Information - based uncertainty decomposition in dual channel microwave remote sensing of soil moisture

Bonan Li[1], Stephen P. Good[1]

[1]Department of Biological & Ecological Engineering, Oregon State University, Corvallis, OR 97330, USA
Correspondence to: Bonan Li (libon@oregonstate.edu)

Abstract. NASA's Soil Moisture Active-Passive (SMAP) mission characterizes global spatiotemporal patterns in surface soil moisture using dual L-band microwave retrievals of horizontal, T_{Bh}, and vertical, T_{Bv}, polarized microwave brightness temperatures through a modeled mechanistic relationship between vegetation opacity, surface scattering albedo ~~(i.e. '*tau-omega*' model)~~and soil effective temperature (T_{eff}). Although this model has been validated against *in situ* soil moisture ~~measurements across sparse validations sites~~, there is lack of systematic characterization of where and why SMAP estimates deviate from the *in situ* observations. Here, ~~soil moisture observations~~ we assess how the information content of *in situ* soil moisture observations from the US Climate Reference Network contrasts with (1) the information contained within raw SMAP observations (i.e. 'informational random uncertainty') derived from T_{Bh}, T_{Bv} and T_{eff} themselves, and (2) with the information contained in SMAP's Dual Channel Algorithm (DCA) soil moisture estimates (i.e. 'informational model uncertainty') derived from the model's inherent structure and parameterizations. ~~are used within a mutual information framework to decompose the overall retrieval uncertainty from SMAPs Modified Dual Channel Algorithm (MDCA) into random uncertainty derived from raw data itself and model uncertainty derived from the model's inherent structure.~~ The results show that, on average, 82% of the information in the *in situ* soil moisture is unexplained ~~uncertainty in~~ by SMAP DCA soil moisture. 36% of the unexplained information is caused by the loss of information in the DCA model process while the remainder is induced by a lack~~inadequacy~~ of additional explanatory power beyond T_{Bh}, T_{Bv} and T_{eff}. Overall, retrieval quality of SMAP DCA soil moisture is negatively correlated with the informational uncertainties, with slight differences across different landcovers. The informational model uncertainty (Pearson correlation of -0.51) was found to be more influential than the informational random uncertainty (Pearson correlation of -0.37). The DCA has a higher informational total uncertainty (88% of unexplained information of *in situ* soil moisture) in shrublands while the informational model uncertainty (31% of the informational total uncertainty) in shrublands is less dominant than other landcovers. ~~We find the fraction of algorithm induced uncertainty is negatively correlated (pearson *r* of -0.48) with correlations between *in situ* observations and MDCA estimates.~~ A decomposition of mutual information between T_{Bh}, T_{Bv} and DCA soil moisture shows that on average 57~~5~~% of ~~the mutual~~ information provided~~is redundantly shared~~ by T_{Bh} and T_{Bv} is redundant. The amount of information redundantly ~~, while the information provided uniquely from both T_{Bh} and T_{Bv} is 15%. The fraction of information redundantly~~ provided by T_{Bh} and T_{Bv} was found to be tightly correlated (~~pearson~~ Pearson correlation of -0.83~~*r* = -0.7~~) to how well the DCA correlated to *in situ* soil moisture~~observations~~. ~~–~~Higher redundant information provided by T_{Bh} and T_{Bv} tends to be found in landcovers with less woody components. The DCA retrieval quality improves as T_{Bh} and T_{Bv} provide more redundant information for the DCA soil moisture. This suggests that the informational redundancy between these remotely sensed observations can be used as independent metric to assess the retrieval quality of algorithms using these data streams. This study provides a baseline approach that can also be applied to evaluate other remote sensing models and understand informational loss as satellite retrievals are translated to end user products.

~~Thus, MDCA overall quality improves as T_{Bh} and T_{Bv} provide more redundant information for the MDCA. This suggests the informational redundancy between these remotely sensed observations can be used as independent metric to assess the overall quality of algorithms using these data streams. This study provides a baseline approach that can also be applied to evaluate other remote sensing models and understand informational loss as satellite retrievals are translated to end user products.~~

1 Introduction

Accurate information on soil moisture is of great importance to understand various of biophysical processes in hydrology, agronomy, and ecosystem sciences (Bassiouni et al., 2020; Uber et al., 2018). The poor spatial representativeness of *in-situ* soil moisture sensors, combined with their labor-intensive installation and maintenance, impedes the application these sensors to understand large scale ecosystem phenomena (Babaeian et al., 2019; Petropoulos et al., 2015). Spaceborne passive microwave remote sensing has been developed as a reliable method to estimate surface soil moisture at large scales (Petropoulos et al., 2015). It leverages the large discrepancies in dielectric properties between liquid water and dry soil that result in a high dependency of soil dielectric constants on soil moisture (Njoku and Entekhabi, 1996). Various microwave frequencies have been available to date, amongst which the L-band (1.4-1.427 GHz) microwave frequencies were found to be more desirable for soil moisture estimation because they can sense soil moisture at a relatively deeper layer (~5cm) and greater vegetation penetration (Njoku and Entekhabi, 1996). Though microwave remote sensing has been investigated for decades, significant uncertainties still exists in both microwave radiometry and in the algorithms used to translate microwave observations to soil moisture estimates.

L-band remote sensing soil moisture estimation uses a radiometer to measure surface emission intensity, which is a linear function of brightness temperature. The brightness temperature is linked with soil moisture and vegetation opacity through the '*tau-omega'* emission model and parameterized by soil and vegetation functions (Njoku and Entekhabi, 1996). The '*tau-omega'* model rationale has been adopted by NASA's Soil Moisture Active-Passive (SMAP) mission~~SMAP~~, which is one of the earth observation missions dedicated to soil moisture estimation at L-band microwave frequency. SMAP implemented two primary algorithms: (1) single channel algorithm (SCA) that uses one polarized brightness temperature as primary input to retrieve soil moisture and (2) the dual channel algorithm (DCA) that can retrieve soil moisture and vegetation opacity simultaneously by taking the polarized brightness temperature information in both horizontal and vertical directions (O'Neill et al., 2020)~~(Peggy O'Neill et al., 2018)~~. There is strong interest in the DCA approach because of~~–~~ its independent estimation of vegetation opacity in lieu of the specified vegetation climatology employed by the SCA~~its independent estimation of vegetation water status~~. Additionally, it has been suggested that using a time-integrated vegetation opacity, as is employed in the multi-temporal dual channel algorithm (MT-DCA) for instance (Piles et al., 2016), improves the estimates of soil and vegetation state. These contrasting approaches, as well as other studies on SMAP's temporal polarized ratio algorithm (TPRA) (Gao et al., 2020) and regularized dual channel algorithm (RDCA) (Chaubell et al., 2019), suggested there is still uncertainty about how SMAP observations of horizontal and vertical brightness temperature can be best translated into estimates of surface properties. Although SMAP can provide spatially explicit soil moisture estimates that have been shown to be useful to understand a set of ecohydrological problems (Jadidoleslam et al., 2019), the soil moisture retrievals are still subject to significant amount of uncertainty due to the imperfection of the model and the forcing datasets. The success of retrieving soil moisture and vegetation opacity are interdependent (Konings et al., 2017) and it is important to consider the how the amount of duplicate information carried within a set of observations limits the number of parameters to be inferred (Konings et al., 2015). Therefore, it is critical to diagnosis and quantify the causality of the uncertainty caused by the SMAP algorithm in order to improve the soil moisture retrieval accuracy.

SMAP soil moisture products have been extensively validated against well-calibrated *in situ* soil moisture using unbiased root mean square error (ubRMSE), bias, RMSE and Pearson correlation coefficients at 'core' and 'sparse' validation sites (Babaeian et al., 2019; Colliander et al., 2017). Additionally, the triple collocation method, which combines *in situ* measurements, SMAP observations, and model fields, has been used to characterize systematic biases and error variances (Chen et al., 2017, 2018). These validation investigations found that SMAP met the required accuracy target (ubRMSE 0.04 ~~c~~m^3/~~c~~m^3) on average, while there exist some locations where the performance of SMAP did not met the expected performance. All these validation studies were focused on finding the general uncertainty of SMAP (which is the deviation of SMAP soil moisture from the *in situ* or reference soil moisture) and cannot diagnose and differentiate where the uncertainty arise. Indeed, the causality of uncertainty of SMAP soil moisture may arise from two aspects: (1) the uncertainty due to the inaccuracies from forcing the datasets and (2) the uncertainty due to poor model ~~form~~ structure and parameterizations. In addition, the evaluation metrics used in these evaluation studies are either heavily depend on *in situ* soil moisture or additional reference dataset, which does not allow for SMAP ~~which challenges SMAP~~ to be validated in some remote and inaccessible areas.

The challenges faced by previous SMAP evaluation investigations can be resolved by leveraging two information quantities (Shannon, 1948): (1) Shannon's entropy, which is the amount of information required to fully describe a random variable and (2) mutual information, which represents the amount information of knowing one variable given the knowledge of another or a set of random variables.~~(1) Shannon's entropy, which describes the inherent uncertainty of a random variable and (2) mutual information, which represents the reduction in uncertainty of one random variable given the knowledge of another random variable.~~ Gong et al. (2013) ~~leveraged~~ estimated these information quantities to partition overall uncertainty in the hydrological modeling process into two categories: (1) random uncertainty that arises by incompleteness of exploratory variable and/or inherent stochasticity of forcing datasets, and (2) model uncertainty that is contributed by poor model parameterization or formulation. The random uncertainty is not resolvable for the given system as it is~~they are~~ only related to the probability distribution of the forcing data itself~~densities~~, while the model uncertainty is reduceable by a better model parameterization.

Given that both horizontal and vertical polarized brightness temperatures are measured by SMAP, it is unclear how each polarization contributes information to the overall performance of the DCA. Recent research on partial information decomposition has provided tremendous opportunities for understanding the nuanced interactions among different variables and model structure. Initially proposed by Williams and Beer (2010) and further advanced by Goodwell and Kumar (2017), this approach has been used to understand environmental processes that links two source variables with a target variable by partitioning multivariate mutual information into unique, redundant and synergistic components.~~. It partitions multivariate mutual information into unique, redundant and synergistic components.~~ The unique information represents the amount of information shared with the target variable from each individual source variable separately~~only from each individual source variable~~. Synergistic information is the information provided to the target while both source variables act jointly. Redundant information is the overlapping information that both source variables redundantly provide to a target. Information partition brings ~~a~~ new insight ~~into~~ by unambiguously characterizing the interdependencies between source variables and a target variable without any underlying modeling assumption. The partitioned components hold potential ~~may be used~~ as a new model evaluation metric that can be used to assess SMAP algorithm performance in remote and inaccessible regions.

In this study, we focus on (1) quantifying the random uncertainty and model uncertainty in SMAP's ~~Modified~~ Dual Channel Algorithm (~~M~~DCA) and understand how model uncertainty is related to ~~M~~DCA retrieval ~~accuracy~~quality; (2) developing an *in situ* and ancillary data independent SMAP ~~M~~DCA evaluation ~~reference~~ metric using partial information decomposition between SMAP ~~M~~DCA soil moisture and horizontally polarized ~~(T_{Bh})~~ and vertically polarized brightness temperature ~~(T_{Bv})~~.

