Multi-objective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface radiometric temperature retrievals

W. T. Crow

Hydrology and Remote Sensing Laboratory, USDA ARS, Beltsville, MD

E. F. Wood and M. Pan

Depart. of Civil and Environmental Engineering, Princeton University,

Princeton, NJ

W.T Crow, Hydrology and Remote Sensing Laboratory, USDA ARS, Rm. 104, Blg. 007, BARC-W, Beltsville, MD 20705, USA. (wcrow@hydrolab.arsusda.gov)

E.F. Wood and M.Pan, Environmental Engineering and Water Resources Program, Depart. of Civil and Environmental Engineering, Princeton University, Princeton, NJ 08544, USA.

Abstract.

Physically-based models of surface water and energy balance processes typically require a large number of soil and vegetation parameters as inputs. Accurate specification of these parameters is often difficult without resorting to calibration of model predictions against independent observations. Along with streamflow observations from gauging stations, spaceborne surface radiometric temperature retrievals offer the only independent observation of land surface model output commonly available at regional spatial scales (i.e. $> 50^2$ km^2). This analysis examines the potential benefits of incorporating spaceborne radiometric surface temperature retrievals and streamflow observations in a multi-objective calibration framework to accurately constrain regional-scale model evapotranspiration predictions. Results for the VIC (Variable Infiltration Capacity) model over the Southern Great Plains of the United States suggest that multi-objective model calibration against radiometric skin temperatures and steamflow observations can reduce error in model monthly evapotranspiration predictions by up to 20% relative to single-objective model calibration against streamflow alone.

1. Introduction

Due to their increasing complexity, ambiguities surrounding parameter selection have emerged as a critical source of error for land surface model predictions of water and energy fluxes [*Franks and Beven*, 1997; *Gupta et al.*, 1999; *Houser et al.*, 2001]. These difficulties are often compounded by the need to apply land surface schemes over relatively coarse-grid cells ($>$ 10 km) for continental- to global-scale applications.

For large-scale applications, obtaining an adequate spatial representation of parameter values through direct observation is almost always impractical. The most common alternative for large-scale simulations is the use of land surface classifications and lookup tables populated with parameter values taken from the literature. Soil hydraulic parameters for distributed hydrological modeling, for instance, are typically derived from empirical relationships between soil texture and hydraulic properties and maps of soil textural classification for a particular region. The significant amount of unexplained variability in these empirical relationships - combined with the spatial inadequacies of the soil texture maps themselves [*Zhu*, 2000] - make the approach approximate at best.

Parameter selection difficulties are often compounded by the presence of land surface heterogeneity at spatial scales finer than the proscribed model grid size. Some quantities such as surface albedo are measurable remotely, have a relatively clear physical meaning at the regional-scale, and are linearly related to model flux predictions. Consequently, effective grid-scale values can be obtained through simple averaging of observable sub-grid-scale heterogeneity. However, many parameters lack a clear physical definition at the grid-cell in the presence of sub-grid-scale heterogeneity and require more complex aggregation strategies (e.g. surface roughness lengths [*Klassen and Claussen*, 1995] or stomatal

resistance [*Blyth et al.*, 1993]). Other parameters are simply difficult to accurately estimate or measure over large spatial scales (e.g. vegetation rooting depths).

As a consequence, it appears increasingly likely that even "physically-based" land surface models will require some calibration against independent observations to accurately represent land surface state variables and fluxes [*Franks and Beven*, 1997]. A central conclusion of the PILPS 2-e study over the United States Southern Great Plains was that land surface schemes that incorporated some type of calibration against observations outperformed those that did not [*Lettenmaier et al.*, 1996]. Developing strategies for transferring calibrated parameters to nearby - or geomorphologically similar basins - is also a critical component of NOAA's Model Parameters Estimation Experiment (MOPEX) study (online at: http://www.nws.noaa.gov/oh/mopex/).

Recent advances in land surface model calibration have focused on examining the issue within a framework that recognizes the inherently multi-objective nature of land surface models [*Gupta et al.*, 1998; *Yapo et al.*, 1998; *Gupta et al.*, 1999; *Houser et al.*, 2001]. An underlying principle of this framework is the recognition that a single parameter set is unlikely to optimize all model outputs. *Gupta et al.* [1999], for instance, demonstrates that the calibration of land surface model predictions against observations of a single output (e.g. evapotranspiration) does not accurately constrain other predictions (e.g. skin temperature or soil moisture) accurately. Instead multi-objective calibration incorporating at least one surface energy flux and one surface state variable is required to ensure adequate calibration of all model predictions. With the exception of *Houser et al.* [2001], most previous multi-objective calibration work has focused on the patch-scale (10-100

m) application of models and not addressed the calibration of land surface models over coarser regional-scales.

Evapotranspiration is typically both the largest component of the terrestrial water balance and the most difficult to measure directly. Consequently, it's accurate prediction is often a central goal of large-scale modeling efforts. Land surface models have generally focused on calibration against streamflow measurements to constrain evapotranspiration predictions. Neglecting variations in interannual soil water storage, calibration against streamflow ensures accurate prediction of annual evapotranspiration, but does not constrain seasonal partitioning between evapotranspiration and soil water storage. Because of its spatial attributes, remote sensing observations have attracted interest as a source of alternative (or complementary) calibration data for land surface models [*Camillo et al.*, 1986; *Burke et al.*, 1997]. Its close conceptual link to terms of the surface energy balance and widespread availability suggests that spaceborne surface radiometric temperature (T_s) retrievals in particular have some calibrational value for evapotranspiration predictions. Recent work has examined the value of remote T_s retrievals as a source of validation data for land surface models [*Jin et al.*, 1997; *Rhoads et al.*, 2001] and demonstrated the utility of T_s observations within the context of land surface data assimilation systems [*Lakshmi*, 2000; *Boni et al.*, 2001]. The goal of this paper is to clarify the potential for improving large-scale ($> 50^2$-km^2) model predictions of evapotranspiration through calibration of a land surface model using remote surface radiometric temperature retrievals. Most centrally, it focuses on whether multi-objective model calibrations involving streamflow and radiometric surface temperature retrievals can outperform traditional single-objective model calibration using streamflow alone.

2. Multi-objective Calibration

Land surface models typically predict a range of land surface state (e.g. soil moisture and soil temperature) and flux (e.g. infiltration, runoff and evapotranspiration) variables. Consequently, it is often advantageous to think of their calibration within a multi-objective framework. Multi-objective calibration is based on the minimization of a set of model performance criteria where each criterion corresponds to a different land surface variable. In general, errors in forcing data, measurement uncertainties, and shortcomings in the physics of the models themselves will prevent a single set of parameters from optimizing all types of land surface model predictions. Instead, multi-objective optimization leads to a set of solutions which captures optimal trade-offs between various types of model predictions.

Figure 2 (adapted from *Houser et al.* [2001]) presents a simple one-parameter (θ) example of multi-objective calibration. In this example, two observations (or objectives) are matched with model predictions. The goodness-of-fit criteria between observations and model predictions are given by $f_1(\theta)$ and $f_2(\theta)$. Figure 2a demonstrates a case where the parameter value required to minimize f_1 (labeled θ_1) provides a poor result for f_2 and vise versa for the minimizing parameter choice for f_2 (labeled θ_2). Between θ_1 and θ_2 in Figure 2a there exists a set of parameters for which it is possible to improve fit to one objective through adjustments to θ only at the expense of fit to another. Following *Gupta et al.* [1998], this set of parameters solutions will be referred to as the Pareto set. Figure 2b shows the same case mapped in f_1 and f_2 fitness space. The curve shows $f_1(\theta)$ and $f_2(\theta)$ values associated with the adjustment of θ within it's feasible range. Bold portions of the trajectory represent results for adjustments of θ within the Pareto set ($\theta_1 < \theta < \theta_2$).

Note that the member of the Pareto set that minimizes $f_1 + f_2$ (A on Figure 2b) is distinct from parameter values that minimize either f_1 (B on Figure 2b) or f_2 (C on Figure 2b) individually.

One method for approximating the Pareto set is to linearly collapse a vector containing multiple fitness criteria ($\mathbf{F}$) into a single scalar criteria (G):

$$G = \mathbf{W} \cdot \mathbf{F} \tag{1}$$

and then minimize G for a finite range of weighting choices in $\mathbf{W}$ [*Gupta et al.*, 1998; *Bastidas et al.*, 1999]. For the one-parameter/two-objective example in Figure 2, $\mathbf{F} = \{f_1, f_2\}$ and $\mathbf{W} = \{W_1, W_2\}$ where W_1 and W_2 are positive real numbers that sum to one. Point A on Figure 2b corresponds to a weighting choice of $\mathbf{W} = \{0.5, 0.5\}$, point B to $\mathbf{W} = \{1, 0\}$, and point C to $\mathbf{W} = \{0, 1\}$. Since it requires a separate optimization calculation for each weighting choice, (1) can be an inefficient method for obtaining the complete Pareto set [*Gupta et al.*, 1998]. Potentially more efficient methods include the multi-objective complex evolution (MOCOM) algorithm introduced by *Yapo et al.* [1998], and optimization algorithms based on genetic analogies [*Kuczera*, 1997; *Seibert*, 2000].

All calibration techniques require the specification of a fitness criterion f to quantify the goodness-of-fit between model predictions (Z) and observations (X). For a series of n observations, the most common fitness criteria is the root mean squares error ($RMSE$):

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1..n} (Z_t - X_t)^2}. \tag{2}$$

Dividing a RMSE criteria by the standard deviation of the observations ($RMSE/\sigma_X$) gives the normalized root mean squared error ($NRMSE$). The quantity $1 - NRMSE^2$ is commonly called the Nash-Sutcliffe coefficient of efficiency for predictions.

