different initial and target resolutions is presented in Table 4, where the target resolution for Cases #1 to #3 is 100 m, while it is 200 m for

Cases #4 to #6. Different DLST methods, each with a particular scaling factor and moving window size, are considered so that the RMSEs between the disaggregated and the fine resolution LSTs are approximately equivalent or even identical (see Table 4). This type of setting suggests that the image quality of the disaggregated LSTs is similar when taking the RMSE as the error index. However, this result is problematic, when considering that the target resolution is kept unaltered but the initial resolutions changes. Indeed, we expect that the method with the largest resolution gap should have the best performance for the DLST methods. Q has a similar behavior with RMSE in that it hardly changes for all the cases.

Instead, the ERGAS is suitable for these model performance evaluations by considering the spatial resolutions of the pre- and post-disaggregated values (Wald et al., 1997). This idea is as well evidenced by Table 4; with the same RMSE and target resolution, the case with a larger resolution gap points to a better model performance, which is confirmed by the ERGAS values. Note that the SIFI is consistent with the ERGAS in these tests, showing a similar capability to evaluate the model performance under this scenario. The SIFI nevertheless differs from the ERGAS in that the decrease rate of SIFI is not proportional to the resolution ratios (a key variable in ERGAS) — it decreases in a smoother way (see Table 4).

Table 4. Comparisons of index values with different resolution ratios between pre- and post-disaggregated LSTs.

Cases*	Resolution (m)**	RMSE (K)	ERGAS	Q	SIFI
#1	200→100	1.77	0.29	0.93	1.16
#2	400→100	1.76	0.14	0.92	0.97
#3	800→100	1.76	0.07	0.93	0.82

#4	400→200	1.76	0.29	0.92	1.04
#5	800→200	1.76	0.14	0.92	1.12
#6	1000→200	1.76	0.11	0.92	0.78

***** Cases #1 to #6 used the ASTER band 11, vegetation fraction, ASTER band 5, albedo, NMDI, and ASTER band 2, respectively, as the scaling factor and 3×3, 7×7, 21×21, 3×3, 9×9, and 7×7 as the moving window size, respectively. DLST was performed over the BJM.

** The numbers on the left and right of the '→' are the resolutions of the original coarse resolution and the disaggregated LSTs, respectively.

6.2.4. *Scenario 4 corresponding to PTL #5*

Under this scenario, the index values are compared when different amounts of thermal details are added to the background LSTs during the DLST process (corresponding to PTL #5). The thermal details were controlled by a multiplier, which, together with the regression coefficients acquired from the relationships between the background LSTs and scaling factors, determines the amount of added details (Zhan et al., 2011). Table 5 illustrates the associated indexed values when different amounts of thermal details were added by varying the multiplier from 0.5 to 1.9. The results show that the RMSE, ERGAS, and Q change with the multiplier, but their values are unable to specify the sharpening statuses. For example, it is very difficult to judge the boundaries among the three sharpening statuses only according the RMSE values, which alter from approximately 1.3 to 3.3 K for Cases #1 to #8. By contrast, the SIFI values are indicative of the sharpening statuses, evidently specifying that the LSTs are under-sharpened for Cases #1 to #4, acceptably over-sharpened for Cases #5 and #6, and unacceptably over-sharpened for Cases #7 and #8. In detail, Case #1 is designated under-sharpened because the rmse (D, B) (the calculated value is 0.28) $<$ rmse (D, B_R)

(the value is 0.74), while Case #5 shows acceptable over-sharpening because rmse (D, B_R) (0.62) < rmse (D, B) (0.74) < rmse (B, B_R) (0.93), and Case #7 has unacceptable over-sharpening because rmse (D, B) (0.96) > rmse (B, B_R) (0.93).

Table 5. Comparisons of index values with different sharpening statuses.

Cases*	multiplier	RMSE (K)	ERGAS	Q	SIFI	status**
#1	0.5	1.32	0.43	0.96	1.12	**USP**
#2	0.7	1.31	0.43	0.96	0.80	**USP**
#3	0.9	1.46	0.48	0.95	0.69	**USP**
#4	1.1	1.74	0.56	0.93	0.67	**USP**
#5	1.3	2.08	0.68	0.91	-0.80	**AOS**
#6	1.5	2.47	0.80	0.88	-0.92	**AOS**
#7	1.7	2.88	0.94	0.85	NaN	**UOS**
#8	1.9	3.31	1.08	0.81	NaN	**UOS**

***** Cases #1 to #6 all used NDVI as the scaling factor and the statistical regression between LST and NDVI was conducted in a global window (i.e., the entire image); the DLST was performed over the BJM from 1000 to 200 m. Note that the method performances for Cases #1 to #6 are determined by the multiplier coefficient (varying from 0.5 to 1.9), which determines the amount of thermal details that are added to the background LSTs.

** USP, AOS, and UOS denote the under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening, respectively.

6.2.5. *Comparison reference to human visual interpretations*

Under this scenario, the compared index values when referencing human visual interpretations (HVIs) and with respect to different DLST methods (as specified by dissimilar scaling factors and moving window sizes) are shown in Table 6. The HVI ranks are also reported based on the disaggregated LSTs given in Fig. 7. The HVI ranks are higher once the image quality is better. These results show that RMSE and ERGAS are inaccurate for these assessments. The Q and the HVI ranks are sometimes inconsistent. For example, the disaggregated LSTs for Case #3 possess the highest Q (i.e., the best image quality or method performance). However, the corresponding LST image (see Fig. 7e) is not the best among the four disaggregated LSTs; its block effect (also termed the grid effect) is considerably more distinct than the LST image for Case #1 (Fig. 7c). By contrast, the estimated SIFI values are consistent with the HVI ranks with no exception, with a lower SIFI corresponding to a higher HVI rank.