2 Material and Methods

2.1 *In situ* soil moisture

The US Climate Reference Network (USCRN) is a systematic and sustained network that is operated and maintained by National Oceanic and Atmospheric Administration (NOAA) to support climate-impact research with continuous high-quality field observed soil moisture, soil temperature and windspeed at different temporal scales (Bell et al., 2013). The USCRN provides soil moisture observations at five different standard depth (5 cm, 10 cm, 20 cm, 50 cm and 100 cm) in 114 locations of Contiguous U.S. (CONUS). The *in situ* datasets have been used for a wide variety of research such as drought monitoring and satellite soil moisture evaluations (Mishra et al., 2017). The hourly soil moisture datasets at the depth of 5 cm were~~as~~ collected from 51 (12 croplands, 30 grasslands, 5 shrublands, 4 mixed)~~8~~ selected USCRN stations (Fig. 1) based on the availability *in situ* soil moisture dataset and the data quality of SMAP pixels in the study period of 03/31/2015 – 10/01/~~2019~~2020.

2.2 ~~MDCA soil moisture~~SMAP Level-2 datasets

In this study, we acquired horizontally polarized brightness temperature (T_{Bh}), vertically polarized brightness temperature (T_{Bv}), ~~and~~ soil effective temperature (T_{eff}), ~~M~~DCA soil moisture and the fraction of landcover at each selected USCRN station from SMAP Level-2 Radiometer Half-Orbit 36 km EASE-Grid Soil moisture, Version 7 data product (O'Neill et al., ~~P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish,~~ 2020)~~t~~ in the same period as the USCRN soil moisture at every station~~he SMAP Enhanced Level-2 Radiometer Half-Orbit 9 km EASE-Grid Soil Moisture (https://nsidc.org/data/smap), Version 3 in the same period of the USCRN soil moisture at every station (Peggy O'Neill et al., 2018)~~. The extracted data series were filter by the ~~their respective~~ quality flags of~~and the~~ T_{Bh}, T_{Bv} and ~~M~~DCA ~~soil moisture values were kept only when they all simultaneous pass~~

~~quality contro~~soil moisture quality flags~~l~~. We only keep the data points when they all simultaneous pass quality control. ~~M~~DCA retrieves soil moisture based on the '*tau-omega'* model (O'Neill et al., ~~, P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish,~~ 2020), which is a well- known radiative transfer-based soil moisture retrieval algorithm in the passive microwave soil moisture community. It requires the brightness temperatures~~ (T_B)~~ as the main inputs, soil effective temperature as an ancillary input, and is parameterized based on~~by~~ overlaying vegetation and soil surface information. The ~~M~~DCA iteratively~~invert~~ feeds the '*tau-omega'* model with initial guesses of ~~surface~~ soil moisture and vegetation optical depth. The retrieved soil moisture is assumed to be close to the real value when the estimated brightness temperatures are similar to the satellite observed brightness temperature (Konings et al., 2015; O'Neill et al., ~~P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish,~~ 2020)~~ The guesses of soil moisture and vegetation optical depth are adjusted iteratively until they minimize the difference between satellite observed T_B and inverted T_B from a least square perspective~~. Compared to the SCAs, the ~~M~~DCA uses a different polarization mixing factor function and different values of vegetation single scattering albedo ~~updates roughness and the polarization mixing parameters~~ (Chaubell et al., 2020).

The SMAP fraction of landcover data field provides the fraction of top three dominate landcovers that were classified by International Geosphere – Biosphere Programme (IGBP) ecosystem surface classification scheme at each pixel (Seitzinger et al., 2015). The IGBP classified land surface into water, evergreen needleleaf forest, evergreen broadleaf forest, deciduous needleleaf forest, deciduous broadleaf forest, mixed forest, closed shrublands, open shrublands, woody savannas, savannas, grasslands, permanent wetlands, croplands, urban and built-up, croplands/natural vegetation mosaics, snow and ice, barren. In this study, the landcover of the study site was classified as the most dominate landcover if the fraction of the most dominate landcover was greater than 50%. Otherwise, the landcover of the study site is classified as the "mixed" landcover. Furthermore, the study sites that are dominated by woody savanna were classified as savannas, by closed/open shrublands were classified as shrublands, by cropland/natural vegetation mosaics were classified as croplands.

2.3 Information - based uncertainty decomposition

~~Shannon's entropy is a quantity that express the inherent uncertainty associated with a random variable. Commonly, modeling efforts are focused on reducing the uncertainty in the variable of interest, which is denoted as $H(Y_{obs})$, using other explanatory variables through some physically- or empirically- based models. Most of models being constructed of natural processes are not perfect, and the model outputs are often not capable of capturing the information of the "truth". In theory, there exists a best achievable model performance that describe the variable of interest the best for a particular system given the available datasets (Gong et al., 2013); yet detailed structure of best achievable model performance is often unknown. Although the detailed structure of best achievable model performance maybe remain unknown, mutual information, denoted as $I(\boldsymbol{X}_{Inputs}; Y_{obs})$ where $\boldsymbol{X}_{Inputs}$ are the available inputs and Y_{obs} is the *in situ* measured variable of interest, can provide a good benchmark measure. The quantity $I(\bullet;\bullet)$ represents the amount of uncertainty reduced due to the knowledge of either variable in this function.~~

~~It should be noted that a model is a formal hypothesis that maps input datasets space to output dataset space in the form of a mathematical function. Therefore, the model hypothesis (function), at least, cannot provide new information. This is expressed as the data processing inequality which states that "no clever manipulation of the data can improve the inferences~~

~~that can be made from the data" (Cover and Thomas, 2005). Formally, if random variables X, Y, Z are said to form a Markov chain (denoted by $X \rightarrow Y \rightarrow Z$), wherein the conditional distribution of Z only depends on Y and is conditionally independent of X, then X can only influence Z via the knowledge of Y and knowing Z can only decrease the amount of X tells about Y. The formula of data processing inequality is defined as:~~

~~$I(X, Y) \quad \geq \quad I(X, Z)$ (1)~~

~~Hence, given the measure of best achievable model performance and data processing inequality, the relationship between input, output, and *in situ* measurements in any modeling processes can be expressed as follows:~~

~~$H(Y_{\text{obs}}) \geq \quad I(\boldsymbol{X}_{\text{Inputs}}; Y_{\text{obs}}) \quad \geq \quad I(Y_{\text{model}}; Y_{\text{obs}})$ (2)~~

~~The relationship equation (2) allow us to differentiate two types of uncertainties, (1) random uncertainty, which unresolvable due to the randomness of the input datasets, that is the difference between $H(Y_{\text{obs}})$ and $I(\boldsymbol{X}_{\text{Inputs}};Y_{\text{obs}})$; (2) model uncertainty, which is resolvable due to the inadequacy of model, that is the information gap between and $I(\boldsymbol{X}_{\text{Inputs}}; Y_{\text{obs}})$ and $I(Y_{\text{model}};Y_{\text{obs}})$.~~

~~In our case, $\boldsymbol{X}_{\text{Inputs}}$ are T_{Bh} and T_{Bv}, Y_{obs} is the *in situ* surface soil moisture, Y_{model} is MDCA soil moisture. The $H(Y_{\text{obs}})$ can be calculated as:~~

~~$H(Y_{\text{obs}}) = -\sum_{y \in Y_{obs}} p(y) \log_2 p(y)$ (3)~~

~~Where $p(y)$ probability mass function of Y_{obs} that is estimated by a fixed bin method (Freedman and Diaconis, 1981). This method calculates $H(Y_{\text{obs}})$ in unit of bits. Previous study has indicated that this method may underestimates the true entropy (Paninski, 2003). Therefore, we leveraged the simple Miller-Madow corrected entropy estimator (Chen et al., 2018a) and applied a normalization method to remove the bias that may cause by the heterogeneity in length of available datasets across all stations. We acknowledge that there exist several entropy correction and estimation methods. However, we pick this Miller-Madow correction based on its simplicity and effectiveness. The corrected and normalized entropy is then expressed as follows:~~

~~$H_{\text{CN}}(Y_{\text{obs}}) = \frac{H(Y_{\text{obs}}) + \frac{K-1}{2n}}{log_{2}^{n}}$ (4)~~

~~Where $H_{\text{CN}}(Y_{\text{obs}})$ is the Miller-Madow corrected and normalized entropy, hereafter entropy, n is the number of data points that were used to calculate the normalized entropy, K is the number of non-zero probabilities associate based on the fixed binned method.~~

~~The computation of two types of uncertainties require the estimation of $I(\text{T}_{\text{Bv}}, \text{T}_{\text{Bh}}; Y_{\text{obs}})$ and $I(Y_{\text{MDCA}}; Y_{\text{obs}})$, which can be computed via the following equation:~~

~~$I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; Y_{\mathrm{obs}}) = H_{\mathrm{CN}}(\mathrm{T_{Bh}}, \mathrm{TB_{v}}) + H_{\mathrm{CN}}(Y_{\mathrm{obs}}) - H_{\mathrm{CN}}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, Y_{\mathrm{obs}})$ (5)~~

~~Where $H_{\mathrm{CN}}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}})$ and $H_{\mathrm{CN}}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, Y_{\mathrm{obs}})$ the estimated joint entropy that describes the uncertainty associated with a set of variables. $H_{\mathrm{CN}}(Y_{\mathrm{MDCA}}; Y_{\mathrm{obs}})$ can be estimated by replacing the $\mathrm{T_{Bh}}$, $\mathrm{T_{Bv}}$ with Y_{MDCA} on both side of the equation. It worth noting that the joint entropies are estimated using equation (3) except they require the estimation of joint probability mass functions that are also estimated using the fixed bin method (Freedman and Diaconis, 1981).~~

The fundamental quantity of information theory is Shannon's entropy, which represents the amount of information required to fully describe a random variable (Cover and Thomas, 2005). Shannon's entropy is the basic building block of computing mutual information and the informational uncertainties. The entropy of a single random variable is defined as

$$H(\mathrm{X}) = -\sum_{x\in X} p(x) log_2\, p(x), \qquad (1)$$

where $p(x)$ is the probability mass function of random variable X. The estimation of $p(x)$ often involves discretizing the values of X into a set of bins and then the $p(x)$ of a specific bin is computed by dividing the total number of datapoints within a specific bin by the total of number of data points of X. The number of bins in this study is estimated by Freedman-Diaconis binning method (Freedman and Diaconis, 1981). The entropy calculated by eq. (1) is in unit of bits.

Previous study has indicated that this method (eq. (1)) may underestimate the true entropy (Paninski, 2003). Therefore, we leveraged the simple Miller-Madow corrected entropy estimator (Zhang and Grabchak, 2013) and applied a normalization method to remove the bias that may cause by the heterogeneity in length of available datasets across all stations. We acknowledge that there exist several entropy correction and estimation methods. However, we select the Miller-Madow correction based on its simplicity and effectiveness. The corrected and normalized entropy is then expressed as

$$H_{CN}(\mathrm{X}) = \frac{H(\mathrm{X}) + \frac{K-1}{2n}}{log_2\, n}, \qquad (2)$$

where $H_{\mathrm{CN}}(\mathrm{X})$ is the Miller-Madow corrected and normalized entropy of random variable X (hereafter entropy), $H(\mathrm{X})$ is the uncorrected entropy from eq. (1), n is the number of data points of X , K is the number of non-zero probabilities (bins contains more than one data point) based on the fixed binned method (Freedman and Diaconis, 1981). In this study, all entropies of single random variables in the later equations (i.e., $H_{CN}(\mathrm{T_{Bh}})$, $H_{CN}(\mathrm{T_{Bv}})$, $H_{CN}(\textit{in situ})$ etc.) are computed using the combination of eq. (1) and eq. (2) with the replacement of $p(\bullet)$ by their individual probability mass functions.