Following notation presented by *Gupta et al.* [1999], the use of bracket { } notation surrounding model output variables is used here to signify calibration of those variables. Results for single-objective calibration (e.g. $\{Z_1\}$) consist of a single optimal parameter set. Multi-objective calibration ($\{Z_1, ..., Z_n\}$ where $n > 1$) leads to a Pareto set of parameter solutions, each reflecting various choices for the weighting vector $\mathbf{W}$ in (1). Calibration results can also be evaluated in terms of their goodness-of-fit to observations. This is possible for cases of direct calibration (e.g. NRMSE for Z_1 given $\{Z_1\}$) and cross-calibration where goodness-of-fit is evaluated for an observation type which is not calibrated against (e.g. NRMSE for Z_3 given $\{Z_1, Z_2\}$).

3. Procedure

The analysis was based on multi-objective calibration of the Variable Infiltration Capacity (VIC) land surface model [*Liang et al.*, 1994; *Cherkauer et al.*, 1999] using a combination of streamflow observations and remote skin temperature retrievals within the model domain shown in Figure 1. Flux towers measurements were used to assess the accuracy of subsequent VIC evapotranspiration predictions. Sections 3.1 and 3.2 detail the acquisition and processing of observational data sets. Sections 3.3 and 3.4 discuss the VIC model and its application to the study domain. Section 3.5 gives specifics concerning the calibration of VIC against observations.

3.1. Water Flux Measurements

The geographic domain for this analysis was taken to be the area of overlap between the Atmospheric Radiation Measurement Cloud and Radiation Testbed (ARM-CART) Southern Great Plains (SGP) site and the state of Oklahoma (see Figure 1). The domain, and its immediate vicinity, contains 17 ARM-CART surface energy flux towers - 5 eddy correlation (ECOR) and 12 energy balance/bowen ratio (EBBR) - and 59 Oklahoma Mesonet stations. In addition, streamflow measurements were obtained from United States Geological Survey (USGS) gages at 18 locations in and around the domain.

Despite a relatively high density of measurement sites, obtaining domain-averaged monthly values for streamflow and evapotranspiration is challenging. Land cover variability within the domain, chiefly the contrast between grassland and winter wheat fields, can lead to large horizontal heterogeneity in growing season surface energy fluxes. The ongoing establishment of ECOR towers on winter wheat fields since 1997 has improved the regional representativeness of ARM-CART flux tower measurements. Nevertheless, ambiguity remains in the aggregation of local evapotranspiration measurements over heterogeneous land cover up to a domain-scale average. For this analysis, monthly flux tower observations were spatially interpolated onto a 10-km grid using r^{-2} weighting, where r is the distance from a given 10-km grid-cell to a given tower. This imagery was then aggregated up to the domain-scale shown in Figure 1.

In addition, adequate sampling of streamflow from the central and western portion of the domain required moving beyond available natural (i.e. unregulated) basins in the region. While every attempt was made to minimize the impact of diversion and regulation, it should be noted that a number of the basins contain flood control structures and/or

non-negligible levels of urban diversion. Domain-averaged, monthly streamflow values were obtained by interpolating streamflow measurements onto a 10-km grid. A simple nearest-neighbor algorithm was used since more complex interpolation procedures (e.g. cubic smoothing, block kriging, and r^{-2} interpolation) yielded very similar results.

Despite the uncertainties inherent in these measurements, domain-averaged values of precipitation, streamflow, and evapotranspiration balance over long time scales. Within the two-year study period (October 1, 1997 to September 30, 1999), total observed precipitation (1972 mm) matched the sum of total observed streamflow (434 mm) and total observed evaporation (1546 mm) to within 8 mm ($< 0.5\%$ of total precipitation). Since measurements of streamflow, evapotranspiration, and precipitation are independently acquired, this balance implies that no long-term biases are present in the observations.

At a finer time scale, month by month evaluations of water flux measurements are possible using soil moisture observations from the 59 Oklahoma Mesonet stations equipped with soil moisture probes within the study domain. Figure 3a shows monthly observed means for the site for the two-year study period and Figure 3b the comparison between spatially averaged monthly changes in 1-m soil moisture measurements (ΔS) at Mesonet sites and the residual of independent precipitation (P), streamflow (Q), and evapotranspiration (E_T) measurements ($P - E_T - Q$). An integrated 1-m soil moisture measurement was derived from weighted averaging of 10-, 25-, 60-, and 75-cm measurements at Mesonet sites. Large differences in Figure 3b underscore the difficulty of closing the domain-scale water budget with observations alone. Some of the difficulty may lie in the exclusive use of 1-m soil moisture measurements to characterize temporal changes in terrestrial water storage. For example, one clear difference in Figure 3b is the larger seasonality in flux

residuals ($P - E_T - Q$) relative to monthly 1-m soil moisture variations. A portion of this difference can be accounted for by the movement of water across the bottom of the 1-m zone sampled by the Mesonet measurements since recharge during the spring and the upward movement of water during the late summer and early fall will dampen the seasonality of soil moisture measurements limited to the root-zone.

3.2. Surface Radiometric Temperature Observations

Surface radiometric temperature observations were taken from both the TIROS Operational Environmental Sounder (TOVS) instrument aboard the NOAA-14 satellite and Geostationary Environmental Satellite (GOES) observations. TOVS surface radiometric temperature estimations were extracted from the TOVS Pathfinder Path A data set [*Susskind et al.*, 1997]. Atmospheric corrections for the Path A data were performed using temperature and moisture profiles derived from the Goddard Earth Observing System - Data Assimilation System (GEOS-DAS) general circulation model. Retrievals are possible for cloud coverage up to 80% and retrievals should be considered a spatial average of cloud-free areas within the scene. When launched in early 1995, NOAA-14 had a local overpass time of 13:30. It subsequently drifted to slightly later overpass times. Consequently, TOVS afternoon surface radiometric temperature observations occurring in 1997 and 1998 were assumed to correspond to a local time of 14:30 and observations in 1999 to a local time of 15:30. At 96° W, local solar times of 14:30 and 15:30 correspond to GMT times of 21 and 22 respectively. TOVS surface radiometric temperature retrievals are available on a 1° x 1° lat/long resolution grid, and a rectangle containing degree boxes between 99° and 96° W and 37° and 35° N was selected to represent the model domain (see Figure 1). Observations within these boxes were averaged to obtain

a single domain-averaged surface radiometric temperature. Only days in which surface radiometric temperature retrievals were successful in at least two of the six-degree boxes were included in the analysis. Comparisons of TOVS T_s observations with ground-based observations typically yield root-mean-squared (RMS) differences of between 4 and 5 ^{o}K for instantaneous measurements (*Lakshmi and Susskind*, 2000). Differences in monthly-averaged T_s values tend to be near 1 to 2 ^{o}K (see e.g. *Drusch and Wood* [2001]) and comparisons over longer time periods suggest the presence of little or no retrieval bias in observations (*Lakshmi and Susskind*, 2000).

GOES observations offer the possibility of surface radiometric temperature observations at a higher temporal frequency (hourly as opposed to daily) and finer spatial resolution (2-km as opposed to 1^o). Following *Czajkowski et al.* [1998], surface temperature retrieval from GOES channel 4 and 5 observations were based on the split-window approach described in *Rhoads et al.* [2001]. GOES radiometric surface temperature products were taken from 1/8^o degree resolution data generated for the North American Land Data Assimilation System (NLDAS). Only days in which radiometric temperature retrieval was successful in at least 200 of the 640 1/8^o pixels that comprised the domain were included in the analysis. Ground-based validating of GOES T_s retrievals with the SGP ARM CART area was performed as part of the NLDAS project. Spatial averages of instantaneous GOES T_s and ground-based measurements within the entire ARM CART site yields a RMS difference of 3.1 ^{o}K during years 1998 and 1999. Averaging T_s values to monthly averages prior to comparison reduces RMS error levels by about half (*Robock et al.*, 2003).

TOVS and GOES T_s retrievals can be intercompared by considering only GOES retrievals that occur within one hour of the daily TOVS overpass time and averaging both

sets of retrievals up to a monthly time scale. Direct comparisons of monthly averaged T_s values yields a RMS difference of 3.8 °K with GOES observations biased 2.3 °K high relative to TOVS retrievals. Some of this difference is likely due to each sensor sampling different days within the month. When determinations of cloud-free scenes are made independently by each sensor, GOES tends to be more conservation and labels fewer days as being sufficiently cloud-free. Forcing monthly averages to be based on exactly the same set of observation times - determined to be cloud-free by both sensors - reduces the observed RMS difference to 2.4 °K.

3.3. VIC Model Set-up

Land surface modeling was based on the Variable Infiltration Capacity (VIC) model [*Liang et al.*, 1994; *Cherkauer et al.*, 1999]. The VIC model was designed to solve the surface water and energy balance over large grid-scales (typically $>$ 10 km) based on observations of rainfall and incoming radiation. Partitioning of rainfall into infiltration and surface runoff is controlled by a variable infiltration capacity curve which implicitly represents sub-grid scale heterogeneity in the infiltrative capacity of the land surface. Vertical water movement occurs within four discrete soil layers through diffusion and drainage processes parameterized by user specified soil hydraulic parameters. Evapotranspiration is predicted using a Penman-Monteith calculation based on observed meteorology, vegetation leaf area index (LAI), and a stomatal conductance formulation which considers the impact of soil water stress. Based on this calculation of evapotranspiration and observations of incoming radiation, the surface energy balance is numerically solved on an hourly time step by iterating on surface temperature. The model has been successfully applied

to the United States Southern Great Plains region by a number of researchers (see e.g. *Lohmann et al.* [1998] or *Abdulla et al.* [1996]) .

Hourly forcing data for VIC simulations were taken from the NLDAS retrospective forcing data set. Rainfall observations were derived from the use of NCAR Climate Prediction Center (CPC) rain gage measurements to bias correct 4-km WSR-88D Doppler radar-based precipitation estimates. Solar radiation forcing was based on the inversion of GOES visible imagery with a short-wave radiative transfer model. Required meteorological observations for VIC (e.g. wind speed, air temperature, relative humidity, and air pressure) were obtained from the NCEP Eta Data Assimilation System (EDAS) and used to estimate incoming longwave radiation. Complete processing details for the NLDAS retrospective data set can be found in *Cosgrove et al.* [2003].