Table 6. Quantitative comparison between SIFI and other indexes for various DLST methods as represented by different scaling factors and moving window sizes.

Cases*	RMSE (K)	ERGAS	Q	SIFI	HVI Rank
#1	1.530	0.124	0.941	0.774	4
#2	1.611	0.131	0.935	0.791	3
#3	1.496	0.122	0.943	0.840	2
#4	1.491	0.121	0.933	0.868	1

***** Cases #1 to #4 used four distinct DLST methods when considering the ASTER band 1, NDVI, NDWI, and $f_v \cdot \varepsilon$, respectively, as the scaling factor and 7×7, 5×5, 9×9, and 21×21 as the moving window sizes. The LSTs were all disaggregated from 800 to 200 m. Note that three significant digits are kept for the index values in this table in particular.

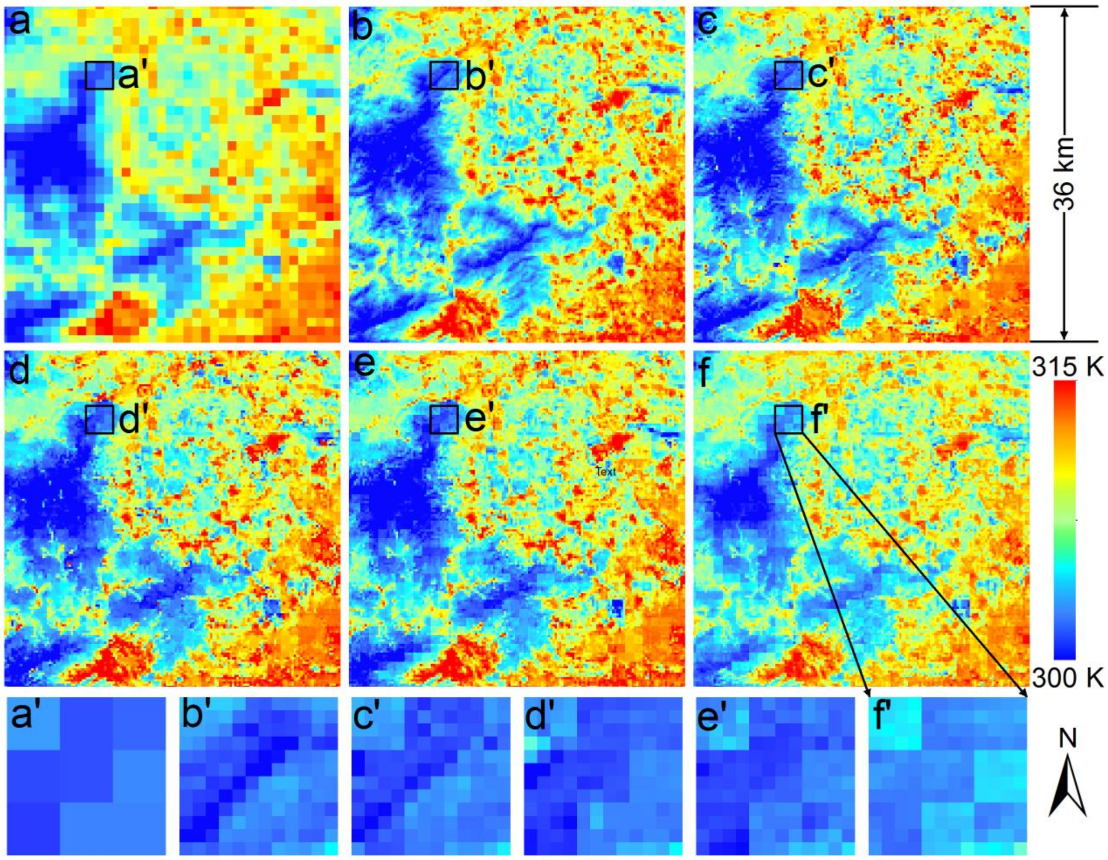

Fig. 7. Coarse resolution, disaggregated, and fine resolution LSTs used for comparison. (a) and (b) are the coarse and fine resolution LSTs, respectively; and (c) to (g) are the disaggregated LSTs corresponding to Cases #1 to #4 in Table 6.

6.3. Conceptual comparisons of the index functionality and structure

The aforementioned results were based on mathematical simulations and real data, and they show that the SIFI assimilates the features of the RMSE, ERGAS, and Q, and abides by all five protocols (see Table 7). For SIFI, once the Euclidean distance is used as the metric as employed in this study, it has the same trait as the RMSE by emphasizing the absolute difference between images (partly corresponding to PTL #1), which is reflected in their symmetry between the under-sharpening and acceptable over-sharpening cases. SIFI also incorporates the trait from Q by combining a normalization process (referred to as $g_d(x)$). This incorporation helps

SIFI be independent of a great portion of the temperature retrieval errors (corresponding to PTL #2) and alleviate the thermal contrast and temperature unit controls (corresponding to PTL #3). SIFI further integrates the ERGAS trait by compensating for the resolution difference between the pre- and post-disaggregated LSTs (PTL #4). The former item achieves this objective by using a triple comparison function (refer to $g_t(x)$), while the latter employs the ratio between the coarse and fine resolutions. SIFI is additionally able to identify the three sharpening statuses through a piecewise function with three different constraints (refer to $g_p(x)$ and corresponding to PTL #5). By contrast, indexes such as the RMSE, ERGAS, and Q comply with only a part of the protocols for DLST (see Table 7). For example, RMSE just partly meets the requirement of PTL #1; it portrays the differences between the disaggregated and reference LSTs but disregards those between the disaggregated and background LSTs; ERGAS only complies with PTLs #1 and #4; and Q is less competent when considering the requirements involved in PTLs #4 and #5.