The joint entropy is a critical intermediate information quantity to calculate these informational uncertainties. It represents the amount of information required to describe a set of random variables. The joint entropy for two random variables is defined as

$$H(\mathrm{X}, \mathrm{Y}) = -\sum_{x\in X}\sum_{y\epsilon Y} p(x,y) log_2\, p(x,y), \qquad (3)$$

where $p(x, y)$ is the joint probability mass function associated with X and Y that is estimated by the same method mentioned above. The same normalization and correction method of eq. (2) is applied to joint entropy of eq. (3). The entropy after the correction and normalization is formulated as

$$H_{CN}(\mathrm{X, Y}) = \frac{H(\mathrm{X,Y}) + \frac{K-1}{2n}}{log_2 n}, \qquad (4)$$

where H_{CN}(X, Y) is the corrected and normalized joint entropy of random variable associated with $\{X, Y\}$, H(X, Y) is the uncorrected entropy from eq. (3), n is the number of data points that were used to calculate the normalized joint entropy (hereafter joint entropy), K is the number of non-zero joint probabilities based on the Freeman and Diaconis method (Freedman and Diaconis, 1981). All the joint entropies that are associated with two or more random variables in the later equations (i.e., H_{CN}(*in situ*, DCA), H_{CN}(T_{Bh}, T_{Bv}, DCA), H_{CN}(T_{Bh}, T_{Bv}, T_{eff}, *in situ*) etc.) are computed using the combination of eq. (3) and eq. (4) with the replacement of p(•) by their joint probability mass functions, respectively.

Commonly, modeling efforts are focused on capturing the information of a random variable of interest via other explanatory variables through some physically- or empirically- based models. However, most of models being constructed of natural processes are not perfect, and the model outputs are often not capable of capturing the exact relationship between the available input variables and the variable of interest (Gupta et al., 1998). In theory, there exists a maximum achievable performance of a model that describes the variable of interest the best for a particular system given the available datasets (Gong et al., 2013); yet the detailed structure of this model is often unknown. Mutual information (Cover and Thomas, 2005), for instance I(A; B), is a measure of the amount information due to the knowledge of knowing either random variable A or B in the function I(•;•). Mutual information between model inputs and *in situ* observations of model output (hereafter *in situ* observations) can be used as a useful and effective measure of best achievable performance model because it links the model inputs and *in situ* observations only through the joint and marginal probability mass functions that do not involve any priori model assumptions (Gong et al., 2013).

The mutual information is defined based on entropy and joint entropy. The mutual information between T_{Bh} and DCA, and the mutual information between T_{Bv} and DCA, are computed as

$$I(\mathrm{T_{Bh}; DCA}) = H_{CN}(\mathrm{T_{Bh}}) + H_{CN}(\mathrm{DCA}) - H_{CN}(\mathrm{T_{Bh}, DCA}) \qquad (5)$$

and

$$I(\mathrm{T_{Bv}; DCA}) = H_{CN}(\mathrm{T_{Bv}}) + H_{CN}(\mathrm{DCA}) - H_{CN}(\mathrm{T_{Bv}, DCA}). \qquad (6)$$

The mutual information between *in situ* and DCA soil moisture is computed as

$$I(\mathrm{DCA}; \textit{in situ}) = H_{CN}(\mathrm{DCA}) + H_{CN}(\textit{ in situ }) - H_{CN}(\mathrm{DCA}, \textit{in situ }). \qquad (7)$$

The mutual information between DCA and *in situ* soil moisture is calculated as

$$I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; \mathrm{DCA}) = H_{CN}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}) + H_{CN}(\mathrm{DCA}) - H_{CN}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{DCA}). \quad (8)$$

The mutual information between $\mathrm{T_{Bh}}$, $\mathrm{T_{Bv}}$, $\mathrm{T_{eff}}$ and *in situ* soil moisture is computed as:

$$I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{T_{eff}}; \textit{in situ}) = H_{CN}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{T_{eff}}) + H_{CN}(\textit{in situ}) - H_{CN}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{T_{eff}}, \textit{in situ}). \quad (9)$$

For a given system in which the inputs and output are linked via mathematical functions, the mutual information between model outputs and *in situ* observation can never exceed the entropy of the *in situ* observations. This information gap is defined as informational total uncertainty (I_{Tot}). The mutual information between the *in situ* observations and the available explanatory variables is also always smaller than the entropy of *in situ* observations. This information gap, defined as informational random uncertainty (I_{Rnd}), is caused by the existence of inherent data uncertainty of the explanatory variables and a lack of complete explanatory variables to fully capture the information in the *in situ* observations. Furthermore, the mutual information between model inputs and *in situ* observations should equal to the mutual information between *in situ* observations and model output if the model hypothesis completely captures or correctly expresses the true relationship between model inputs and *in situ* observations. However, it's commonly known that "All models are wrong, but some are useful" (Peters and Kok, 2016) and model assumptions typically cannot fully express the true relationship between the explanatory variables and *in situ* observations. Hence, the mutual information between model output and *in situ* observation is expected to be smaller than the mutual information between model inputs and *in situ* observations. This information gap, defined as informational model uncertainty (I_{Mod}) is induced by poor model assumption, formulations, and/or inappropriate model parameterizations. Therefore, the informational total uncertainty (I_{Tot}) is the sum of the informational random uncertainty and informational model uncertainty come naturally given the explicitly definition of these informational uncertainties. In this study, the explanatory variables of DCA are $\mathrm{T_{Bh}}$, $\mathrm{T_{Bv}}$ and the $\mathrm{T_{eff}}$. The *in situ* observation and model output are *in situ* USCRN soil moisture and DCA soil moisture, respectively.

Leveraging eq. (7) and eq. (9), the DCA informational random uncertainty (I_{Rnd}), DCA informational model uncertainty (I_{Mod}), and DCA total informational uncertainty (I_{Tot}) calculated are calculated as:

$$I_{Rnd} = H_{CN}(\textit{in situ}) - I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{T_{eff}}; \textit{in situ}), \quad (10)$$

$$I_{Mod} = I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{T_{eff}}; \textit{in situ}) - I(\mathrm{DCA}; \textit{in situ}), \quad (11)$$

and

$$I_{Tot} = H_{CN}(\textit{in situ}) - I(\mathrm{DCA}; \textit{in situ}) = I_{Rnd} + I_{Mod}. \quad (12)$$

2.4 Partial information decomposition

~~This method partitions multivariate shared information to unique, redundant and synergistic components. The decomposed information components on the model inputs and outputs maybe indicative on understand informational loss as model inputs are translated to end user products and these components may have the potential for evaluating model performance. The partial information decomposition of MDCA can be expressed as follows:~~

~~$I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; Y_{\mathrm{MDCA}}) = U_1(\mathrm{T_{Bh}}; Y_{\mathrm{MDCA}}) + U_2(\mathrm{T_{Bv}}; Y_{\mathrm{MDCA}}) + R(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; Y_{\mathrm{MDCA}}) + S(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; Y_{\mathrm{MDCA}})$ (6)~~

~~Where U_1 and U_2 are unique information of $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ shared with Y_{MDCA}, respectively. S and R are the synergistic information and redundant information that $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ shared with Y_{MDCA}, respectively. All the decomposed components are non-negative real values.~~

~~The individual mutual information between $\mathrm{T_{Bh}}$, $\mathrm{T_{Bv}}$ and Y_{MDCA} can be expressed as follows:~~

~~$I(\mathrm{T_{Bh}}; Y_{\mathrm{MDCA}}) = U_1(\mathrm{T_{Bh}}; Y_{\mathrm{MDCA}}) + R(\mathrm{T_{Bh}},\mathrm{T_{Bv}}; Y_{\mathrm{MDCA}})$ (7)~~

~~$I(\mathrm{T_{Bv}}; Y_{\mathrm{MDCA}}) = U_1(\mathrm{T_{Bv}}; Y_{\mathrm{MDCA}}) + R(\mathrm{T_{Bh}},\mathrm{T_{Bv}}; Y_{\mathrm{MDCA}})$ (8)~~

~~U_1, U_2, S and R are unknowns in the systems of equations (6) - (8). Therefore, additional information is need to fully estimated one of these unknowns. We used the approach proposed by Goodwell and Kumar (2017) to estimate R as follows:~~

~~$R = R_{\mathrm{min}} + I_s{*}(R_{\mathrm{MMI}} - R_{\mathrm{min}})$ (9)~~

~~Where R_{min} is represents a lower bound for R that is expressed as:~~

~~$R_{\mathrm{min}} = \max(0, -II)$ (10)~~

~~The inter-dependency of $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ represented by I_s and computed as:~~

~~$I_s = \frac{I(\mathrm{T_{Bh}};\mathrm{T_{Bv}})}{\min\{H(\mathrm{T_{Bh}}); H(\mathrm{T_{Bv}})\}}$ (11)~~

~~II is interaction information that can be positive or negative. II is computed as:~~

~~$II = I(\mathrm{T_{Bh}}; Y_{\mathrm{MDCA}}|\mathrm{T_{Bv}}) - I(\mathrm{T_{Bh}}; Y_{\mathrm{MDCA}})$ (12)~~

The distinct informational contributions of $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ to the DCA outputs are be assessed through a decomposition of the information. This method partitions multivariate mutual information to unique, redundant and synergistic components (Williams and Beer, 2010). The decomposed information components on the DCA model inputs and outputs are expected to indicative of informational flow as model inputs are translated to end user products, and these components may have potential for evaluating model performance. The partial information decomposition of $I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; \mathrm{DCA})$ can be expressed as

$$I(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; \mathrm{DCA}) = U_1(\mathrm{T_{Bh}}; \mathrm{DCA}) + U_2(\mathrm{T_{Bv}}; \mathrm{DCA}) + R(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; \mathrm{DCA}) + S(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}; \mathrm{DCA}), \qquad (13)$$

where U_1 and U_2 are unique information of $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ shared with DCA, respectively. S and R are the synergistic information and redundant information that $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ shared with DCA estimates, respectively. All the decomposed components are non-negative real values.

The mutual information between $\mathrm{T_{Bh}}$ and DCA and mutual information between $\mathrm{T_{Bv}}$ and DCA were defined as

$$I(\mathrm{T_{Bh}}; \mathrm{DCA}) = U_1(\mathrm{T_{Bh}};\mathrm{DCA}) + R(\mathrm{T_{Bh}},\mathrm{T_{Bv}}; \mathrm{DCA}) \qquad (14)$$

and

$$I(\mathrm{T_{Bv}}; \mathrm{DCA}) = U_2(\mathrm{T_{Bv}}; \mathrm{DCA}) + R(\mathrm{T_{Bh}},\mathrm{T_{Bv}}; \mathrm{DCA}), \qquad (15)$$

where U_1, U_2, S and R are unknowns in the systems of equations (13) - (15). Goodwell and Kumar (2017) showed that the R can be formulated as

$$R = R_{\mathrm{min}} + I_{\mathrm{s}}{*}(R_{\mathrm{MMI}} - R_{\mathrm{min}}), \qquad (16)$$

where

$$I_s = \frac{I(\mathrm{T_{Bh}};\mathrm{T_{Bv}})}{\min\{H_{CN}(\mathrm{T_{Bh}}); H_{CN}(\mathrm{T_{Bv}})\}}, \qquad (17)$$

$$R_{MMI} = \min(I(\mathrm{T_{Bh}}; \mathrm{DCA}), I(\mathrm{T_{Bv}}; \mathrm{DCA})) \qquad (18)$$

and

$$R_{\mathrm{min}} = \max(0, -II) \qquad (19)$$

The II is the interaction information of $\mathrm{T_{Bh}}$, $\mathrm{T_{Bv}}$, DCA and can be computed as:

$$\begin{aligned} II = I(\mathrm{T_{Bh}}; \mathrm{DCA}|\, \mathrm{T_{Bv}}) - I(\mathrm{T_{Bh}}; \mathrm{DCA}) = H_{CN}(\mathrm{T_{Bh}}, \mathrm{DCA}) + H_{CN}(\mathrm{T_{Bv}}, \mathrm{DCA}) + \\ H_{CN}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}) - H_{CN}(\mathrm{T_{Bh}}) - H_{CN}(\mathrm{T_{Bv}}) - \mathrm{H_{CN}}(\mathrm{DCA}) - H_{CN}(\mathrm{T_{Bh}}, \mathrm{T_{Bv}}, \mathrm{DCA}) \end{aligned} \qquad (20)$$