Forcing data were used to drive VIC predictions on an hourly time step in full energy-balance mode. The hourly resolution of the simulations and forcing data was essential for facilitating direct comparisons between instantaneous satellite measurements of T_s and model predictions at various points along the diurnal cycle. For each land cover component (bare soil, grassland, and winter wheat), VIC derived a surface temperature estimate by iteratively solving the surface energy balance. Surface temperature estimates for each land cover type were aggregated into a single surface radiometric temperature (T_s) through weighted averaging of surface temperature to the fourth power [*Norman et al.*, 1995]:

$$T_s = [\sum_{i=1}^{3} f_i T_{s,i}^4]^{1/4} \tag{3}$$

where f_i refers to the fractional extent of each land cover type.

3.4. VIC Calibration Parameters

Table 1 lists the calibration parameters and their minimum and maximum possible values. The seven were selected both for their importance in driving evapotranspiration predictions as well as the high level of uncertainty in their specification at coarse spatial scales. With the exception of the maximum baseflow rate (ds_{max} - see Table 1), parameters for runoff and baseflow processes were taken from the VIC calibration work of *Abdulla et al.* [1996] within the Red-Arkansas River basin. Various constitutive relationships were also employed to relate parameters in Table 1 to additional VIC model parameters. Soil sand ($\%_{sand}$ - see Table 1) and clay percentages ($\%_{clay}$ - see Table 1) were converted into soil hydraulic conductivity, pore size distribution parameter, bubbling pressure, and porosity values using single-variable regression relationships presented in *Cosby et al.* [1984]. Soil quartz content was assumed equal to $\%_{sand}$ and values of critical and wilting point soil moisture were derived from soil suction curves parameterized with pore size distribution indices and bubbling pressures derived from the *Cosby et al.* [1984] regression relationships. Based on land cover classifications of the region, vegetation cover within the study domain was classified as 85% grassland and 15% winter wheat fields. Areas with summer crop and forage land cover types were lumped with the grassland classification and small areas with tree and brush cover were neglected. The surface area of grass and winter wheat roots was assumed to decay exponentially with a folding length equal to the root density decay parameter (k^{-1} - see Table 1). This exponential relationship was integrated within the three soil layers (0-15 cm, 15-45 cm, and 45-145 cm) to give the fraction of root area in each soil layer. Typical monthly LAI cycles for both grass and winter fields were taken from field measurements [*Verma and Berry*, 1999]. These observed

annual cycles were rescaled such that their annual maximum (May for winter wheat and July for grasslands) was equal to the calibrated maximum LAI parameter (LAI_{max} - see Table 1). Winter wheat fields were assumed to be fully vegetated between December and June and completely bare between July and November. In contrast, a constant fraction of the grassland fields (f_{veg} - see Table 1) was specified to be bare soil. Except for the post-harvest conversion of winter wheat fields to bare soil, roughness lengths (z_o - see Table 1) were assumed to be seasonally constant and equal for both grassland and winter wheat land covers types.

Due to computational constraints, simulations were run as a single grid-cell containing the entire study domain shown in Figure 1. The representation of sub-grid heterogeneity was limited to the statistical representation of sub-grid land cover variations (i.e. winter wheat, grassland, and bare soil surfaces) and the treatment of sub-grid infiltration capacity fundamental to the VIC modeling concept. While relatively coarse, such a grid-size is consistent with the original design specifications of the VIC model.

3.5. VIC Calibration Procedure

VIC calibration was against monthly summed (for fluxes) and averaged (for temperatures) observations during the two-year period between October 1, 1997 and September 31, 1999. All observations and model predictions were spatially averaged over the entire domain shown in Figure 1. To ensure that monthly averages of T_s retrievals were based on the same temporal support as model predictions, modeled mean monthly surface temperature values were obtained by averaging hourly model predictions only for time steps at which remote observations were available. For TOVS observations, this meant that only hourly model predictions corresponding to the single daily NOAA-14 overpass time

were considered. All other VIC T_s predictions were discarded. Automated calibration was performed with the Shuffled Complex Evolution algorithm developed at the University of Arizona (SCE-UA) [*Duan et al.*, 1992]. Following (1), G was defined to be the weighted sum of normalized root mean square errors (NRMSE - see Section 2) in VIC model predictions relative to both surface temperature (T_s) and streamflow (Q) observations:

$$G = W_{T_s}\ NRMSE_{T_s} + W_Q\ NRMSE_Q \tag{4}$$

where the weights W_{T_s} and W_Q are positive and sum to unity. Using the SCE-UA algorithm, VIC parameters were optimized within the ranges specified in Table 1 to minimize G. The optimization procedure was repeated for a range of W_{T_s} and W_Q choices, and remote T_s retrievals acquired from both TOVS and GOES, to approximate the Pareto set for multi-objective $\{Q, T_{s,\mathrm{GOES}}\}$ and $\{Q, T_{s,\mathrm{TOVS}}\}$ calibration. Members these Pareto sets were evaluated based on the accuracy of their cross-calibration E_T predictions. In addition to the automated SCE-UA calibration, a Monte Carlo calibration method based on 100,000 random samples of the parameters listed in Table 1 was employed. Each random parameter value was independently sampled from uniform distributions bounded between the feasible parameter extremes listed in Table 1 and used to initiate a VIC model simulation which was evaluated in terms of the misfit between its predictions and observations of Q and T_s. Instead of returning a single optimal parameter set, these simulations calculated T_s and Q NRMSE fitness for a large number of randomly selected parameter sets. Using (4), these parameter sets were then ranked for a range of W_{T_s} and W_Q combinations.

4. Results

Sections 4.1 and 4.2 compare the accuracy of VIC E_T and Q predictions derived from single-objective $\{Q\}$ and $\{T_s\}$ calibrations to E_T and Q VIC predictions for parameter sets within the Pareto set associated with multi-objective $\{Q, T_s\}$ calibration. Section 4.1 focuses on calibration using the SCE-UA algorithm. Section 4.2 examines results for the Monte Carlo method based on the random generation of a large number of parameter sets.

4.1. SCE-UA Algorithm Calibration

Figure 4a describes results for the cross-calibration of VIC E_T predictions using the SCE-UA algorithm and Q and T_s observations described in Sections 3.1 and 3.2. The abscissa shows weighting values corresponding to W_{T_s} and W_Q in (4). As a result, the figure relates cross-calibrated E_T model errors for a sample of parameter solutions within the Pareto set derived by $\{Q, T_s\}$ calibration. The goodness-of-fit criteria (i.e. **F** in (1)) is the NRMSE measure described in Section 2.

Single objective calibration results are located at the edges of the plots in Figure 4. $\{Q\}$, $\{T_{s,\mathrm{GOES}}\}$, and $\{T_{s,\mathrm{TOVS}}\}$ calibration lead to cross-calibrated E_T accuracies of 15.8 mm, 22.7 mm, and 24.3 mm, respectively. When restricted to single-objective calibration, Q measurements constrain E_T predictions more effectively than remote T_s observations. Nevertheless, since direct $\{E_T\}$ calibration leads to an E_T RMSE of 8.1 mm, the use of Q measurements as a surrogate for direct E_T calibration effectively doubles model E_T error (15.8 mm versus 7.7 mm). E_T predictions are improved by multi-objective calibration the incorporates both Q and $T_{s,\mathrm{GOES}}$ observations. Appropriate weighting of $T_{s,\mathrm{GOES}}$ and Q NRMSE values ($W_{Ts,\mathrm{GOES}} = 0.75$ and $W_Q = 0.25$) leads to a 2.8 mm reduction in

RMSE for VIC E_T predictions relative to $\{Q\}$ calibration results. This represents an 18% reduction in total model error (2.8 mm / 15.4 mm) and a 37% reduction in the fraction of model error attributable to calibration shortcomings (2.8 mm / 7.7 mm).

Previous calibration studies of VIC in the SGP region have noted that single-objective $\{Q\}$ calibration produces good E_T predictions on a average annual basis but may lead to seasonal errors [*Abdulla et al.*, 1996]. The monthly time series of VIC E_T results in Figure 5 for $\{Q\}$ calibration are consistent with results in *Abdulla et al.* [1996] with the exception that VIC overestimation of E_T occurs in the spring (as opposed to mid-winter in *Abdulla et al.* [1996]) and VIC underestimation in centered on late-summer (as opposed to fall). Seasonal E_T biases occur because Q observations do not contain any information concerning the partitioning of $P-Q$ into E_T and changes in soil water storage. However, comparison of VIC E_T results for $\{Q\}$ calibration to the best $\{Q, T_{s,\mathrm{GOES}}\}$ calibration result (i.e. $W_{Ts,\mathrm{GOES}} = 0.75$ and $W_Q = 0.25$) demonstrates that multi-objective calibration incorporating T_s observations makes about a 20% reduction in RMSE levels associated with $\{Q\}$ calibration. Since T_s levels generally rise as E_T falls (and vise versa), any seasonal E_T bias should be associated with a opposing bias in VIC T_s predictions. Figures 5 suggests that this T_s bias is remotely detectable using GOES and that a partial correction of seasonal VIC E_T predictions is a by-product of calibrating model parameters to minimize both Q and T_s error.

Reducing the calibration weighting for Q observations eventually lowers the accuracy of VIC Q predictions. However, Figure 4b demonstrates that decreasing Q weighting relative to T_s observations significantly degrades Q predictions only for very low Q weights. At the minimum seen for GOES results in Figure 4a ($W_{Ts,\mathrm{GOES}} = 0.75$ and $W_Q = 0.25$), Q

RMSE for $\{Q, T_{s,\mathrm{GOES}}\}$ calibration is only 0.3 mm greater than that observed for single-objective $\{Q\}$ calibration. This suggests that the improved cross-calibration of E_T does not come at the expense of Q accuracy and multi-objective calibration improves VIC's overall representation of the terrestrial water cycle.

The cross-calibrational value of $T_{s,\mathrm{GOES}}$ observations for VIC E_T predictions demonstrated in Figure 4a is not reflected in $\{Q, T_{s,\mathrm{TOVS}}\}$ calibration results. $\{Q, T_{s,\mathrm{TOVS}}\}$ calibration produces consistently inferior E_T results relative to single-objective $\{Q\}$ calibration. Key parameter contrasts underscoring the difference in TOVS and GOES results in Figure 4 are discussed in Section 4.2.2.