In addition, SIFI further incorporates the local mean, i.e., the local details, by combining with the detail-retrieval procedure (refer to $g_d(x)$). In this way, the SIFI provides quantitative assessments on the intensity of the block effect in the disaggregated LSTs (Anderson et al., 2011; Zhan et al., 2013). However, the other three indexes only estimate the global statistical variables of the compared images (e.g., the mean and covariance), which render fewer local details.

Table 7. Conceptual comparisons on index functionality in reference to the proposed five protocols.

Protocol*	RMSE	ERGAS	Q	SIFI
PTL #1	✓**	✓	✓	✓✓
PTL #2			✓✓	✓✓

PTL #3			✓✓	✓✓
PTL #4		✓✓		✓✓
PTL #5				✓✓

* PTL #1 requires that the index should simultaneously measure the similarity between the disaggregated and reference LSTs as well as the dissimilarity between the disaggregated and background LSTs. PTLs #2 to #5 correspond to the indexes that should be independent of the temperature retrieval error (PTL #2), thermal contrast and temperature unit control (PTL #3), and resolution ratio control (PTL #4), and should be indicative of the sharpening statuses (PTL #5).

** '✓' indicates that the index abides by some of the requirements of the specific protocol, while '✓✓' indicates that the index completely complies with the protocol.

Researchers often use an indirect validation strategy; LSTs are first upscaled to coarser resolutions, which are then disaggregated again to the original fine resolution, at which point an intercomparison becomes possible (Agam et al., 2007; Rodriguez-Galiano et al., 2012). For this type of validation, the temperature retrieval errors vanish because the coarse resolution LSTs and the fine resolution LSTs used for validation are from an identical source. When this validation strategy is used, simple indexes such as the RMSE are mostly feasible for comparing methods with great differences in performance.

For communications in the DLST community, one may need to judge the performance of a single method many times simply through a single value, and the SIFI will help for this case. Although the SIFI as illustrated here has shown many advantages, in practice, one may also need to know the absolute differences (e.g., the RMSE) between the disaggregated and reference LSTs for practical applications such as the remote sensing of surface fluxes. For example, the widespread use of the Taylor diagram to evaluate the predicated and the reference geophysical variables underlines the importance of summarizing various aspects of the model performance by plotting the RMSE, standard deviation, and correlation coefficient in a diagram (Taylor, 2001). From this perspective, we therefore recommend these indexes, such as the RMSE, ERGAS, Q, correlation coefficient, and SIFI, even along with the estimated distances for calculating the SIFI (i.e., $m(D, R)$, $m(D, B)$, $m(D, B_R)$, and $m(B, B_R)$), which are used collectively to assess a newly proposed method or compare several methods for DLST. For these reasons, we believe the maximum benefit ultimately lies in this approach.

6.4. Problems and prospects

(1) *Problems*

First, the procedures used in the design of the SIFI are able to remove (or alleviate) the linear and systematic process errors/controls; these procedures are nevertheless unable to eliminate the random and highly nonlinear process errors, e.g., the mismatch between the fine resolution scaling factors and the coarse and fine resolution LSTs (i.e., the co-registration error). Practitioners need to be careful to interpret the index values because the co-registration error is fickle. Second, the initial scaling factors always possess a spatial resolution that is much higher than that of fine resolution LSTs. The spatial upscaling of the scaling factors to the resolution of the fine LSTs should consider the point spread function of the sensor (Zhan et al., 2013). The aim of this consideration is to make sure that the true resolution of scaling factor do have the same resolution with the fine LSTs. Otherwise, the RMSE between the disaggregated and reference LSTs will no longer be zero, even if the temperature retrieval as well as the DLST processes are error-free.

(2) *Prospects*

The SIFI proposed in this study is only one alternative that satisfies the proposed protocols. We need to clarify that other strategies that conform to the protocols can also be applied to help design an index even better than SIFI. First, SIFI employs the strategies given by Eqs. (3) and (4) to remove the linear or the locally/globally systematic process errors and controls. Other normalization schemes that are able to remove these associated errors and controls are also feasible. Second, this study mainly uses the Euclidean distance (i.e., the RMSE) as the metric for calculating SIFI. It is expected that distance metrics such as the general Minkowski distance (Han et al., 2011) may generate a parallel capability for assessments. Nevertheless, metrics such as the Euclidean or Minkowski distances give a high weight to outliers and may make the resultant SIFI less indicative of method performances. Therefore, researchers should try to avoid outliers through setting thresholds during the DLST. One may infer that distance metrics that are insensitive to outliers (e.g., the angle cosine distance) are applicable for performance assessments. However, researchers need be very careful to use such metrics because they more emphasize the structural similarity between two images, that is, they ignore the information retained in the absolute values between images, which is yet important for the quantitative applications of DLST. Third, SIFI is calculated for an entire image, i.e., a single SIFI value is calculated for a single evaluation. SIFI may be modified to be dependent on pixel location – a series of local SIFI values can be obtained by setting moving windows on a LST image. By this modification, method performances can be evaluated for different parts within a single image.