It is important to acknowledge that we used the point based *in situ* soil moisture as the ground truth in this analysis. Due to course spatial resolution of SMAP products, we acknowledge that *in situ* soil moisture may not be able to represent the spatial averaged soil moisture well. Although the nominal sensing depth of L-band SMAP soil moisture is 5 cm, the penetration depth was found to be even shallower in wetter regions (Shellito et al., 2016). In fact, the L-band sensing depth was found to as little as ~1cm (Jackson et al., 2012) and can be more sensitive to surface meteorological conditions and more random than the actual *in situ* soil moisture. The heterogeneity in each pixel relative to the *in situ* observations together with the sensing depth disparity may negatively influence the results of this study and result in an overestimate the actual informational uncertainties. We also acknowledge the existence of upscaling methods for matching the *in situ* soil moisture to satellite footprint (Crow et al., 2012). However, most of upscaling methods are achieved under the assistance of additional reference soil moisture datasets. This process introduces additional pieces of information in the DCA system making the separation of the uncertainty induced by the upscaling algorithm or additional dataset from other informational uncertainties much harder. Additionally, we used the hourly *in situ* data to best match the SMAP DCA soil moisture retrievals in time (within an hour). Therefore, it is hard to find a reference dataset at with high frequency in time domain and good spatial coverage. Here we consider the informational uncertainty caused by the spatial mismatch and sensing depth mismatch between *in situ* and DCA soil moisture as part of the informational random uncertainty (I_{Rnd}). Because the DCA essential is a mathematical function and does not inherently require the inputs to be at a specific resolution. The spatial resolution is often the inherent attribute of the data. The sensing depth is more of imperfection L-band sensor. The reader should also keep these in mind while interpreting and adopting the results in this study.

3 Results

3.1 Information quantities and system informational uncertainties

Figure 2 shows the estimated entropies~~y~~ ~~and mutual information quantities~~ across all the study sites while Figure 3 shows the mutual information quantities. The H_{CN}(T$_{\mathrm{Bh}}$) and H_{CN}(T$_{\mathrm{Bv}}$) general follow the same pattern with both having an average value of ~0.37. Although the patterns of H_{CN}(T$_{\mathrm{Bh}}$) and H_{CN}(T$_{\mathrm{Bv}}$) are similar, the H_{CN}(T$_{\mathrm{Bh}}$) is slightly more variable than H_{CN}(T$_{\mathrm{Bv}}$) with the coefficients of variation (CV) being 0.05 and 0.04, respectively. H_{CN}(T$_{\mathrm{eff}}$) shares the same average with H_{CN}(T$_{\mathrm{Bh}}$) and H_{CN}(T$_{\mathrm{Bv}}$), whereas the patterns of H_{CN}(T$_{\mathrm{eff}}$) is quite different (Fig. 2). On average, the H_{CN}(*in situ*) is 0.35, while H_{CN}(DCA) and 0.38. In general, H_{CN}(DCA) follows the pattern of H_{CN}(in situ) with the CV of H_{CN}(DCA) (0.05) being smaller than the CV of H_{CN}(*in situ*) (0.08). As shown in Figure 4a, H_{CN}(T$_{\mathrm{Bh}}$), H_{CN}(T$_{\mathrm{Bv}}$) and H_{CN}(DCA) are significantly correlated with H_{CN}(*in situ*), while no significant correlation is found between H_{CN}(*in situ*) and H_{CN}(T$_{\mathrm{eff}}$). The H_{CN}(DCA) has the strongest correlation strength with H_{CN}(*in situ*) compared with other entropy quantities (Fig. 4a).

~~It is shown that the joint entropy of T$_{\mathrm{Bh}}$ and T$_{\mathrm{Bv}}$ (H_{CN}(h,v)) are always the largest compared to other information quantities. On average, H_{CN}(h,v) is 0.53 bits, which is greater than the entropies of MDCA soil moisture, H_{CN}(MDCA), and *in situ* soil moisture, H_{CN}(*in situ*), (0.38 and 0.35, respectively). Although the pattern of H_{CN}(MDCA) and H_{CN}(*in situ*) are similar, the H_{CN}(*in situ*) is more variable than H_{CN}(MDCA) with the coefficients of variation (CV) being 0.08 and 0.05, respectively. Mutual information between T$_{\mathrm{Bh}}$, T$_{\mathrm{Bv}}$ and *in situ* soil moisture, I(h,v; *In situ*), and mutual information between MDCA soil moisture and *in situ* soil moisture, I(MDCA; *In situ*), are the least information quantities, as they are expected to be. I(h,v; *In situ*) follows the pattern of I(MDCA; *In situ*) with the mean values being 0.09 and 0.06, respectively.~~

The mutual information quantities (Fig. 3) are shown to be generally smaller than the entropy quantities (Fig. 2). On average, I(T$_{\mathrm{Bh}}$,T$_{\mathrm{Bv}}$; DCA) is 0.14, while the I(DCA; *in situ*) and I(T$_{\mathrm{Bh}}$,T$_{\mathrm{Bv}}$, T$_{\mathrm{eff}}$; *in situ*) are 0.06 and 0.17 (Fig. 3), respectively. I(T$_{\mathrm{Bh}}$,T$_{\mathrm{Bv}}$, T$_{\mathrm{eff}}$; *in situ*) is significantly correlated (0.58) with H_{CN}(*in situ*), while no significant correlation is found for other two mutual information quantities (Fig. 4b). It is noticeable that there exists a large information gap (Fig. 2 and Fig. 3) between H_{CN}(*in situ*) and I(T$_{\mathrm{Bh}}$,T$_{\mathrm{Bv}}$, T$_{\mathrm{eff}}$; *in situ*) and I(T$_{\mathrm{Bh}}$,T$_{\mathrm{Bv}}$, T$_{\mathrm{eff}}$; *in situ*) and I(DCA; *in situ*). These information gaps confirm the existence of informational random uncertainty (I_{Rnd}) and informational model uncertainty (I_{Mod}) in the SMAP DCA system. On average, the SMAP DCA explains 18% of the H_{CN}(*in situ*) leaving 82% of the H_{CM}(*in situ*) that is unexplained (Table 1) as informational total uncertainty (I_{Tot}). 36% (Table 1) of the I_{Tot} is caused by I_{Mod}, while the rest is induced by I_{Rnd}. The information uncertainties vary slightly across different landcovers. On average, the SMAP DCA system is capable of capturing more information of H_{CN}(*in situ*) at croplands (Table 1). Grasslands and Mixed landcover have largest absolute I_{Rnd} (0.20) than other landcovers, while shrublands has the largest proportion of I_{Rnd} to I_{Tot} (Table 1). The shrublands have the largest I_{Mod} in absolute value, while grasslands have the largest proportion of I_{Mod} to I_{Tot} (Table 1).

~~It is noticeable that there exists large information gaps (Fig. 2) between H_{CN}(*in situ*) and I(h,v; *in situ*) and I(h,v; *In situ*) and I(MDCA; *in situ*). H_{CN}(*in situ*) represent the amount of information that is required to fully characterize the "true "soil moisture, while I(h,v; *in situ*) indicates the available information contained in the system input variable about the "true" soil moisture. The information gap between H_{CN}(*in situ*) and I(MDCA; *in situ*) is the overall SMAP uncertainty in which 88% is contributed by the random uncertainty in the systems explanatory variables (Fig. 3). The information gap between I(h,v; *In situ*) and I(MDCA; *in situ*) represents the MDCA model uncertainty, which contributes 12% of the total uncertainty (Fig. 3).~~

3.2 ~~Model uncertainty and retrieval accuracy~~Informational uncertainties and retrieval quality

The relationship between different informational uncertainties and the Pearson correlation coefficients between *in situ* and SMAP DCA output, a commonly adopted relative model evaluation metric in SMAP studies (Chen et al., 2017; Colliander et al., 2017), was evaluated. ~~Figure 4 shows the relationship between the fraction of model uncertainty against different commonly adopted absolute (Fig. 4a) and relative model evaluation metrics (Fig. 4b). The model uncertainty is shown to be tightly related to these metrics.~~The I_{Tot}, I_{Mod} and I_{Rnd} are shown to be related how well the SMAP DCA soil moisture is correlated with *in situ* soil moisture (Fig. 5). I_{Tot} is found to be negatively correlated (r = -0.66, Fig. 5a) with the Pearson correlation between *in situ* soil moisture and SMAP DCA soil moisture. Similarly, I_{Mod} and I_{Rnd} are also shown to be negatively (-0.51 and -0.37 respectively) related to the Pearson correlation between *in situ* soil moisture and SMAP DCA soil moisture with I_{Mod} being more influential than I_{Rnd} (Fig. 5b - 5c). The negative relationship between SMAP DCA informational uncertainties are in line with general expectations since SMAP tends to capture more information about the *in situ* soil moisture when it retrieves high quality datasets. ~~It is observed that the fraction of MDCA induced uncertainty is positively correlated (r = 0.28) with RMSE of *in-situ* soil moisture and MDCA soil moisture (Fig. 4a). An obvious negative relationship is found when it comes to the relationship between the fraction of MDCA induced uncertainty and r of MDCA soil moisture and *in situ* soil moisture (r = -0.48). Both the positive and negative relationship are in line with general expectations since model uncertainty should go up when the retrieval accuracy is poor and vice versa.~~

3.3 Partial information decomposition of ~~M~~DCA

The partial information decompositions were assessed on a site basis and are shown in Figure 6. The fractional contribution of each component to that site's mutual information between brightness temperatures and DCA estimates, $I(\mathrm{T_{Bh}},\mathrm{T_{Bv}}$; DCA), was also calculated and are given in Table 2. Generally, the majority of $I(\mathrm{T_{Bh}},\mathrm{T_{Bv}}$; DCA) is redundantly ($R$) shared by $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$, which is about 0.08 (57% of $I(\mathrm{T_{Bh}},\mathrm{T_{Bv}}$; DCA)) on average (Table 2). The mean values of unique information of $\mathrm{T_{Bh}}$ (U_h) and synergistic information (S) of $\mathrm{T_{Bh}}$ and $\mathrm{T_{Bv}}$ are 0.026 (19% of $I(\mathrm{T_{Bh}},\mathrm{T_{Bv}}$; DCA)) and 0.019 (14% of $I(\mathrm{T_{Bh}},\mathrm{T_{Bv}}$; DCA)), respectively (Table 2). Compared to other decomposed information components, U_v is the smallest, but is of similar magnitude with *S*, with its mean being 0.013 (10% of $I(T_{Bh},T_{Bv}; DCA)$). Croplands have the highest *R* in absolute value 0.095 (68% of $I(T_{Bh},T_{Bv}; DCA)$), while mixed landcover has the highest fraction of *R* (74% of $I(T_{Bh},T_{Bv}; DCA)$) (Table 2). In general, the DCA system is mainly dominated by *R*. This indicates that both T_{Bh} and T_{Bv} provide similar information within the DCA. ~~Figure 5 illustrates that majority of the mutual information between T_{Bh}, T_{Bv} and MDCA ($I(T_{Bh},T_{Bv}; MDCA)$) is redundantly shared by T_{Bh} and T_{Bv}, which take about 0.55 of $I(T_{Bh},T_{Bv}; MDCA)$ on average (Fig. 5). U_h is comparable to S with a mean value of 0.15, respectively. Compared to other decomposed information components, U_v is the smallest but is of similar magnitude with U_h and S with mean being 0.14. Although the R is the largest information component, it has the smallest CV (0.35) compared to U_h (CV = 0.58), U_v (CV = 0.52) and S (CV = 0.63). In general, the MDCA system is dominated by R. This indicates that both T_{Bh} and T_{Bv} provide information regarding the soil moisture estimations, but these two variables are themselves highly dependent.~~