4.2. Monte Carlo Results

Results in Figures 4 and 5 describe cross-calibration results for optimal fits to various fitness criteria but give no consideration to the impact of measurement uncertainty. For instance, Figure 6 plots a point cloud for VIC Q and E_T RMSE responses derived from 100,000 randomly selected parameters values (see Section 3.5). The dotted vertical line represents an arbitrarily chosen threshold that defines a set of points considered statistically indistinguishable from the best Q fit. Note that indistinguishable Q responses are associated with a wide range of E_T responses. Consequently, two parameter sets can give essentially identical results for one objective but vastly different cross-calibration results for a another objective not accounted for during calibration. Since results in Figures 4 and 5 reflect only the single best fit to observations of T_s and Q, it is unclear whether cross-calibrated E_T results are robust to reasonable levels of observational error.

4.2.1. E_T Cross-calibration Results.

Figures 6 and 7 examine the cross-calibration advantages of T_s measurements in a more robust framework. Both figures are based on the random generation of 100,000 model parameters sets from Table 1 and the random sampling procedure discussed in Section 3.4. Figure 6b plots E_T and $T_{s,\mathrm{GOES}}$ RMSE results for the "indistinguishable" Q responses that fall to the left of the line in Figure 6a. A statistically significant trend is observed which suggests that $T_{s,\mathrm{GOES}}$ observations have the potential to sort previously indistinguishable parameter sets into ones exhibiting good E_T responses from those that do not. The trend is a general property for all parameters sets exhibiting good Q fits and cannot be ascribed to an anomalous result for any single parameter set.

Figure 7 is analogous to Figure 4 except that "optimal" parameter solutions for each of the various weighting combinations are defined to be the 1% of the randomly selected parameter sets which give the lowest G value in (4) for various choices of W_{T_s} and W_Q. This set of parameter solutions can be run through VIC to derive a corresponding distribution of E_T accuracies. The median, 25th, and 75th quartiles of E_T RMSE distributions are plotted in Figure 7. Consequently, E_T cross-calibration results are based not just on the single best result, but for an entire set of near-optimal, and effectively indistinguishable, parameter solutions. Several minor differences exist between results in Figures 4 and 7. For example, $\{T_{s,\mathrm{GOES}}\}$ and $\{T_{s,\mathrm{TOVS}}\}$ calibration is associated with lower E_T RMSE relative to SCE-UA results in Figure 4, and the slight local minimum in TOVS results seen near $W_Q = 0.35$ in Figure 4 is not reflected in Figure 7. Despite these discrepancies, both figures demonstrate the same basic qualitative trends. Therefore, we conclude that the cross-calibration advantages associated with the incorporation of $T_{s,\mathrm{GOES}}$ observations in Figures 4 and 5 are robust in nature and not an anomaly associated with the single,

potentially unrepresentative, optimal parameter set selected by the SCE-UA algorithm. As in Figure 4, no comparable advantages are observed for $T_{s,\mathrm{TOVS}}$ observations.

The choice of 1% threshold defining "indistinguishable" is somewhat arbitrary, but results do not qualitatively changes for other thresholds. In addition to results presented in Figure 7, a series of smaller Monte Carlo simulations (10,000 members) were run to gauge the impact of raising the threshold of cloud-free T_s retrievals required to define a single domain-averaged T_s retrieval (see Section 3.2). Doubling the required threshold (from 2 out of 6 pixels to 4 out of 6 pixels for TOVS and from 200 out of 640 pixels to 400 out of 640 pixels for GOES) produced results essentially identical to those shown in Figure 7. In fact, results are stable up to the point where thresholds become too stringent to allow for more than a handful of usable retrievals in some months (6 out of 6 pixels for TOVS and ∼550 out of 640 pixels for GOES).

4.2.2. Parameter Results.

Monte Carlo results can also be evaluated in terms of the model parameters found to facilitate a good fit between VIC model predictions and various observation types. Instead of returning a single "calibrated" parameter set, these results examine the range of parameter values that can be associated with good fits to various observation types. Figure 8 displays box and whisker plots for the 1% of 100,000 randomly selected parameter sets with the lowest G in (4) for the relative weighting of Q and $T_{s,GOES}$ at the E_T RMSE minimum in Figure 7 ($W_Q = 0.20$ and $W_{T_{s,\mathrm{GOES}}} = 0.80$). Also plotted are the 1% of randomly selected parameter sets exhibiting the best single-objective fit to E_T and Q observations. The ordinate range for plots in Figure 8 corresponds to the maximum and minimum parameter values listed in Table 1.

A clear tendency in Figure 8 is for calibration against different observation types to return substantially different parameter values. This is a common phenomenon in hydrologic modeling which reflects errors in observations as well as inherent structural shortcomings in a model's approximation of reality (*Gupta et al.*, 1998). Isolating the source, and impact, of these parameter differences is frequently an ambiguous process. However, a key difference between parameters associated with $\{Q\}$ and $\{Q, T_{s,\mathrm{GOES}}\}$ calibration appears to be the tendency for multi-objective $\{Q, T_{s,\mathrm{GOES}}\}$ calibration to predict lower LAI_{max} values and a sandier soil texture. High LAI values for $\{Q\}$ calibration lead to excessively high springtime E_T predictions and, consequently, more severe water limitations on E_T later in the growing season. These water limitations are exacerbated by excessively clayey soils whose high wilting point further restricts late summer E_T magnitudes. Seasonal biases in VIC E_T predictions manifest themselves as excessively cool (warm) T_s VIC predictions in spring (summer). The critical contribution of the $T_{s,\mathrm{GOES}}$ observations within multi-objective $\{Q, T_{s,\mathrm{GOES}}\}$ calibration appears to be detection of seasonal T_s errors and the positive impact on E_T predictions of correcting VIC T_s values by lowering LAI_{max} and $\%_{clay}$ values while raising $\%_{sand}$. Unlike $T_{s,\mathrm{GOES}}$ observations, multi-objective calibration incorporating TOVS T_s observations (not shown) is unable to detect seasonal biases in VIC T_s associated with poor E_T predictions. Except for extremely low weighting of Q observations, multi-objective $\{Q, T_{s,\mathrm{TOVS}}\}$ calibration yields parameters, and VIC E_T predictions, that do not differ greatly from single-objective $\{Q\}$ calibration.

Comparison of E_T model predictions derived with randomly selected parameter sets to observations demonstrates that relatively well-defined parameter values can be associated with accurate E_T predictions (see Figure 8). However, there is significantly more spread

in parameter sets exhibiting good fits to Q observations. *Beven* [1993] coined the term "equifinality" is described the phenomenon of equally accurate model predictions arising from widely varying parameter choices. Multi-objective calibration using a weighted combination of Q and T_s does little to reduce parameter equifinality relative to single objective $\{Q\}$ calibration. Moving from $\{Q\}$ to $\{Q, T_{s,\mathrm{GOES}}\}$ calibration decreases the interquartile spread of parameters for only 3 out of 7 VIC parameters in Figure 8. In addition, with the exception of LAI_{max} and $\%_{clay}$ (see discussion above), there is no trend of $\{Q, T_{s,\mathrm{GOES}}\}$ calibration results becoming more consistent with $\{E_T\}$-specified parameter ranges than $\{Q\}$ calibration. Consequently, while the incorporation of $T_{s,\mathrm{GOES}}$ into a multi-objective framework calibration allows for slightly improved E_T predictions, it does not does lead to a consistent convergence of model parameters towards those found through $\{E_T\}$ calibration nor does it significantly alleviate parameter equifinality difficulties associated with single-objective $\{Q\}$ calibration.

5. Discussion and Conclusions

Previous calibration studies using VIC have focused on obtaining baseflow, runoff, and soil parameters through single-objective calibration against streamflow observations [*Abdulla et al.*, 1996; *Nijssen et al.*, 2001]. Less emphasis has been placed on calibrating parameters related to vegetation and energy balance processes. Within the SGP region, relatively low runoff ratios ($\sim$15%) and a pronounced seasonal cycle in soil moisture (see Figure 2b or *Lohmann et al.* [1998]) suggest that comparisons to streamflow observations alone may not accurately constrain evapotranspiration predictions. This study focuses on the potential for improving the calibration of large-scale land surface models by incorporating satellite-derived surface radiometric temperature retrievals into a multi-objective

calibration framework with streamflow. The approach is evaluated over the intersection of the SGP ARM-CART site with the state of Oklahoma (see Figure 1). The site is notable for its quality and density of evapotranspiration observations.

Computational constraints played a large role in determining the methodology applied to this analysis. Comparison to instantaneous surface radiometric temperature (T_s) observations required an hourly model resolution and 17,520 time steps (24*365*2) to cover the two-year simulation period. Furthermore, construction of Figures 4 and 7 necessitated a large number ($> 50{,}000$) of model simulations. Two principle sacrifices were necessary to maintain computational feasibility. First, sub-domain scale heterogeneity in dynamic model forcings (e.g. solar insolation and precipitation) and soil properties was neglected. Second, the Pareto set for multi-objective $\{Q, T_s\}$ calibration was sampled only at several various discrete weightings of Q and T_s. The lack of a complete Pareto set retrieval limits plotting of E_T RMSE results in Figure 4 to discrete points along the abscissa. Nevertheless, the inferred curves appear smooth enough to offer some confidence that a partial representation is sufficient to capture critical trends in E_T fitness among members of the $\{Q, T_s\}$ Pareto set.

A second limitation was the accuracy of the ground-truth data sets used to evaluate and/or calibrate VIC predictions. Comparisons of model output to observations were limited to monthly averages for T_s and monthly sums for streamflow (Q) and evapotranspiration (E_T) to minimize the impact of observational error. The observational quality of TOVS T_s retrievals ($T_{s,\mathrm{TOVS}}$), for instance, has been questioned at time scales below monthly [*Drusch and Wood*, 2001], and daily or weekly comparison to E_T or Q observations would likely incorporate significant errors related to the spatial sampling of E_T

and/or the routing of Q in some of the larger basins. Consequently, comparisons at a finer time scales - while increasing the degrees of freedom and accuracy of criteria calculations - come at the cost of greater observational uncertainty. A monthly time-scale was therefore chosen as a compromise between these two considerations.