SIFI has potential to be further applied to disaggregation/downscaling assessments. The recent rise of the spatio-temporal DLST requires the assessment of sequential fine resolution LSTs rather than a LST image at a single moment. Such assessments, therefore, may be performed by combining the sequential SIFIs within a certain cycle (e.g. the diurnal cycle) (Göttsche & Olesen, 2009). SIFI may be further enhanced to facilitate the assessments of method performances when *in situ* measurements on LST are available. Finally, SIFI and the associated design philosophies may also be used to assess model performances for disaggregation of other satellite products such as precipitation and soil moisture.

7. Conclusions

At present, the performances of DLST methods are evaluated by simple indexes (e.g., RMSE) or more complicated ones that are adapted from optical image fusion (e.g., ERGAS and Q). These indexes are insufficient because not only do they include all the errors involved in the complete process from thermal radiance to LSTs (termed the process error) but also because they are susceptible to process controls, including differences in the thermal contrast, temperature units, and resolution ratios. In addition, these indexes are unable to differentiate among the three sharpening statuses, the under-sharpening, acceptable over-sharpening, and unacceptable over-sharpening statuses. These deficiencies make evaluating the performance of the DLST methods far from precise under different scenarios. It is therefore of great urgency to design a better index for these evaluations.

In considering this issue, five standard protocols were proposed with which a suitable index should be assigned. In being guided under these protocols, a simple yet flexible index (SIFI) was designed. SIFI incorporates the following four procedures: (1) the detail-retrieval procedure $g_d(x)$ that is primarily used to remove the impacts from the temperature retrieval error; (2) the Gaussian normalization $g_n(x)$ primarily aimed at attenuating the controls on the differences in the thermal contrasts and temperature units; (3) the triple comparison $g_t(x)$, which is scheduled to lessen the controls on the difference in resolution ratios; and (4) the piecewise comparison $g_p(x)$ and several provisos (given by several inequalities to indicate the three sharpening statuses).

Comprehensive evaluations show that indexes that include the RMSE, ERGAS, and Q abide by only part of the requirements denoted in the five protocols. The new index SIFI instead complies with all the proposed protocols. This SIFI is able to capture the model performance more accurately; it can remove the impacts from the process errors and controls on evaluations and can indicate the sharpening statuses such that a disaggregation is under- or over-sharpened. Further analysis illustrates that the SIFI attaches more importance to the scenario in which the DLST is poorly performed and therefore is sensitive to the grid effect in the DLST. Note that it remains difficult for the SIFI to remove the highly nonlinear process errors, such as the mismatch error, and there may be better procedures than those used in this study. Nevertheless, SIFI facilitates the comparison of model performances and therefore helps in the further enhancement of methods for DLST. In addition, we believe that the design philosophies of the SIFI are likely applicable to the model performance comparisons for the disaggregation of other geophysical variables.

Acknowledgements

This work is jointly supported by the Key Research and Development Programs for Global Change and Adaptation under Grant number 2016YFA0600201, the National Natural Science Foundation of China under Grant number 41671420, and the Key Research and Development Programs for Global Change and Adaptation under Grant number 2017YFA0603604. We are also grateful for the financial support provided by the DengFeng Program-B of Nanjing University.

Appendix A: Illustration of the three sharpening statuses in high-dimension

The conceptual sharpening statuses for a single pixel block (i.e., a single dimension) was provided in Section 2.3. However, disaggregation of LSTs is conducted for all the pixels of an entire LST image rather than a single pixel block. In other words, the precise three sharpening statuses should be displayed in a high-dimension. Fig. A1 demonstrates the conceptual illustration of the *under-sharpening* (USP), *acceptable over-sharpening* (AOS), and *unacceptable over-sharpening* (USP) in two dimensions, and illustration of even higher dimensions is similar to the two-dimension case.

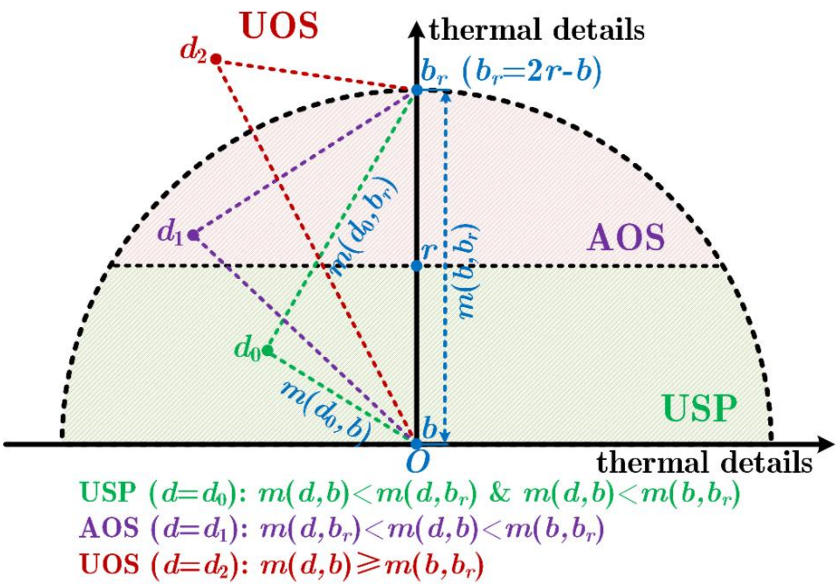

Fig. A1. Two-dimensional conceptual description of the three sharpening statuses. The coarse resolution, disaggregated, and fine resolution LST images are denoted by the three dots *b*, *d*, and *r*, respectively; b_r is the mirror image of *b* when using *r* as the center of symmetry. d_0, d_1, and d_2 represent the cases of *under-sharpening* (USP), *acceptable over-sharpening* (AOS), and *unacceptable over-sharpening* (USP), respectively. $m(d, r)$, $m(d, b)$, $m(d, b_r)$, and $m(b, b_r)$ are the distances between the associated two LST images.