3.4 Partial information decomposition and retrieval accuracy

Through this analysis, it is shown (Fig. 7) that there are strong relationships between SMAP retrieval quality and decomposed information components. In general, the DCA tends to retrieve high quality soil moisture when U_h, U_v and S are low (Fig. 7a – Fig. 7c). This is demonstrated by a negative correlation of these component with the Pearson correlation between *in situ* and DCA soil moisture (Fig. 7a – Fig. 7c). In contrast, *R* shows the strongest positive correlation (Fig. 7d) with the relative model evaluation metric (r = 0.83). This indicates that *R* could potentially be a reference metric for DCA evaluation that does not require *in situ* and ancillary datasets. ~~Figure 6 shows the relationship between different decomposed information components and the RMSE of *in situ* and MDCA soil moisture. In general, only U_h is significantly negatively correlated (r = −0.28) with the RMSE of *in situ* and MDCA soil moisture (Fig. 6a), while relationships between RMSE and other components are not statistically significant (Fig. 6b – Fig. 6d). Figure 7 shows the relationship between different information components and the r of in situ and MDCA soil moisture. This demonstrates that all the information components are significantly correlated with the correlation, r, of *in situ* and MDCA soil moisture. U_h, U_v and S are negatively (Fig. 7a – Fig. 7c) correlated with r, while R is positively correlated with r. R shows the strongest correlation (Fig. 7d) with the relative model evaluation metric (r = 0.7). This indicates that R could potentially be a reference metric for MDCA evaluation. It does not require *in situ* soil moisture and shows a better performance than simply using r (Fig. 7d inset).~~

4 Discussion

4.1 ~~Random uncertainty and model uncertainty~~DCA informational uncertainty

The first objective of this study is to leverage information theory to quantitatively decompose the informational total uncertainty into informational random uncertainty and informational model uncertainty in the DCA as an approach to understand where retrieval errors arise.~~the overall uncertainty to random uncertainty and model uncertainty in the MDCA as an approach to understand where retrieval errors arise.~~ This information theory approach can add considerable power to SMAP modeling diagnosis. Mutual information can provide a way to unambiguously define the best achievement performance of a model ~~model performance~~ that is able to completely transform the available information to the desired target given a set of the input data.

~~In this study, any~~Any model based on the ~~M~~DCA model structure is a hypothesis that relates the input datasets ~~T_{Bh} and T_{Bv}~~ to soil moisture based on prior physical knowledge. The essence of the model is a procedure of processing the input dataset in order ~~T_{Bh} and T_{Bv}~~ to estimate ~~get~~ soil moisture. Thus, models, even the one performs the best, can only reduce the available information in its inputs and are not capable of adding new information about the “true” soil moisture. ~~The modeled soil moisture is deemed as an estimate of “true” soil moisture and a Markov chain is formed from T_{Bh}, T_{Bv} via MDCA soil moisture to *in situ* soil moisture. Any model, even the one performs the best, can only reduce the available information in its primary inputs (T_{Bh} and T_{Bv}) and is not capable of add new information about the “true” soil moisture.~~ Hence, there is no chance of building a model that is better than the one with the best achievable performance of the input data themselves ~~benchmark one~~ (yet even achieving this theoretically limit is nearly impossible) ~~if no freedom is given to the available datasets~~. If, however, ~~given~~ more freedom on available datasets to incorporate is given, it is possible to build models that outperform the aforementioned best achievable model performance by adding new explanatory variables which ~~will~~ may lead to a family of models that have completely different model structure. Based on Table 1, we found that the DCA has more informational uncertainty in shrublands than grasslands and croplands which is consistent with previous study (Zhang et al., 2019). This might be due to stronger variability in vegetation types for shrublands while grasslands and croplands tend to be more uniform and homogeneous. Furthermore, shrublands tend to be relatively less sensitive to changes in water availability while grasslands are more sensitive to the soil moisture dynamics in the condition of drought (Geruo et al., 2017). It is worth to noting that these finding are based on lumping our studied sites into different landcover categories, and results may be different while comparing two specific sites from different landcovers.

~~Additionally, the fraction that random uncertainty contributes to the overall uncertainty is quite significant (88% on average) in this study. The random uncertainty in the system may arises from the inherent error due to calibration of T_{Bh} and T_{Bv} in the locations and the presence water body. If poorly calibrated, the soil moisture estimations can be exacerbated due to the error propagation that hinders the correct information being transformed. Therefore, for example, a better and robust calibration strategy of T_{Bh} and T_{Bv} to the presence of water body might need. Furthermore, a better quality-control method or additional data screening metric with respect to water corrected T_{Bh} and T_{Bv} is also required to further reduce the random uncertainty.~~

The fraction that informational random uncertainty contributes to the informational total uncertainty is quite significant (64% on average) in this study. The informational random uncertainty in the system may arises from the inherent error due to calibration of T_{Bh} and T_{Bv} in the locations, the mismatch in the scale of observations, and the presence water bodies. If poorly calibrated, the soil moisture estimations can be exacerbated due to the error propagation that hinders the correct information being expressed. Furthermore, SMAP attempts to use the T_{eff} to capture both soil and canopy temperature because the differences between canopy and soil temperature are minimized in the morning and dawn. The T_{eff} is computed based on a model that uses the information from average soil temperature of first layer (5cm -15cm) and second layer (15cm - 35cm) and interpolated in time in order to match SMAP morning and dawn observations (O’Neill et al., 2020). These interpolation and modeling processes may produce erroneous T_{eff} dataset and hence contribute the informational random uncertainty of DCA. Therefore, a better and robust calibration strategy of T_{Bh} and T_{Bv} to the presence of water bodies and a comprehensive assessment of T_{eff} may be needed to reduce some of the information random uncertainty.

~~Apart from random uncertainty, the model uncertainty contribution is also a significant amount the total (12% on average). This model uncertainty may arise from poor model parameterizations. It's known that the '*tau-omega*' model in MDCA is parameterized by landcover based parameters. The values of these parameters are derived from past studies, past experience and some information discussions with subject matter experts, which could be biased and inaccurate (Peggy O'Neill et al., 2018). In addition, these parameter values are differentiated by landcover and do not vary in time and microwave polarization directions. In fact, these parameters may not vary in short time (days or weeks) but could vary from a long-term perspective (month or years) and the parameter associated with vegetation structure may vary correspondent to different phenology phases.~~ Informational model uncertainty contributes an unneglectable portion to the informational total uncertainty (36% on average). This model uncertainty may arise from poor model parameterizations. It's known that DCA is parameterized with a set of surface and vegetation parameters such as vegetation single scattering albedo (ω), surface height standard deviation s, etc. These parameter values are landcover dependent are derived from past studies as well as prior experience and some information discussions with experts, all of which could be biased and inaccurate (O'Neill et al., 2020). These parameter values also are differentiated by landcover microwave polarization directions (Wigneron et al., 2004) and were assumed to be constant in time. It is possible that these parameters (such as ω) vary in time and shift during senescence or harvesting seasons (Konings et al., 2017). It is observed that the proportion of the informational model uncertainty is slightly smaller in shrublands (Table 1), while these proportions are larger in croplands and grasslands (Table 1). This might because the model parameterizations are more reasonable in shrublands than other landcovers. In addition, croplands and grasslands may have seasonal harvesting and therefore may more subject to changes in these values, while shrublands may not.

To summarize, this is the first attempt of leveraging mutual information approach to quantitively analyze the uncertainty components in microwave remote sensing models. The results of this study can be further used as a foundation guidance of SMAP algorithm assessing approach that can quantitively identify where information lost in the process of SMAP soil moisture modeling. This analysis, though focused on ~~M~~DCA soil moisture, can be transferred and extended to analyze ~~any~~ other remote sensing algorithms.~~models.~~

4.2 Model evaluation from another perspective

The second objective of this study was to ~~was to~~ demonstrate ~~is~~ that partitioned information components can be used as a new ~~M~~DCA model evaluation metric that does not depend on *in situ* soil moisture and other ancillary datasets. We found a strong linear relationship between redundant information (*R*) of the polarized brightness temperatures and Pearson correlation between ~~*r* of M~~DCA and *in situ* soil moisture, which indicated that T_{Bh} and T_{Bv} are highly dependent. *R* is also the dominant component relative to others quantified here. In general, it is more likely to observe higher *R* in the less woody landcovers (croplands and grasslands) where the range of brightness temperature may possible be greater. From an information perspective, higher or complete *R* indicates that one variable is a function of the other, or they share the same source. T_{Bv} and T_{Bh} are known to be highly correlated. It's important to note that the decomposed information component *R* is dependent on the DCA parameterizations that determines how strong the T_{Bh} and T_{Bv} are linked with the DCA. This stronger linkage is indicated by a higher value of *R* relative to other components. ~~From an information perspective, higher or complete *R* indicates that one source variable is a function of the other, or they share the same source. It can be observed from Figure 7d (inset) that there is~~

~~a strong linear relationship between TB$_v$ and TB$_h$ ($r \geq 0.94$). Therefore, it is expected that a higher redundancy in the MDCA system. The MDCA takes T_{Bv} and T_{Bh} as primary inputs while T_{Bh} and T_{Bv} share a lot of redundancy. Therefore, it is not surprising that the MDCA soil moisture underperforms SMAP SCA soil moisture due to the error accumulation and error propagation from both channels.~~

We found that DCA model performance, as characterized by the correlation between Person correlation between DCA and *in situ* soil moisture, improves with larger values of *R* that T_{Bh} and T_{Bv} share with DCA estimates. Given the strength of this relationship, the *R* could be potentially used as a DCA evaluation metric that doesn't depend on *in situ* measurement and ancillary dataset. It is also useful for SMAP DCA soil moisture users to have a rough estimation of how high the quality of the obtained DCA soil moisture without actually knowing the *in situ* soil moisture. However, this depends on specific user requirements for data quality. In general, the DCA soil moisture tends to be in high end in term retrieval quality (~ 0.75 in Pearson correlation) when the *R* is greater 0.1.

~~To summarize, the redundant information shown a strong correlation with *r*, which could be potentially used as a MDCA evaluation metric. This metric only involves T_{Bh}, T_{Bv} and MDCA soil moisture and doesn't depend on *in situ* measurement and ancillary dataset. Compared to another *in situ* independent metric, such as pearson *r* of T_{Bh} and T_{Bv}, it shows a better performance (0.70 vs 0.52). This is potentially due to numerous non-linear processes acting within the MDCA, which are not well captured by linear metrics such as the pearson *r* of T_{Bh} and T_{Bv}.~~

5 ~~Conclusion and~~ Approach Limitations

While we expect that this approach can be generalized to analyze other remote sensing models, it may be difficult to compute the joint probability density functions for models with high-dimensional inputs. Difficulty in determining the joint probability density functions hinders the estimation of high dimensional joint entropy and mutual information components. Although there exist serval data dimension reduction techniques, these dimension reduction techniques are mostly based some assumptions (Xu et al., 2019). In practice, most of the systems with high dimension inputs tend to be complex. Therefore, there is a strong risk of introducing additional uncertainty if one chooses an inappropriate technique.

This study was conducted only at locations where *in situ* soil moisture is readily available. The problem of how to leverage information theory to evaluate the error components in the locations without *in situ* soil moisture measurements is challenging and could be an interesting topic for future works. Finally, we would expect the informational uncertainty analysis to provide asymptotic estimation of random and model uncertainties. The best performance we can expect from this current uncertainty analysis is to use all of the available datasets we have; yet we believe that uncertainty estimations of this approach should be stabilized given adequate representative locations and data records.