Despite these limitations, Figures 4 to 7 demonstrate a modest level of improvement associated with VIC E_T predictions when GOES T_s observations ($T_{s,\mathrm{GOES}}$) are included with streamflow observations within a multi-objective model calibration framework. This improvement is detectable in both automatically optimized SCE-UA results based only on the single best fit to observations (Figure 4) and Monte Carlo results derived from the top 1% of 100,000 random parameter choices (Figure 7). No analogous improvements are noted for TOVS T_s observations. TOVS and GOES T_s products differ in their spatial resolution, temporal frequency, and methods for correction of atmospheric effects. Additional work is required to identify which of these differences are critical to the contrast in results for TOVS and GOES T_s retrievals observed in Figures 4 and 7.

No attempt is typically made in multi-objective calibration to differentiate between various members of the Pareto set associated with a particular model and set of observations. Instead, it is assumed that all members of the set are equally appropriate and made distinguishable only by the arbitrary selection of a particular weighting vector $\mathbf{W}$ (see e.g. *Gupta et al.* [1998]). Our analysis deviates from this tendency by evaluating members of the Pareto set for $\{Q, T_{s,\mathrm{GOES}}\}$ calibration based on the accuracy of their E_T predictions. This emphasis on E_T prediction in not arbitrary. Rather, it reflects the difficulty of measuring E_T at large scales, and a desire to focus on particular applications of the VIC model where the partitioning of net radiation into surface energy fluxes is

the single critical model determination (e.g. coupling of VIC with a short-term weather prediction model). More hydrological applications of VIC, which focus on the accuracy of model Q predictions, may be better served by single-objective calibration against Q observations. Nevertheless, is it worth noting that improvements in E_T associated with shifting calibrational weighting from Q to $T_{s,\mathrm{GOES}}$ do not generally occur at the direct expense of Q accuracy (Figures 4b and 7b).

In parameter space, Figure 8 suggests that relatively large ranges of parameters can be associated with acceptable $\{Q\}$ fits. Direct $\{E_T\}$ calibration leads to more tightly constrained parameter ranges. One desirable quality in multi-objective $\{Q, T_s\}$ calibration would be an ability to focus relatively diffuse $\{Q\}$ parameter results into ranges that more closely approximate $\{E_T\}$ calibration. Unfortunately, this is not the case. Calibration against both Q and $T_{s,\mathrm{GOES}}$ does little to reduce apparent parameter equifinality problems associated with $\{Q\}$ calibration. Therefore, despite the advantages in criteria space associated with the inclusion of $T_{s,GOES}$ observations, no clear advantageous could be identified in parameter space. This implies a complex response surface for E_T where improved E_T performance is not necessarily associated with convergence to parameter sets derived from direct $\{E_T\}$ calibration.

Acknowledgments. Soil moisture observations from the Oklahoma Mesonet network were made available through the purchase of data by NASA's Land Surface Hydrology Program and NOAA's Office of Global Programs for their funded investigations. This research was partially supported by NOAA grant NA86GP0258 and NASA grant NAG5-9414.

References

Abdulla, F.A., D.P. Lettenmaier, E.F. Wood, and J.A. Smith, Application of a macroscale hydrologic model to estimate the water balance of the Arkansas-Red River Basin, *J. Geophys. Res., 101*, 7449-7459, 1996.

Bastidas, L.A., H.V. Gupta, S. Sorooshian, W.J. Schuttleworth, and Z.L. Yang, Sensitivity analysis of a land surface scheme using multicriteria methods, *J. Geophys, Res, 104*, 19481-19490, 1999.

Beven, K.J., Prophecy, reality, and uncertainty in distributed hydrological modelling, *Adv. in Water Resources, 16*, 41-51, 1993.

Blyth, E.M., A.J. Dolman, and N. Wood, Effective resistance to sensible and latent heat flux in heterogeneous terrain, *Q.J.R. Meteorol., 119*, 423-442, 1993.

Boni, G., D. Entekhabi, and F. Castelli, Land data assimilation with satellite measurements for the estimation of surface energy balance components and surface control on evaporation, *Wat. Resour. Res., 37*, 1713-1722, 2001.

Burke, E.J., R.J. Gurney, L.P. Simmons, and T.J. Jackson, Calibrating a soil water and energy budget model with remotely sensed data to obtain quantitative information about the soil, *Wat. Resour. Res., 33*, 1689-1697, 1997.

Camillo, R.J., P.E. O'Neil, and R.J. Gurney, Estimating soil hydraulic parameters using passive microwave data, *IEEE Trans. on Geosci. and Remote Sens., 24*, 930-936, 1986.

Cherkauer, K. A. and D. P. Lettenmaier, Hydrologic effects of frozen soils in the upper Mississippi River basin, *J. Geophys. Res., 104*, 19599-19610, 1999.

Cosby, B.J., G.M. Hornberger, R.B. Clapp, and T.R. Ginn, A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils, *Wat.*

Resour. Res., 20, 682-690, 1984.

Cosgrove, B.A., D. Lohmann, K. Mitchell, R. Houser, E. Wood, A. Robock, J. Schaake, J. Sheffield, Q. Duan, D. Lettenmeir, R. Pinker, D. Tarpley, and W. Higgins, Realtime and retrospective forcing in the North American Land Data Assimilation System (NLDAS), submitted to *J. Geophys. Res.*, 2002.

Czajkowski, K.P., S.N. Goward, and H. Ouaidrari, Impact of AVHRR filter functions on surface temperature estimation from the split window approach, *Int. J. Remote Sens., 19*, 2007-2012, 1998.

Drusch M and E.F. Wood, On the accuracy of TOVS surface temperatures: A comparison between measured, modeled, and satellite derived data. In *Remote Sensing and Hydrology 2000,* M. Owe, K. Brubaker, J. Ritchie and A. Rango (eds), IAHS Publ. 267, IAHS Press, Wallingford, Oxon, UK, 202-206, 2001.

Duan, Q., S. Sorooshian, V. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res., 28*, 1015-1031, 1992.

Franks, S.W., and K.J. Beven, Bayesian estimation of uncertainty in land surface- atmosphere flux predictions, *J. of Geophys. Res., 102*, 23991-23999, 1997.

Gupta, H.V., S. Soroosh Sorooshian, and P.O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res., 34*, 751-763, 1998.

Gupta, H.V., L.A. Bastidas, S. Sorooshian, W.J. Schuttleworth, and Z.L. Yang, Parameter estimation of a land surface scheme using multicriteria methods, *J. Geophys. Res., 104*, 19491-19503, 1999.

Houser, P.R., H.V. Gupta, W.J. Schuttleworth, and J.S. Famiglietti, Multiobjective calibration and sensitivity of a distributed land surface water and energy balance model, *J. Geophys. Res., 106*, 33421-33434, 2001.

Jin, M., R.E. Dickinson, and A.M. Vogelmann, A comparison of CCM2-BATS skin temperature and surface-air temperature with satellite and surface observations, *J. Climate, 10*, 1505-1524, 1997.

Klassen, W. and M. Claussen, Landscape variability and surface flux parameterization in climate models, *J. Hydrol., 73*, 181-188, 1995.

Kuczera, G., Efficient subspace probabilistic parameter optimization for catchment models, *Water Resour. Res., 22*, 177-185, 1997.

Lakshmi, V., A simple temperature assimilation scheme for use in land surface models, *Wat. Resour. Res., 36*, 3687-3700, 2000.

Lakshmi, V., and J. Susskind, Comparison of TOVS-derived land surface variables with ground observations, *J. Geophys. Res., 105*, 2179-2190, 2000.

Lettenmaier, D.P., D. Lohmann, E.F. Wood, and X. Liang, PILPS-2c workshop report, Princeton University, Princeton NJ, October 1996.

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges, A simple hydrologically based Model of land surface water and energy fluxes for GCMs, *J. Geophys. Res., 99*, 14415-14428, 1994.

Lohmann, D., and many others, The project for intercomparison of land-surface parameterization schemes (PILPS) phase 2(c) Red-Arkansas River basin experiment: 3. Spatial and temporal analysis of water fluxes, *Global and Planetary Change, 19*, 161-179, 1998.

Nijssen, B.N., G.M. O'Donnell, D.P. Lettenmaier and E.F. Wood, Predicting the discharge of global rivers, *J. Clim. 14,* 3307-3323, 2001.

Norman, J.M., W.P. Kustas, and K.S. Humes, A two-source approach for estimating soil and vegetation energy fluxes from observations of directional radiometric surface temperature, *Agric. Forest Meteorol., 77*, 263-293, 1995.

Rhoads, J., R. Dubayah, D. Lettenmeir, G. O'Donnell, and V. Lakshmi, Validation of land surface models using satellite-derived surface temperature, *J. Geophys. Res., 106*, 20085-20099, 2001.

Robock, A., and many others, Evaluation of the North American Land Data Assimilation System over the Southern Great Plains during the warm season, Submitted to *J. Geophys. Res.*, 2003.

Seibert, J., Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrol. and Earth. Sys. Sci., 4*, 215-224, 2000.

Susskind, J., P. Praino, L. Rokke, L. Iredell, and A. Mehta, Characteristics of the TOVS Pathfinder Path A Dataset, *Bull. of the Amer. Met. Soc., 78*, 1449-1472, 1997.

Verma, S., and J. Berry, Net exchange of carbon dioxide in grassland and agricultural ecosystems in the ARM-CART region; modeling and year-round measurements, *Annual Progress Report of FY 1997/1998*, National Institute for Global Environmental Change, 1999.

Yapo, P.O., H.V. Gupta, S. S. Sorooshian, Multi-objective global optimization for hydrologic models, *J. of Hydrol., 204*, 83-97, 1998.

Zhu, A-X, Mapping soil landscape as spatial coninua: The neural network approach, *Wat. Resour. Res., 36*, 663-677, 2000.