The constraints for differentiating the statuses in the two-dimension are consistent with those in the one-dimension. (1) LSTs are under sharpened ($d = d_0$) when $m(d, b) < m(d, b_r)$ & $m(d, b) < m(b, b_r)$, and therefore the USP can be geometrically represented by the light green region (refer to Fig. A1). (2) LSTs are acceptably over sharpened ($d = d_1$) when $m(d, b) > m(d, b_r)$ & $m(d, b) < m(b, b_r)$, and therefore the AOS corresponds to the light red region. (3) LSTs are unacceptably over sharpened ($d = d_2$) when $m(d, b) > m(b, b_r)$, which corresponds to the region beyond the semicircle.

APPENDIX B: PROOF OF THE SIFI'S INDEPENDENCE OF LINEAR AND SYSTEMATIC ERRORS OR CONTROLS

The standard definition of SIFI given by Eq. (6) is able to remove the linear and/or systematic process errors using the Gaussian normalization. Let us consider two variables T_1 and T_2, and there is a linear relationship between these two variables, given as:

$$T_1 = a_1 \cdot T_2 + a_0 \tag{B1}$$

where a_1 and a_0 are the linear coefficient and constant. The Gaussian normalization of T_1 can be given as follows:

$$g_n(T_1) = \frac{T_1 - u_1}{\sigma_1} = \frac{(a_1 \cdot T_2 + a_0) - (a_1 \cdot u_2 + a_0)}{a_1 \cdot \sigma_1} = \frac{T_2 - u_2}{\sigma_2} = g_n(T_2) \tag{B2}$$

where $g_n(\cdot)$ is a normalization function given by Eq. (4); u_1, u_2, σ_1, and σ_2 are the means and standard deviations for T_1 and T_2, respectively. We therefore prove that SIFI is capable of eliminating the linear and systematic process errors and controls (e.g., a great portion of the temperature retrieval error).

APPENDIX C: SIMPLIFICATION OF SIFI UNDER PARTICULAR ASSUMPTIONS

Once the Euclidean distance (i.e., the RMSE) is employed as the metric and the over-sharpening does not occur, the standard definition of SIFI given by Eq. (6) can be simplified into the following equation:

$$\mathrm{SIFI} = \mathrm{RMSE}(D,R)\cdot[\mathrm{RMSE}(D,B)]^{-1} \tag{C1}$$

where B, D, and R are given by the following:

$$\begin{cases} B = g_{\mathrm{n}}\left(g_{\mathrm{d}}(b)\right) = \sigma_b^{-1}(b-\bar{b}-u_{b-\bar{b}}) \\ D = g_{\mathrm{n}}\left(g_{\mathrm{d}}(d)\right) = \sigma_b^{-1}(d-\bar{d}-u_{d-\bar{d}}) \\ R = g_{\mathrm{n}}\left(g_{\mathrm{d}}(r)\right) = \sigma_b^{-1}(r-\bar{r}-u_{r-\bar{r}}) \end{cases} \tag{C2}$$

where $g_{\mathrm{n}}(x)$ and $g_{\mathrm{d}}(x)$ are given by Eq. (7); b, d, and r are the background coarse resolution, disaggregated, and reference fine resolution LSTs, respectively; σ_x, u_x, and $\bar{x}$ are the standard deviation, mean, and the aggregated coarse resolution LSTs of a LST image (i.e., x). Combining Eqs. (C1) and (C2), the following equations can be deduced:

$$\begin{aligned} \mathrm{SIFI} &= \mathrm{RMSE}(D,R)\cdot[\mathrm{RMSE}(D,B)]^{-1} \\ &= \frac{\mathrm{RMSE}\left(\sigma_b^{-1}(d-\bar{d}-u_{d-\bar{d}}),\sigma_b^{-1}(r-\bar{r}-u_{r-\bar{r}})\right)}{\mathrm{RMSE}\left(\sigma_b^{-1}(d-\bar{d}-u_{d-\bar{d}}),\sigma_b^{-1}(b-\bar{b}-u_{b-\bar{b}})\right)} \\ &= \frac{\mathrm{RMSE}\left(d-\bar{d}-u_{d-\bar{d}},r-\bar{r}-u_{r-\bar{r}}\right)}{\mathrm{RMSE}\left(d-\bar{d}-u_{d-\bar{d}},b-\bar{b}-u_{b-\bar{b}}\right)} \end{aligned} \tag{C3}$$

Once the fine and coarse resolution LSTs are from a same source, i.e., the disaggregated LSTs are validated by the aggregation-and-then-disaggregation strategy, the aggregated coarse resolution LSTs for b, d, and r will be the identical; and the mean of the thermal details for b, d, and r will also be equal to zero. We thus have the following equations:

$$\begin{cases} \bar{b} = \bar{d} = \bar{r} \\ u_{b-\bar{b}} = u_{d-\bar{d}} = u_{r-\bar{r}} = 0 \end{cases} \tag{C4}$$

Combining Eqs. (C3) and (C4), we obtain the final equation:

$$\text{SIFI}=\frac{\text{RMSE}\left(d-\bar{d}, r-\bar{r}\right)}{\text{RMSE}\left(d-\bar{d}, b-\bar{b}\right)}=\frac{\text{RMSE}\left(d, r\right)}{\text{RMSE}\left(d, b\right)} \tag{C5}$$

The above proof therefore finally demonstrates that $\text{RMSE}(D,R)\cdot[\text{RMSE}(D,B)]^{-1}$ is equivalent to $\text{RMSE}(d,r)\cdot[\text{RMSE}(d,b)]^{-1}$.