~~This study attempts to differentiate and quantify the uncertainty sources in MDCA using information theoretic. We found that on average 88% of the uncertainty is contributed by the inadequacy of explanatory variables of SMAP or uncertainties in the estimated brightness temperature, while the rest of the uncertainty is induced by inaccurate MDCA parameterizations. The fraction of the model uncertainty to the overall uncertainty is negatively correlated with the pearson *r* of *in situ* and MDCA~~

~~soil moisture ($r = -0.48$) while positively correlated with the error between in situ and MDCA soil moisture ($r = 0.28$). The decomposition of the mutual information has shown that all decomposed components are correlated with the pearson *r* between in situ and MDCA soil moisture with the redundant information being the tightest ($r = 0.7$). The uncertainty decomposition analysis opens a new window for SMAP algorithm uncertainty diagnosis. The result of mutual information decomposition analysis can be adopted as a new *in situ* independent SMAP soil moisture evaluation reference metric.~~

~~We acknowledge the existence of limitations of this study. First, we expect that this approach can be generalized to analyze other remote sensing models. However, it may be difficult to compute the joint probability density function for models with high-dimensional inputs, and thus also difficult to estimate the joint entropy and mutual information components. Though there exist several approaches for computing joint entropy and mutual information, the caveat here is that it is not guaranteed that the estimated mutual information can be exactly the entropy and joint entropy that fulfils the equality of, for instance, equation (5). Second, this study was conducted at locations where *in situ* soil moisture readily available. The problem of how to leverage information theory to evaluate the error components in the locations without *in situ* soil moisture measurements is challenging and could be an interesting topic for future works. Third, we would expect that the information theoretic to provide asymptotic estimation of random and model uncertainties, the best performance we can expect from this current uncertainty analysis is to use all of the available datasets we have; yet we believe that uncertainty estimations of this approach should be stabilized given adequate representative locations and data records.~~

Code availability

The code regarding the SMAP dataset ~~s~~ time series ~~extraction~~, mutual information and partial information decomposition calculation can be obtained from https://github.com/libonancaesar/HESS_Information_Uncertainty ~~is upon request.~~

Data availability

SMAP Enhanced Level-2 Radiometer Half-Orbit ~~9~~ 36 km EASE-Grid Soil Moisture, Version ~~3~~ 7 is acquired from US National Snow and Ice Data Center (https://nsidc.org/data/smap). The *in situ* soil moisture is accessible through U.S. Climate Reference Network (https://www.ncdc.noaa.gov/crn/).

Author contribution

Bonan Li: conceptualization; data acquisition; formal analysis; methodology; original draft writing, editing; ***Stephen P. Good***: conceptualization; methodology; draft writing, editing, revisions; supervision.

Competing interests

The authors declare no conflicts of interest.

Acknowledgments

This project was supported by The National Aeronautics and Space Administration under grant NNX16AN13G.

References

Babaeian, E., Sadeghi, M., Jones, S. B., Montzka, C., Vereecken, H. and Tuller, M.: Ground, Proximal, and Satellite Remote Sensing of Soil Moisture, Rev. Geophys., 57(2), 530–616, doi:10.1029/2018RG000618, 2019.

Bassiouni, M., Good, S. P., Still, C. J. and Higgins, C. W.: Plant Water Uptake Thresholds Inferred From Satellite Soil Moisture, Geophys. Res. Lett., 47(7), doi:10.1029/2020GL087077, 2020.

Bell, J. E., Palecki, M. A., Baker, C. B., Collins, W. G., Lawrimore, J. H., Leeper, R. D., Hall, M. E., Kochendorfer, J., Meyers, T. P., Wilson, T. and Diamond, H. J.: U.S. Climate Reference Network Soil Moisture and Temperature Observations, J. Hydrometeorol., 14(3), 977–988, doi:10.1175/JHM-D-12-0146.1, 2013.

Chaubell, J., Yueh, S., Chan, S., Dunbar, S., Colliander, A., Entekhabi, D. and Chen, F.: Smap Regularized Dual-Channel Algorithm for the Retrieval of Soil Moisture and Vegetation Optical Depth, in IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 5312–5315, IEEE., 2019.

Chaubell, M. J., Yueh, S. H., Scott Dunbar, R., Colliander, A., Chen, F., Chan, S. K., Entekhabi, D., Bindlish, R., O’Neill, P. E., Asanuma, J., Berg, A. A., Bosch, D. D., Caldwell, T., Cosh, M. H., Collins, C. H., Martinez-Fernandez, J., Seyfried, M., Starks, P. J., Su, Z., Thibeault, M. and Walker, J.: Improved SMAP Dual-Channel Algorithm for the Retrieval of Soil Moisture, IEEE Trans. Geosci. Remote Sens., 58(6), 3894–3905, doi:10.1109/TGRS.2019.2959239, 2020.

Chen, F., Crow, W. T., Colliander, A., Cosh, M. H., Jackson, T. J., Bindlish, R., Reichle, R. H., Chan, S. K., Bosch, D. D., Starks, P. J., Goodrich, D. C. and Seyfried, M. S.: Application of Triple Collocation in Ground-Based Validation of Soil Moisture Active/Passive (SMAP) Level 2 Data Products, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 10(2), 489–502, doi:10.1109/JSTARS.2016.2569998, 2017.

Chen, F., Crow, W. T., Bindlish, R., Colliander, A., Burgin, M. S., Asanuma, J. and Aida, K.: Global-scale evaluation of SMAP, SMOS and ASCAT soil moisture products using triple collocation, Remote Sens. Environ., 214(March), 1–13, doi:10.1016/j.rse.2018.05.008, 2018.

Colliander, A., Jackson, T. J., Bindlish, R., Chan, S., Das, N., Kim, S. B., Cosh, M. H., Dunbar, R. S., Dang, L., Pashaian, L., Asanuma, J., Aida, K., Berg, A., Rowlandson, T., Bosch, D., Caldwell, T., Caylor, K., Goodrich, D., al Jassar, H., Lopez-Baeza, E., Martínez-Fernández, J., González-Zamora, A., Livingston, S., McNairn, H., Pacheco, A., Moghaddam, M., Montzka, C., Notarnicola, C., Niedrist, G., Pellarin, T., Prueger, J., Pulliainen, J., Rautiainen, K., Ramos, J., Seyfried, M., Starks, P., Su, Z., Zeng, Y., van der Velde, R., Thibeault, M., Dorigo, W., Vreugdenhil, M., Walker, J. P., Wu, X., Monerris, A., O’Neill, P. E., Entekhabi, D., Njoku, E. G. and Yueh, S.: Validation of SMAP surface soil moisture products with core validation sites, Remote Sens. Environ., 191, 215–231, doi:10.1016/j.rse.2017.01.021, 2017.

Cover, T. M. and Thomas, J. A.: Elements of Information Theory, Wiley., 2005.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D. and Walker, J. P.: Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, Rev. Geophys., 50(2), 634, doi:10.1029/2011RG000372, 2012.

Freedman, D. and Diaconis, P.: On the histogram as a density estimator:L 2 theory, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57(4), 453–476, doi:10.1007/BF01025868, 1981.

Gao, L., Sadeghi, M., Ebtehaj, A. and Wigneron, J.-P.: A temporal polarization ratio algorithm for calibration-free retrieval of soil moisture at L-band, Remote Sens. Environ., 249, 112019, doi:10.1016/j.rse.2020.112019, 2020.

Geruo, A., Velicogna, I., Kimball, J. S., Du, J., Kim, Y., Colliander, A. and Njoku, E.: Satellite-observed changes in vegetation sensitivities to surface soil moisture and total water storage variations since the 2011 Texas drought, Environ. Res. Lett., 12(5), doi:10.1088/1748-9326/aa6965, 2017.

Gong, W., Gupta, H. V., Yang, D., Sricharan, K. and Hero, A. O.: Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach, Water Resour. Res., 49(4), 2253–2273, doi:10.1002/wrcr.20161, 2013.

Goodwell, A. E. and Kumar, P.: Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables, Water Resour. Res., 53(7), 5920–5942, doi:10.1002/2016WR020216, 2017.

Gupta, H. V., Sorooshian, S. and Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34(4), 751–763, doi:10.1029/97WR03495, 1998.

Jackson, T. J., Bindlish, R., Cosh, M. H., Zhao, T., Starks, P. J., Bosch, D. D., Seyfried, M., Moran, M. S., Goodrich, D. C., Kerr, Y. H. and Leroux, D.: Validation of Soil Moisture and Ocean Salinity (SMOS) Soil Moisture Over Watershed Networks in the U.S., IEEE Trans. Geosci. Remote Sens., 50(5), 1530–1543, doi:10.1109/TGRS.2011.2168533, 2012.

Jadidoleslam, N., Mantilla, R., Krajewski, W. F. and Goska, R.: Investigating the role of antecedent SMAP satellite soil moisture, radar rainfall and MODIS vegetation on runoff production in an agricultural region, J. Hydrol., 579, 124210, doi:10.1016/j.jhydrol.2019.124210, 2019.

Konings, A. G., McColl, K. A., Piles, M. and Entekhabi, D.: How Many Parameters Can Be Maximally Estimated From a Set of Measurements?, IEEE Geosci. Remote Sens. Lett., 12(5), 1081–1085, doi:10.1109/LGRS.2014.2381641, 2015.

Konings, A. G., Piles, M., Das, N. and Entekhabi, D.: L-band vegetation optical depth and effective scattering albedo estimation from SMAP, Remote Sens. Environ., 198, 460–470, doi:10.1016/j.rse.2017.06.037, 2017.

Mishra, A., Vu, T., Veettil, A. V. and Entekhabi, D.: Drought monitoring with soil moisture active passive (SMAP) measurements, J. Hydrol., 552(January 2015), 620–632, doi:10.1016/j.jhydrol.2017.07.033, 2017.

Njoku, E. G. and Entekhabi, D.: Passive microwave remote sensing of soil moisture, J. Hydrol., 184(1–2), 101–129, doi:10.1016/0022-1694(95)02970-2, 1996.

O'Neill, P. E., S. Chan, E. G. Njoku, T. Jackson, R. Bindlish, and J. C.: SMAP L2 Radiometer Half-Orbit 36 km EASE-Grid Soil Moisture, Version 7. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center,~~,~~ doi:https://doi.org/10.5067/F1TZ0CBN1F5N, 2020.

Paninski, L.: Estimation of Entropy and Mutual Information, Neural Comput., 15(6), 1191–1253, doi:10.1162/089976603321780272, 2003.

~~Peggy O'Neill, Rajat Bindlish, Steven Chan, Eni Njoku and Tom Jackson: Soil Moisture Active Passive (SMAP) Algorithm Theoretical Basis Document SMAP L2 & L3 Radar Soil Moisture (Active) Data Products, Jet Propuls. Lab., Calif. Inst. Technol., Pasadena, CA, USA, JPL D-66480, 2018.~~

Peters, G.-J. Y. and Kok, G.: All models are wrong, but some are useful: a comment on Ogden (2016), Health Psychol. Rev., 10(3), 265–268, doi:10.1080/17437199.2016.1190658, 2016.

Petropoulos, G. P., Ireland, G. and Barrett, B.: Surface soil moisture retrievals from remote sensing: Current status, products & future trends, Phys. Chem. Earth, 83–84, 36–56, doi:10.1016/j.pce.2015.02.009, 2015.

Piles, M., Entekhabi, D., Konings, A. G., McColl, K. A., Das, N. N. and Jagdhuber, T.: Multi-temporal microwave retrievals of Soil Moisture and vegetation parameters from SMAP, in 2016 IEEE International Geoscience and Remote Sensing

Symposium (IGARSS), pp. 242–245, IEEE., 2016.