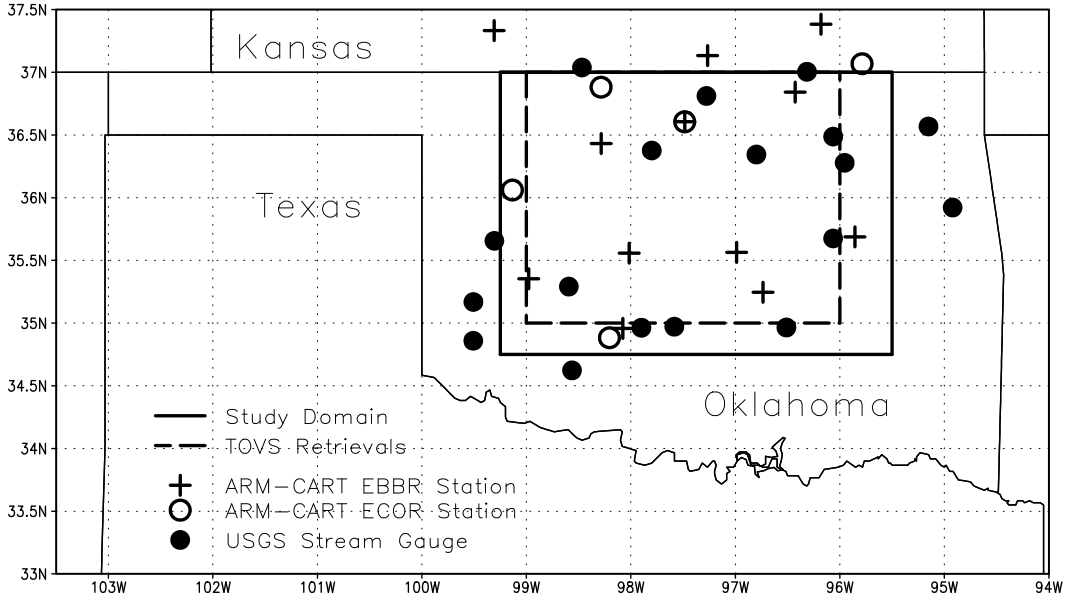

Figure 1. Location of the study domain, ARM-CART surface energy flux tower stations, gauged basin outlets and the TOVS retrieval rectangle within the United States Southern Great Plains. The study domain is taken to be the intersection of the DOE ARM-CART area with the state of Oklahoma.

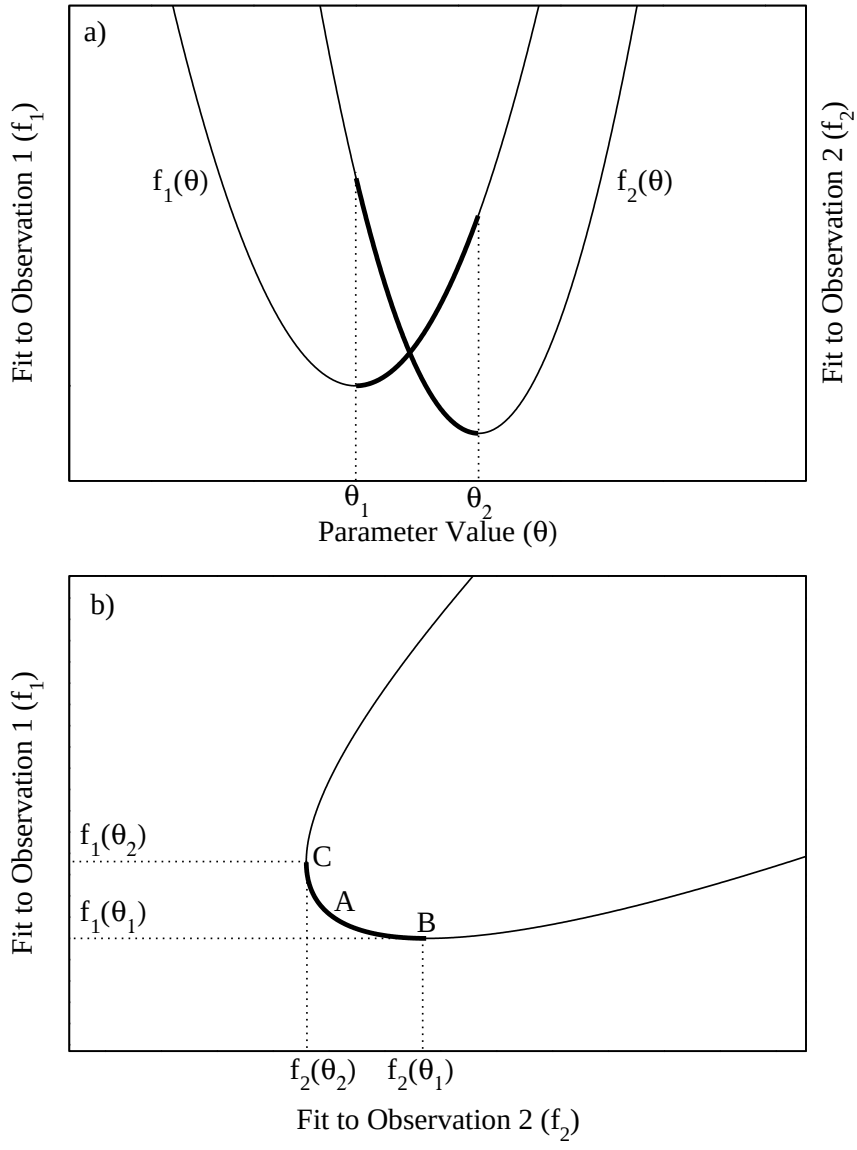

Figure 2. a) Graph of parameter value (θ) versus fitness criteria (f_1 and f_2) for a simple one-parameter multi-objective calibration case. b) Transformation of same case into f_1 versus f_2 fitness-space. Labeled points B and C correspond to fitness criteria results for single-objective calibration against f_1 and f_2 respectively. Labeled point A corresponds to the equal weighting of f_1 and f_2 for multi-objective calibration. Both graphs are modeled after Section 5.0 of *Houser et al.* [2001].

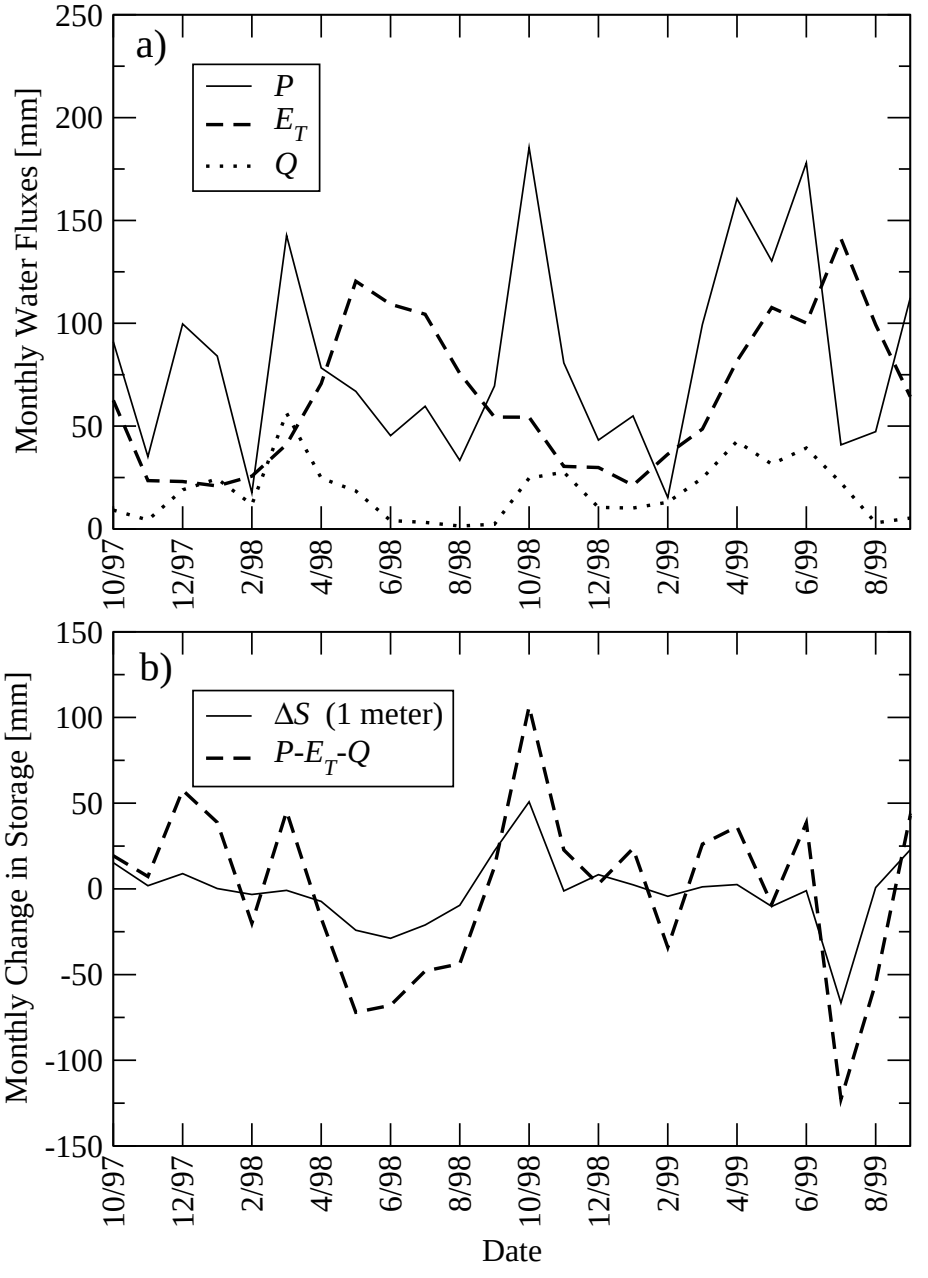

Figure 3. a) Observed monthly totals of precipitation (P), streamflow (Q) and evapotranspiration (E_T) within the study domain shown in Figure 1. b) Residual of measured fluxes (P-E_T-Q) and average observed change in top 1-meter soil moisture (ΔS) over 59 Oklahoma Mesonet sites within the model domain.