References

Addesso, P., Longo, M., Maltese, A., Restaino, R., & Vivone, G. (2015). Batch methods for resolution enhancement of TIR Image sequences. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*, 3372-3385, doi:10.1109/JSTARS.2015.2440333.

Agam, N., Kustas, W. P., Anderson, M. C., Li, F., & Neale, C. M. (2007). A vegetation index based technique for spatial sharpening of thermal imagery. *Remote sensing of Environment, 107*, 545-558, doi:10.1016/j.rse.2006.10.006.

Anderson, M. C., Allen, R. G., Morse, A., & Kustas, W. P. (2012). Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. *Remote sensing of Environment, 122*, 50-65, doi:10.1016/j.rse.2011.08.025.

Anderson, M. C., Kustas, W. P., Norman, J. M., Hain, C. R., Mecikalski, J. R., Schultz, L., González-Dugo, M. P., Cammalleri, C., d'Urso, G., Pimstein, A., & Gao, F. (2011). Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. *Hydrology and Earth System Sciences*, 15, 223–239, doi:10.5194/hess-15-223-2011.

Barsi, J., Hook, S. J., Schott, J. R., Raqueno, N. G., & Markham, B. L. (2007). Landsat-5 thematic mapper thermal band calibration update. *IEEE Geoscience and Remote Sensing Letters*, 4(4), 552-555, doi:10.1109/LGRS.2007.896322.

Bechtel, B., Zakšek, K., & Hoshyaripour, G. (2012). Downscaling land surface temperature in an urban area: A case study for Hamburg, Germany. *Remote Sensing*, 4, 3184-3200, doi:10.3390/rs4103184.

Bindhu, V. M., Narasimhan, B., & Sudheer, K. P. (2013). Development and verification of a non-linear disaggregation method (NL-DisTrad) to downscale MODIS land surface temperature to the spatial scale of Landsat thermal data to estimate evapotranspiration. *Remote Sensing of Environment*, 135, 118-129, doi:10.1016/j.rse.2013.03.023.

Bisht, G., Venturini, V., Islam, S., & Jiang, L. (2005). Estimation of the net radiation using MODIS (Moderate Resolution Imaging Spectroradiometer) data for clear sky days. *Remote sensing of Environment*, *97*, 52–67, doi:10.1016/j.rse.2005.03.014.

Chander, G., Markham, B. L., & Helder, D. L. (2009). Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote sensing of Environment*, 113, 893-903, doi:10.1016/j.rse.2005.03.014.

Chen, Y., Zhan, W., Quan, J., Zhou, J., Zhu, X., & Sun, H. (2014). Disaggregation of remotely sensed land surface temperature: a generalized paradigm. *IEEE Transactions on Geoscience and Remote Sensing*, *52*, 5952-5965, doi:10.1109/TGRS.2013.2294031.

Despini, F., Teggi, S., & Baraldi, A. (2014). Methods and metrics for the assessment of Pan-sharpening algorithms. *In, SPIE Remote Sensing (pp. 924403-924403-924414): International Society for Optics and Photonics*, doi:10.1117/12.2067316.

Dominguez, A., Kleissl, J., Luvall, J. C., & Rickman, D. L. (2011). High-resolution urban thermal sharpener (HUTS). *Remote Sensing of Environment*, 115(7), 1772-1780, doi:10.1016/j.rse.2011.03.008.

Gao, L., Zhan, W., Huang, F., Quan, J., Lu, X., Wang, F., Ju, W., & Zhou, J. (2017). Localization or globalization? Determination of the optimal regression window for disaggregation of land surface temperature. *IEEE Transactions on Geoscience and Remote Sensing*, 55, 477-490, doi: 10.1109/TGRS.2016.2608987.

Gevaert, C. M., & García-Haro, F. J. (2015). A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. *Remote Sensing of Environment*, *156*, 34-44, doi:10.1016/j.rse.2014.09.012.

Ghosh, A., & Joshi, P. K. (2014). Hyperspectral imagery for disaggregation of land surface temperature with selected regression algorithms over different land use land cover scenes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 96, 76-93, doi:10.1016/j.isprsjprs.2014.07.003.

Gillespie, A., Rokugawa, S., Matsunaga, T., Cothern, J. S., Hook, S., & Kahle, A. B. (1998). A temperature and emissivity separation algorithm for Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) images. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4), 1113-1126, doi:10.1109/36.700995.

Göttsche, F. M., & Olesen, F. S. (2009). Modelling the effect of optical thickness on diurnal cycles of land surface temperature. *Remote Sensing of Environment*, 113(11), 2306-2316, doi:10.1016/j.rse.2009.06.006.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques. (3th ed.). Elsevier.

Jiménez-Muñoz, J. C., & Sobrino, J. A. (2003). A generalized single-channel method for retrieving land surface temperature from remote sensing data. *Journal of Geophysical Research: Atmospheres*, *108*, ACL 2-1, doi:10.1029/2003JD003480.