Seitzinger, S. P., Gaffney, O., Brasseur, G., Broadgate, W., Ciais, P., Claussen, M., Erisman, J. W., Kiefer, T., Lancelot, C., Monks, P. S., Smyth, K., Syvitski, J. and Uematsu, M.: International Geosphere–Biosphere Programme and Earth system science: Three decades of co-evolution, Anthropocene, 12, 3–16, doi:10.1016/j.ancene.2016.01.001, 2015.

Shannon, C. E.: A Mathematical Theory of Communication, Bell Syst. Tech. J., 27(3), 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x, 1948.

Shellito, P. J., Small, E. E., Colliander, A., Bindlish, R., Cosh, M. H., Berg, A. A., Bosch, D. D., Caldwell, T. G., Goodrich, D. C., McNairn, H., Prueger, J. H., Starks, P. J., van der Velde, R. and Walker, J. P.: SMAP soil moisture drying more rapid than observed in situ following rainfall events, Geophys. Res. Lett., 43(15), 8068–8075, doi:10.1002/2016GL069946, 2016.

Uber, M., Vandervaere, J.-P., Zin, I., Braud, I., Heistermann, M., Legoût, C., Molinié, G. and Nord, G.: How does initial soil moisture influence the hydrological response? A case study from southern France, Hydrol. Earth Syst. Sci., 22(12), 6127–6146, doi:10.5194/hess-22-6127-2018, 2018.

Wigneron, J.-P., Parde, M., Waldteufel, P., Chanzy, A., Kerr, Y., Schmidl, S. and Skou, N.: Characterizing the Dependence of Vegetation Model Parameters on Crop Structure, Incidence Angle, and Polarization at L-Band, IEEE Trans. Geosci. Remote Sens., 42(2), 416–425, doi:10.1109/TGRS.2003.817976, 2004.

Williams, P. L. and Beer, R. D.: Nonnegative Decomposition of Multivariate Information, , 1–14 [online] Available from: http://arxiv.org/abs/1004.2515, 2010.

Xu, X., Liang, T., Zhu, J., Zheng, D. and Sun, T.: Review of classical dimensionality reduction and sample selection methods for large-scale data processing, Neurocomputing, 328, 5–15, doi:10.1016/j.neucom.2018.02.100, 2019.

Zhang, R., Kim, S. and Sharma, A.: A comprehensive validation of the SMAP Enhanced Level-3 Soil Moisture product using ground measurements over varied climates and landscapes, Remote Sens. Environ., 223, 82–94, doi:10.1016/j.rse.2019.01.015, 2019.

Zhang, Z. and Grabchak, M.: Bias Adjustment for a Nonparametric Entropy Estimator, Entropy, 15(12), 1999–2011, doi:10.3390/e15061999, 2013.

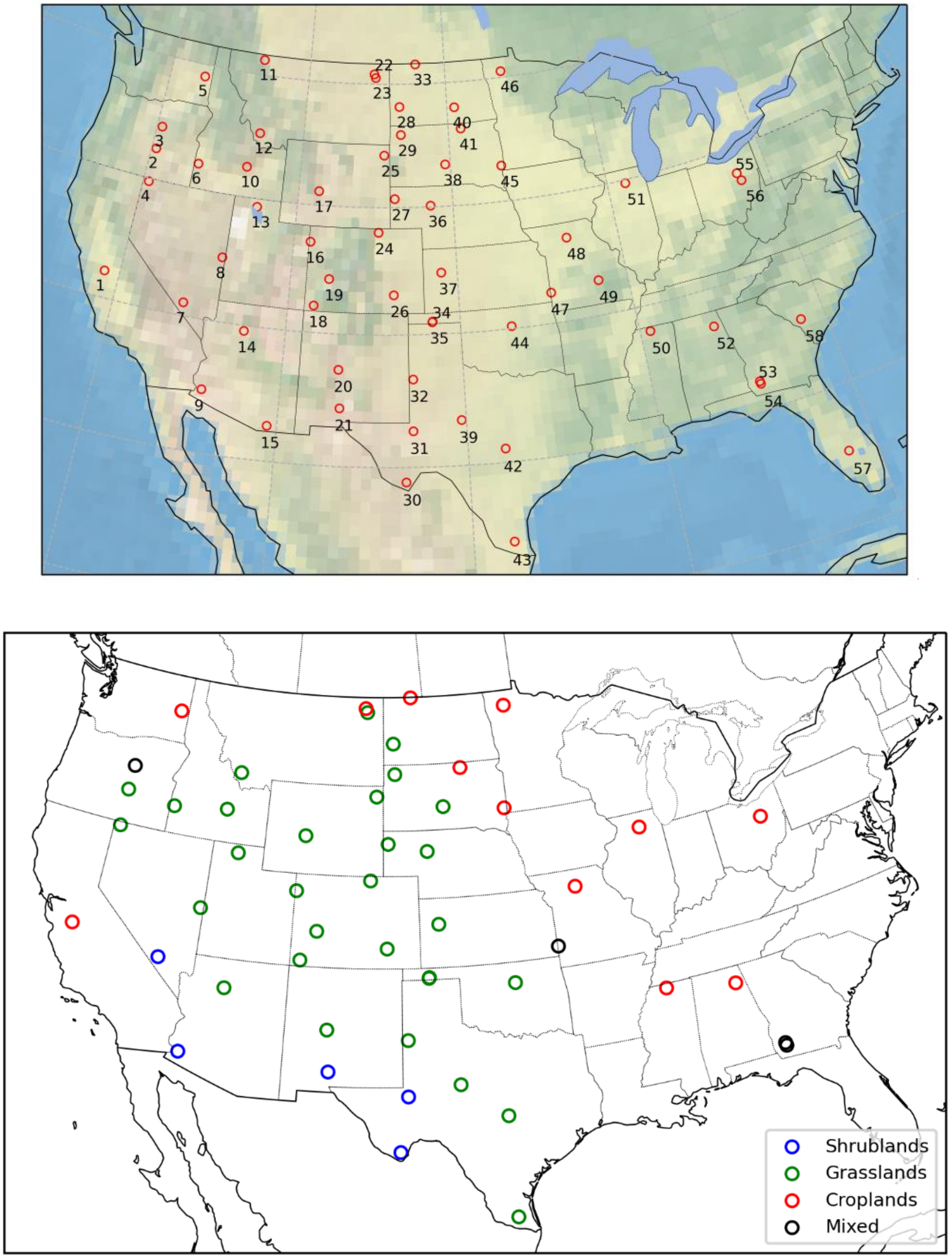
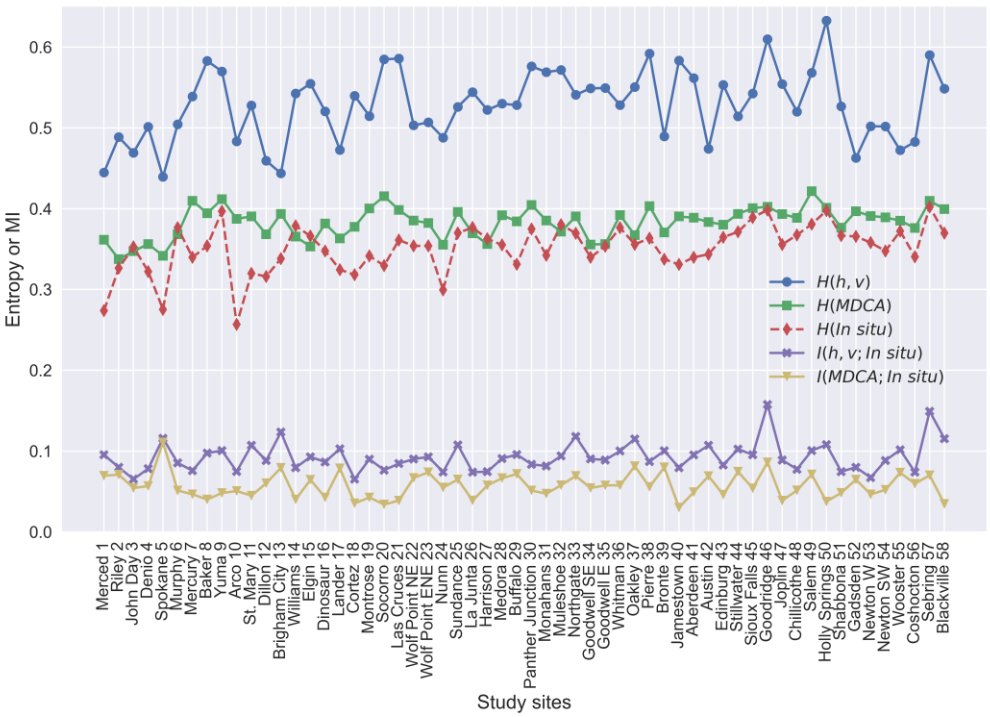


Figure 1 Spatial distribution of selected USCRN stations classified by landcovers~~from west to east. See figure 2 caption for names of individual sites based on numbering.~~


Entropy or MI
0.0
0.1
0.2
0.3
0.4
0.5
0.6
$H(h, v)$
$H(MDCA)$
$H(In\ situ)$
$I(h, v; In\ situ)$
$I(MDCA; In\ situ)$
Merced 1
Riley 2
John Day 3
Denio 4
Spokane 5
Murphy 6
Mercury 7
Baker 8
Yuma 9
Arco 10
St. Mary 11
Dillon 12
Brigham City 13
Williams 14
Elgin 15
Dinosaur 16
Lander 17
Cortez 18
Montrose 19
Socorro 20
Las Cruces 21
Wolf Point NE 22
Wolf Point ENE 23
Nunn 24
Sundance 25
La Junta 26
Harrison 27
Medora 28
Buffalo 29
Panther Junction 30
Monahans 31
Muleshoe 32
Northgate 33
Goodwell SE 34
Goodwell E 35
Whitman 36
Oakley 37
Pierre 38
Bronte 39
Jamestown 40
Aberdeen 41
Austin 42
Edinburg 43
Stillwater 44
Sioux Falls 45
Goodridge 46
Joplin 47
Chillicothe 48
Salem 49
Holly Springs 50
Shabbona 51
Gadsden 52
Newton W 53
Newton SW 54
Wooster 55
Coshocton 56
Sebring 57
Blackville 58
Study sites