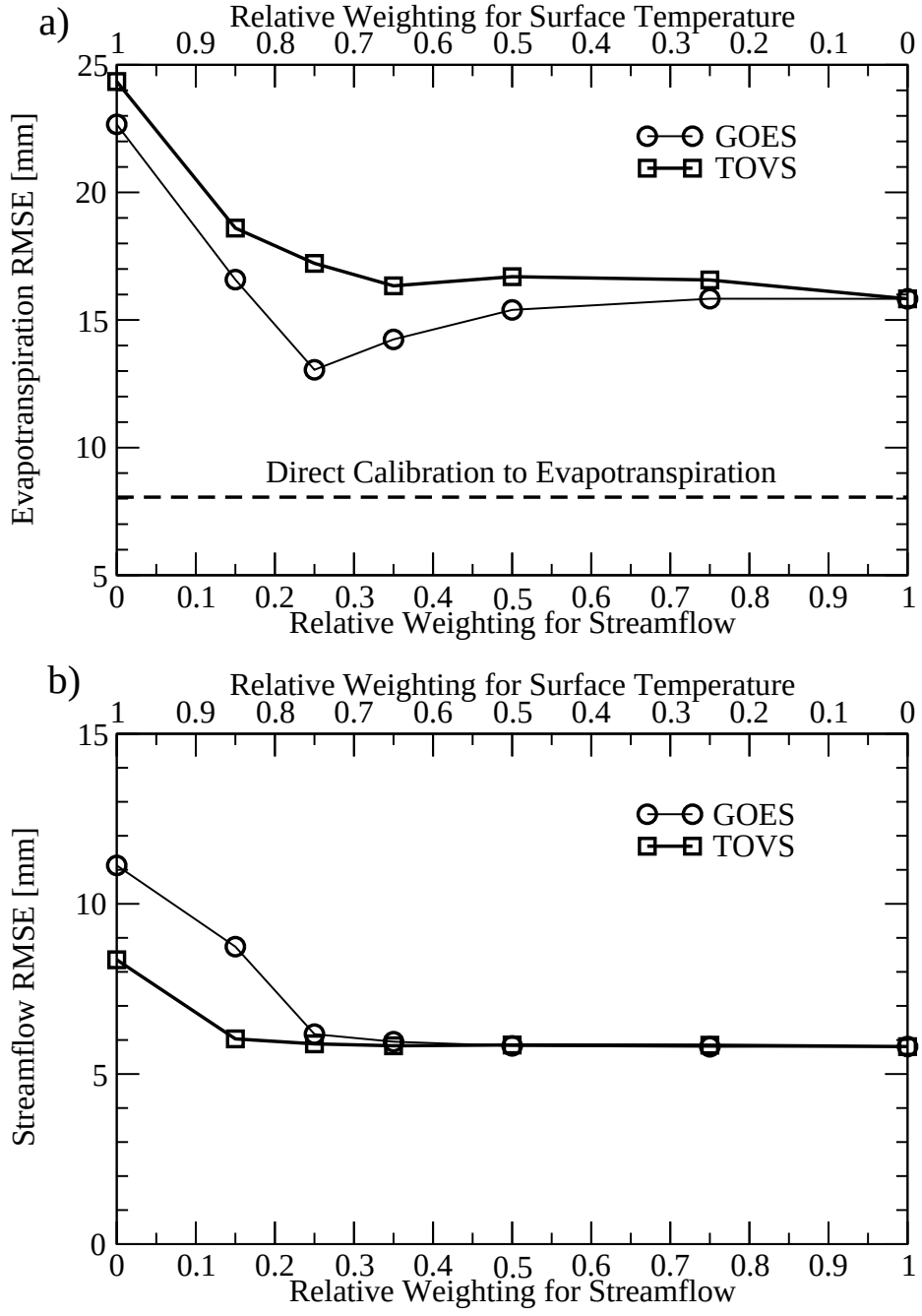

Figure 4. a) Cross-calibration E_T RMSE values for members of the Pareto set derived from $\{Q, T_{s,\mathrm{GOES}}\}$ calibration using the SCE-UA algorithm. Results are shown for both TOVS and GOES T_s retrievals. b) Same but for Q RMSE values.

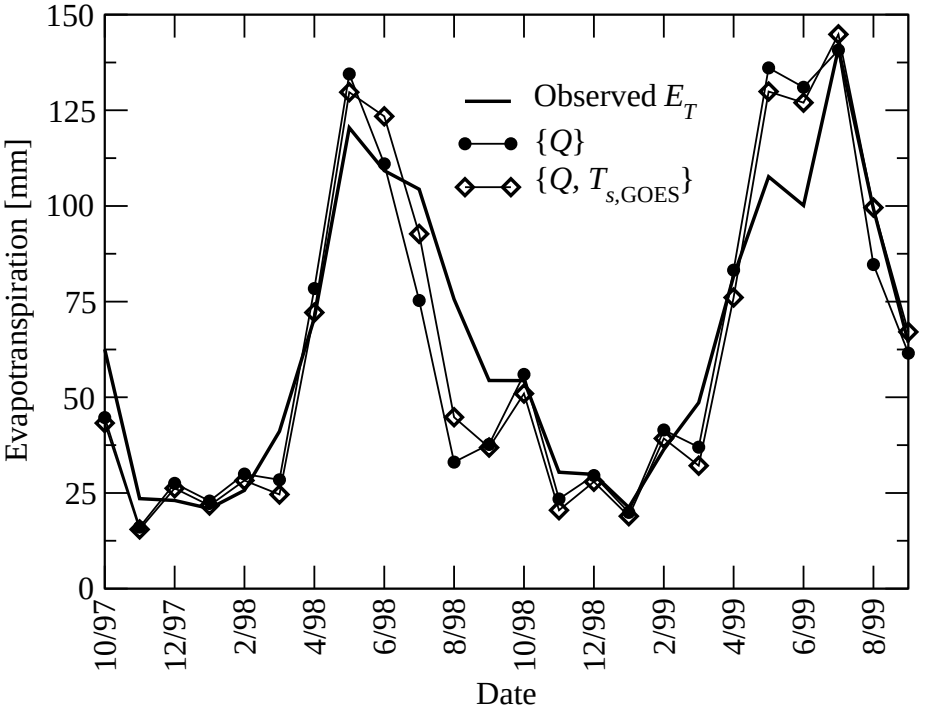

Figure 5. Time series of monthly VIC model E_T predictions derived from single-objective $\{Q\}$ calibration and multi-objective $\{Q, T_{s,\mathrm{GOES}}\}$ calibration using $W_{Ts,\mathrm{GOES}} = 0.75$ and $W_Q = 0.25$.

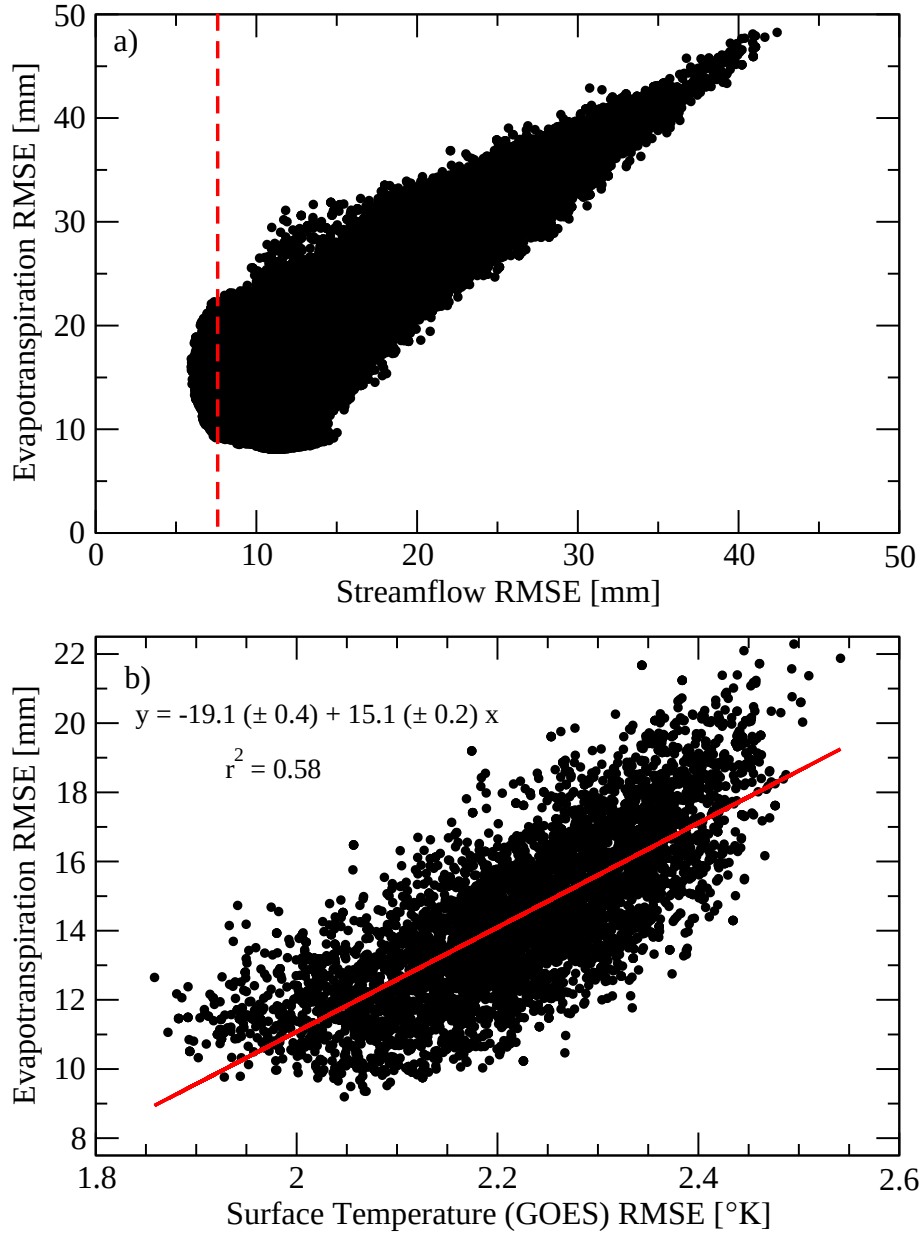

Figure 6. a) Q and E_T RMSE errors for 100,000 VIC model simulations using randomly selected parameters. b) Correlation between $T_{s,\mathrm{GOES}}$ and E_T RMSE for parameter sets with Q fitness criteria judged to be indistinguishable from the best fit (i.e. points to left of the dashed line in Figure 6a).