Jiménez-Muñoz, J. C., Mattar, C., Sobrino, J. A., & Malhi, Y. (2016). Digital thermal monitoring of the Amazon forest: an intercomparison of satellite and reanalysis products. *International Journal of Digital Earth*, 9(5), 477-498, doi:10.1080/17538947.2015.1056559.

Keramitsoglou, I., Kiranoudis, C. T., & Weng, Q. (2013). Downscaling geostationary land surface temperature imagery for urban analysis. *IEEE Geoscience and Remote Sensing Letters*, 10, 1253-1257, doi:10.1109/LGRS.2013.2257668.

Kustas, W. P., Norman, J. M., Anderson, M. C., & French, A. N. (2003). Estimating subpixel surface temperatures and energy fluxes from the vegetation index–radiometric temperature relationship. *Remote sensing of Environment, 85*, 429-440, doi:10.1016/S0034-4257(03)00036-1.

Lagouarde, J.-P., Bach, M., Sobrino, J. A., Boulet, G., Briottet, X., Cherchali, S., Coudert, B., Dadou, I., Dedieu, G., & Gamet, P. (2013). The MISTIGRI Thermal Infrared project: scientific objectives and mission specifications. *International Journal of Remote Sensing, 34*, 3437-3466, doi:10.1080/01431161.2012.716921.

Lagouarde, J.-P., Hénon, A., Kurz, B., Moreau, P., Irvine, M., Voogt, J., & Mestayer, P. (2010). Modelling daytime thermal infrared directional anisotropy over Toulouse city centre. *Remote sensing of Environment, 114*, 87-105, doi:10.1016/j.rse.2009.08.012.

Liang, S. (2005). *Quantitative remote sensing of land surfaces*: John Wiley & Sons.

Li, Z.-L., Tang, B.-H., Wu, H., Ren, H., Yan, G., Wan, Z., Trigo, I. F., & Sobrino, J. A. (2013). Satellite-derived land surface temperature: Current status and perspectives. *Remote sensing of Environment, 131*, 14-37, doi:10.1016/j.rse.2012.12.008.

Liu, D., & Zhu, X. (2012). An enhanced physical method for downscaling thermal infrared radiance. *IEEE Geoscience and Remote Sensing Letters*, 9(4), 690-694, doi:10.1109/LGRS.2011.2178814.

Liu, J., & Moore, J. M. (1998). Pixel block intensity modulation: adding spatial detail to TM band 6 thermal imagery. *International Journal of Remote Sensing, 19*, 2477-2491, doi:10.1080/014311698214578.

McFeeters, S. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing, 17*, 1425-1432, doi:10.1080/01431169608948714.

Mechri, R., Ottlé, C., Pannekoucke, O., & Kallel, A. (2014). Genetic particle filter application to land surface temperature downscaling. *Journal of Geophysical Research: Atmospheres*, 119, 2131-2146, doi:10.1002/2013JD020354.

Merlin, O., Duchemin, B., Hagolle, O., Jacob, F., Coudert, B., Chehbouni, G., Dedieu, G., Garatuza, J., & Kerr, Y. (2010). Disaggregation of MODIS surface temperature over an agricultural area using a time series of Formosat-2 images. *Remote sensing of Environment*, 114, 2500-2512, doi:10.1016/j.rse.2010.05.025.

Moosavi, V., Talebi, A., Mokhtari, M. H., Shamsi, S. R. F., & Niazi, Y. (2015). A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature. *Remote sensing of Environment*, 169, 243-254, doi:10.1016/j.rse.2015.08.015.

Mukherjee, S., Joshi, P., & Garg, R. (2014). A comparison of different regression models for downscaling Landsat and MODIS land surface temperature images over heterogeneous landscape. *Advances in Space Research, 54*, 655-669, doi:10.1016/j.asr.2014.04.013.

Nichol, J. (2009). An emissivity modulation method for spatial enhancement of thermal satellite images in urban heat island analysis. *Photogrammetric Engineering & Remote Sensing*, *75*, 547-556, doi:http://dx.doi.org/10.14358/PERS.75.5.547.

Nishii, R., Kusanobu, S., & Tanaka, S. (1996). Enhancement of low spatial resolution image based on high resolution-bands. *IEEE Transactions on Geoscience and Remote Sensing*, *34*, 1151-1158, doi:10.1109/36.536531.

Pardo-Igúzquiza, E., Chica-Olmo, M., & Atkinson, P. M. (2006). Downscaling cokriging for image sharpening. *Remote sensing of Environment*, *102*, 86-98, doi:10.1016/j.rse.2006.02.014.

Pohl, C., & Van Genderen, J. L. (1998). Review article multisensor image fusion in remote sensing: concepts, methods and applications. International Journal of Remote Sensing, 19(5), 823-854, doi:10.1080/014311698215748.

Qin, Z.-H., Karnieli, A., & Berliner, P. (2001). A mono-window algorithm for retrieving land surface temperature from Landsat TM data and its application to the Israel-Egypt border region. *International Journal of Remote Sensing*, *22*, 3719-3746, doi:10.1080/01431160010006971.

Roberts, D. A., Quattrochi, D. A., Hulley, G. C., Hook, S. J., & Green, R. O. (2012). Synergies between VSWIR and TIR data for the urban environment: An evaluation of the potential for the Hyperspectral Infrared Imager (HyspIRI) Decadal Survey mission. *Remote sensing of Environment*, *117*, 83-101, doi:10.1016/j.rse.2011.07.021.