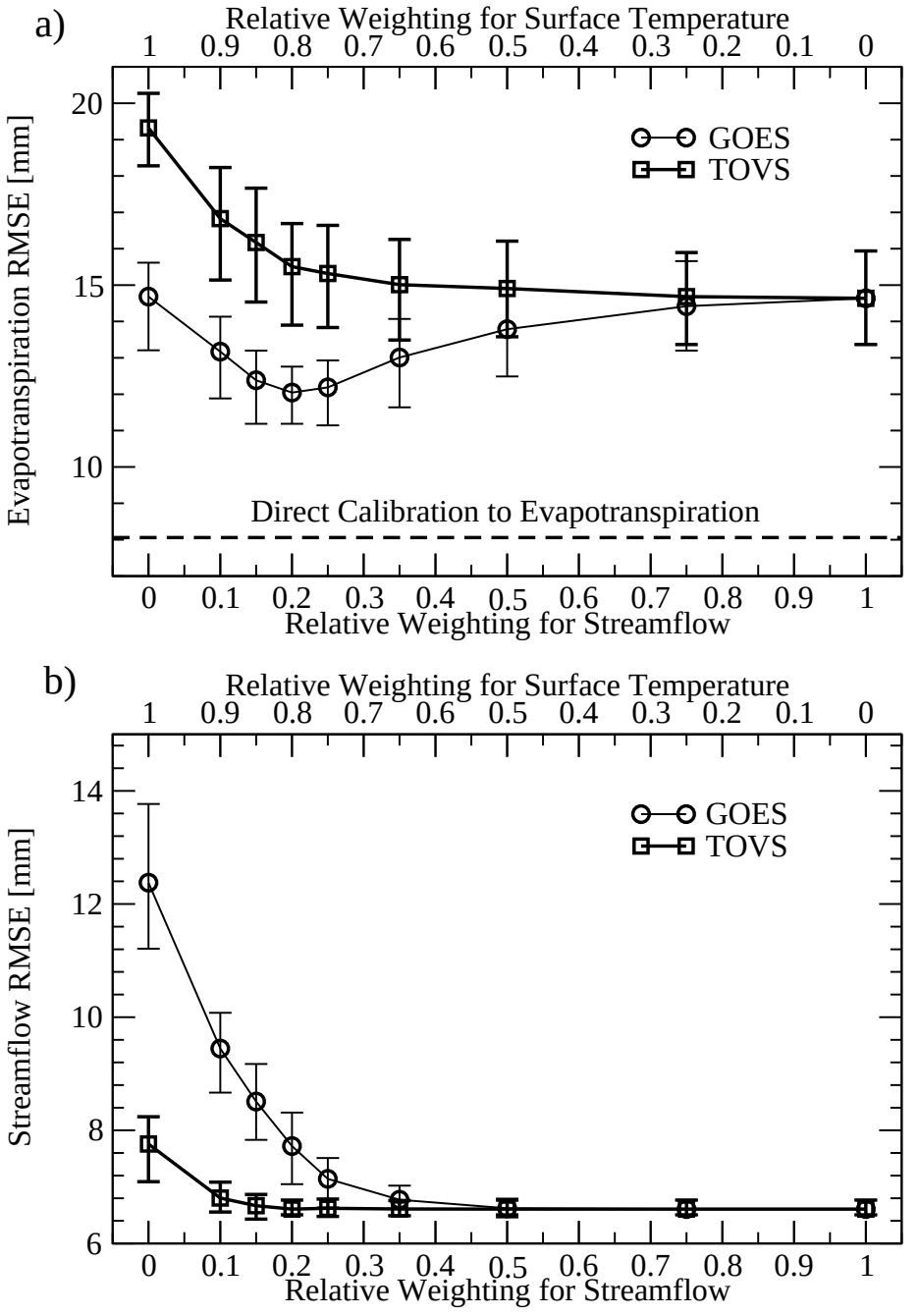

Figure 7. a) Cross-calibration E_T RMSE values for the 1% of randomly selected parameter sets with the lowest G values in (4) for a range of W_{Ts} and W_Q choices. Results are shown for both TOVS and GOES T_s retrievals. b) Same but for Q RMSE values.

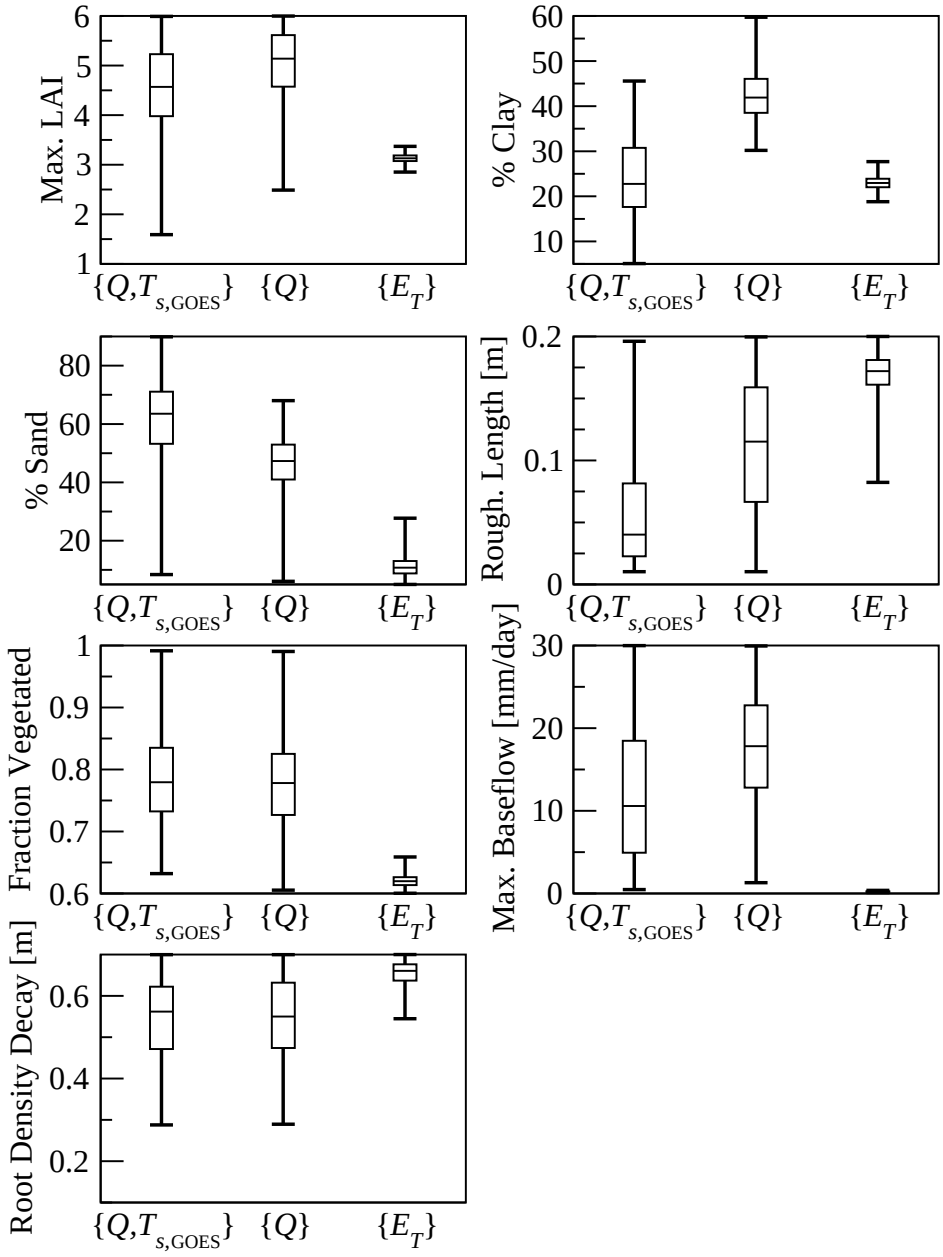

Figure 8. Box and whisker plot for the 1% of randomly selected parameter with the highest fitness (i.e. lowest NRMSE) for Monte-Carlo based $\{Q\}$, $\{E_T\}$, and $\{Q, T_{s,\mathrm{GOES}}\}$ calibration. $\{Q, T_{s,\mathrm{GOES}}\}$ results are based on the relative weighting of Q and $T_{s,\mathrm{GOES}}$ NRMSE in (4) that led to the minimum for E_T error in Figure 7 ($W_Q = 0.20$ and $W_{T_{s,\mathrm{GOES}}} = 0.80$).

Table 1. Calibrated parameters and their maximum and minimum possible values. The given maximum/minimum range represents limited knowledge of parameter values prior to calibration. Values for LAI_{max}, z_o, and k^{-1} are assumed equal for both grassland and winter wheat land cover types while f_{veg} values apply only to grassland areas. See Section 3.4 for details.

Parameter	Explanation	Units	Minimum	Maximum
LAI_{max}	Annual maximum LAI	-	1.0	6.0
$\%_{sand}$	Soil sand percentage	-	5.0	90
$\%_{clay}$	Soil clay pPercentage	-	5.0	60
z_o	Surface roughness length	m	0.01	0.20
f_{veg}	Vegetation fraction	-	0.60	1.00
k^{-1}	Root density decay	m	0.10	0.70
ds_{max}	Maximum baseflow rate	mm day^{-1}	0.01	30.0