Rodriguez-Galiano, V., Pardo-Iguzquiza, E., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2012). Downscaling Landsat 7 ETM+ thermal imagery using land surface temperature and NDVI images. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 515-527, doi:10.1016/j.jag.2011.10.002.

Sandholt, I., Rasmussen, K., & Andersen, J. (2002). A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status. *Remote sensing of Environment, 79*, 213-224, doi:10.1016/S0034-4257(01)00274-7.

Sobrino, J. A., Jiménez-Muñoz, J. C., El-Kharraz, J., Gómez, M., Romaguera, M., & Soria, G. (2004). Single-channel and two-channel methods for land surface temperature retrieval from DAIS data and its application to the Barrax site. *International Journal of Remote Sensing*, 25(1), 215-230, doi:10.1080/0143116031000115210.

Sobrino, J. A., Gómez, M., Jiménez-Muñoz, J. C., & Olioso, A. (2007). Application of a simple algorithm to estimate daily evapotranspiration from NOAA–AVHRR images for the Iberian Peninsula. *Remote sensing of Environment*, 110(2), 139-148, doi:10.1016/j.rse.2007.02.017.

Sobrino, J. A, Oltra-Carrió, R., Sòria, G., Bianchi, R., & Paganini, M. (2012). Impact of spatial resolution and satellite overpass time on evaluation of the surface urban heat island effects. *Remote Sensing of Environment, 117*, 50-56, doi:10.1016/j.rse.2011.04.042.

Stathopoulou, M., & Cartalis, C. (2009). Downscaling AVHRR land surface temperatures for improved surface urban heat island intensity estimation. *Remote sensing of Environment*, *113*, 2592-2605, doi:10.1016/j.rse.2009.07.017.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106, 7183-7192.

Teggi, S. (2012). A technique for spatial sharpening of thermal imagery of coastal waters and of watercourses. *International journal of remote sensing*, 33(10), 3063-3089, doi:10.1080/01431161.2011.627888.

Teggi, S., & Despini, F. (2014). Estimation of subpixel MODIS water temperature near coastlines using the SWTI algorithm. *Remote sensing of Environment*, 142, 122-130, doi:10.1016/j.rse.2013.11.011.

Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., Restaino, R., & Wald, L. (2014). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, *53*, 2565 - 2586, doi:10.1109/TGRS.2014.2361734.

Wald, L., Ranchin, T., & Mangolini, M. (1997). Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing, 63, 691-699,* doi:https://hal.archives-ouvertes.fr/hal-00365304.

Wang, Z., & Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9, 81-84, doi:10.1109/97.995823.

Wang, F., Qin, Z., Li, W., Song, C., Karnieli, A., & Zhao, S. (2014). An efficient approach for pixel decomposition to increase the spatial resolution of land surface temperature images from MODIS thermal infrared band data. *Sensors*, 15, 304-330, doi:10.3390/s150100304.

Weng, Q., Fu, P., & Gao, F. (2014). Generating daily land surface temperature at Landsat resolution by fusing Landsat and MODIS data. *Remote Sensing of Environment*, *145*, 55-67, doi:10.1016/j.rse.2014.02.003.

Wu, P., Shen, H., Zhang, L., & Göttsche, F.-M. (2015). Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature.

Remote sensing of Environment, 156, 169-181, doi:10.1016/j.rse.2014.09.013.

Zakšek, K., & Oštir, K. (2012). Downscaling land surface temperature for urban heat island diurnal cycle analysis. *Remote sensing of Environment*, 117, 114-124, doi:10.1016/j.rse.2011.05.027.

Zhan, W., Chen, Y., Zhou, J., Wang, J., Liu, W., Voogt, J., Zhu, X., Quan, J., & Li, J. (2013). Disaggregation of remotely sensed land surface temperature: Literature survey, taxonomy, issues, and caveats. *Remote sensing of Environment, 131*, 119-139, doi:10.1016/j.rse.2012.12.014.

Zhan, W., Huang, F., Quan, J., Zhu, X., Gao, L., Zhou, J., & Ju, W. (2016). Disaggregation of remotely sensed land surface temperature: A new dynamic methodology. *Journal of Geophysical Research: Atmospheres*, 121(18), 10538–10554, doi: 10.1002/2016JD024891.

Zhan, W., Chen, Y., Zhou, J., Li, J., & Liu, W. (2011). Sharpening thermal imageries: A generalized theoretical framework from an assimilation perspective. *IEEE Transactions on Geoscience and Remote Sensing*, 49(2), 773-789, doi:10.1109/TGRS.2010.2060342.

Zhou, J., Chen, Y., Zhang, X., & Zhan, W. (2013). Modelling the diurnal variations of urban heat islands with multi-source satellite data. *International Journal of Remote Sensing, 34*, 7568-7588, doi:10.1080/01431161.2013.821576.

Zhou, J., Li, M., Liu, S., Jia, Z., & Ma, Y. (2015). Validation and performance evaluations of methods for estimating land surface temperatures from ASTER Data in the middle reach of the Heihe River Basin, Northwest China. *Remote Sensing*, 7(6), 7126-7156, doi:10.3390/rs70607126.

Zhou, J., Liu, S., Li, M., Zhan, W., Xu, Z., & Xu, T. (2016). Quantification of the Scale Effect in Downscaling Remotely Sensed Land Surface Temperature.

Remote Sensing, 8, 975, doi:10.3390/rs8120975.