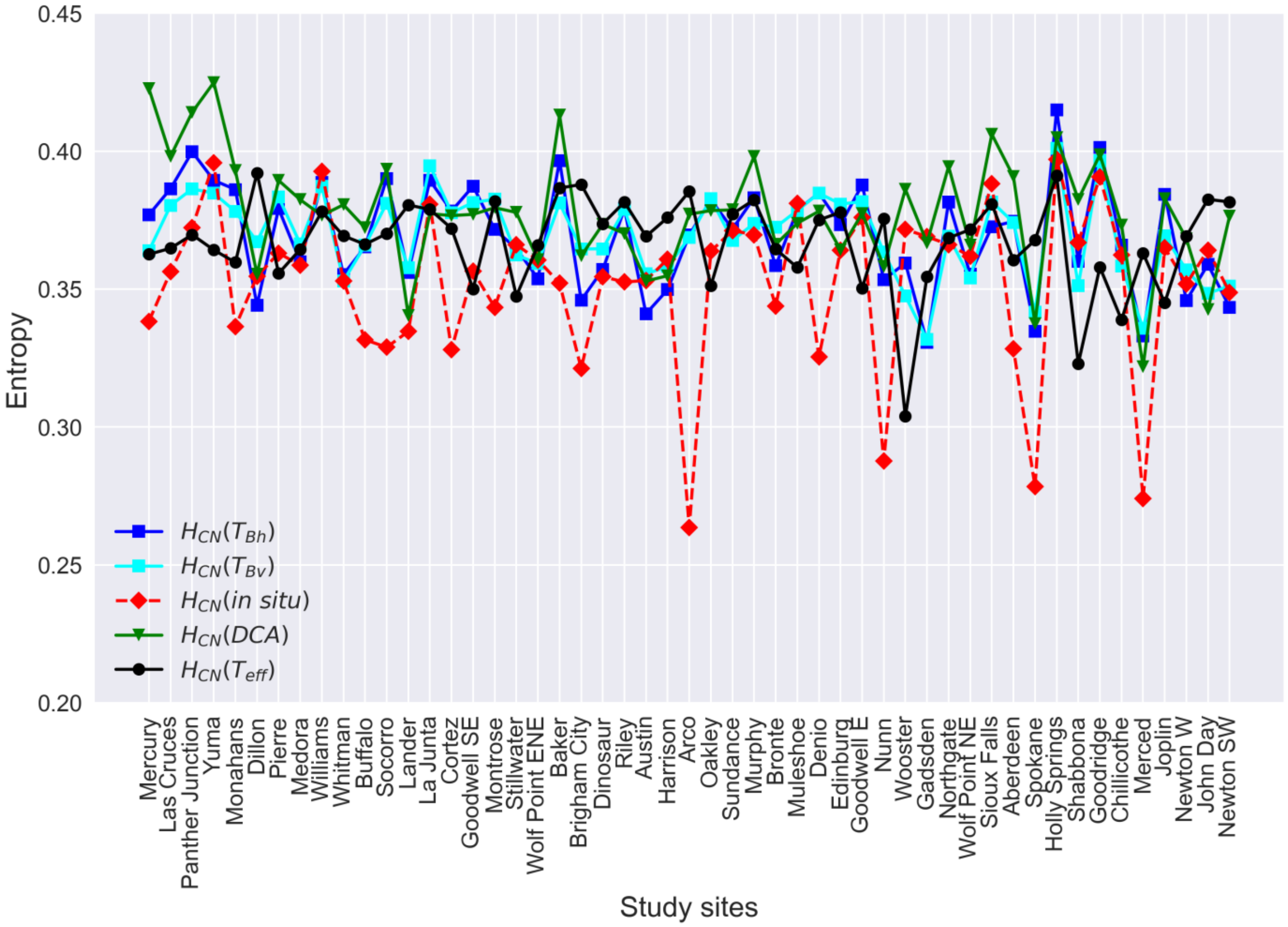


Figure 2 Entropies of *in situ* soil moisture, horizontally polarized brightness temperature (T_{Bh}), vertically polarized brightness temperature (T_{Bv}), soil effective temperature (T_{eff}) and DCA soil moisture across the study sites.

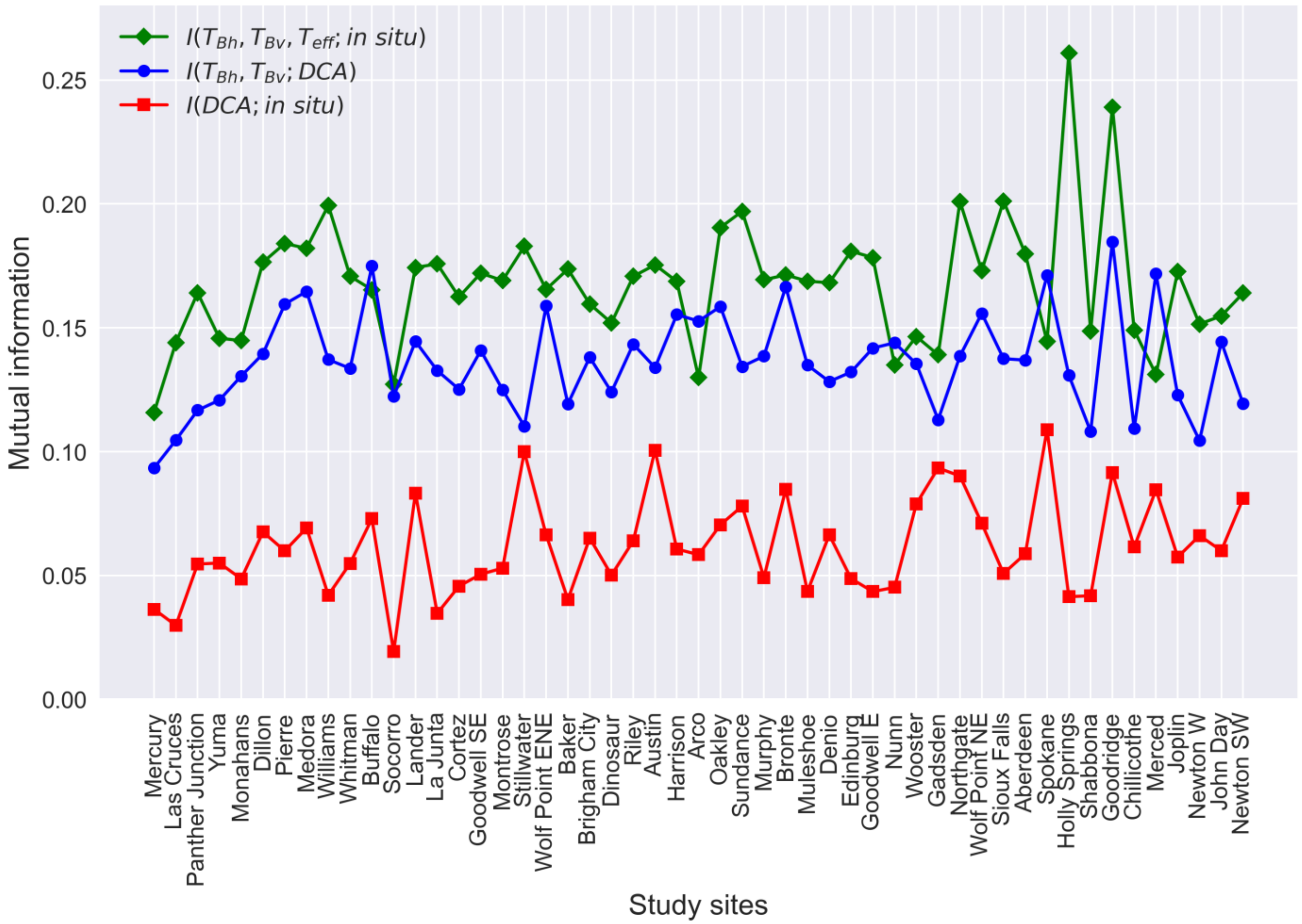
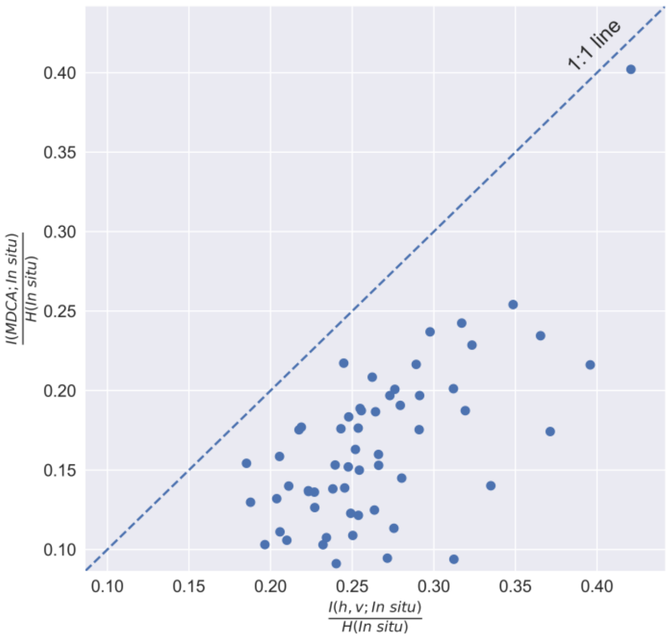


Figure 3 Mutual information between horizontally polarized brightness temperature (T_{Bh}), vertically polarized brightness temperature (T_{Bv}), soil effective temperature (T_{eff}) and *in situ* soil moisture, mutual information between horizontally polarized brightness temperature (T_{Bh}), vertically polarized brightness temperature (T_{Bv}) and DCA soil moisture, mutual information between DCA soil moisture and *in situ* soil moisture.

~~**Information quantities of *in situ* soil moisture, T_{Bh}, T_{Bv} and MDCA soil moisture across the study sites.**~~



~~Figure 3 Mutual information between MDCA soil moisture and *in situ* soil moisture against mutual information between T_{Bh}, T_{Bv} and *in situ* soil moisture.~~

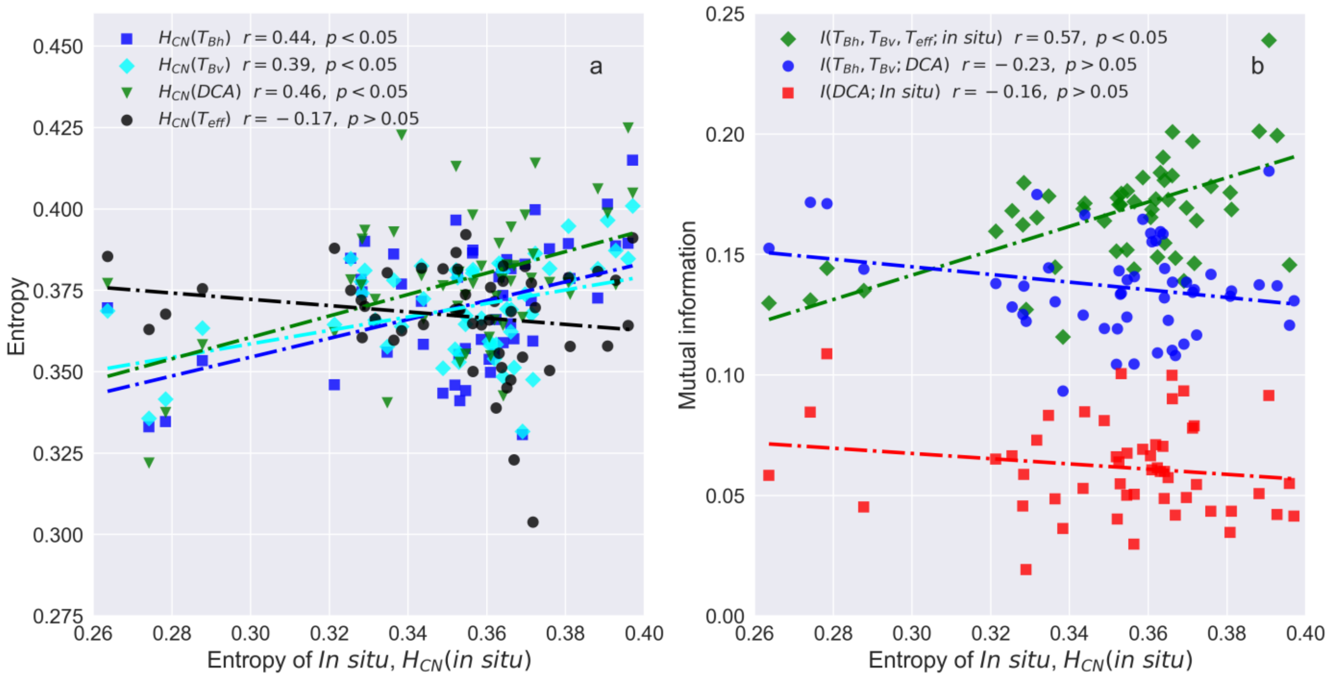


Figure 4 Entropy of *in situ* soil moisture against the entropies of DCA soil moisture, horizontally polarized brightness temperature (T_{Bh}), vertically polarized brightness temperature (T_{Bv}) and soil effective temperature (T_{eff}) (a) and mutual information quantities (b).

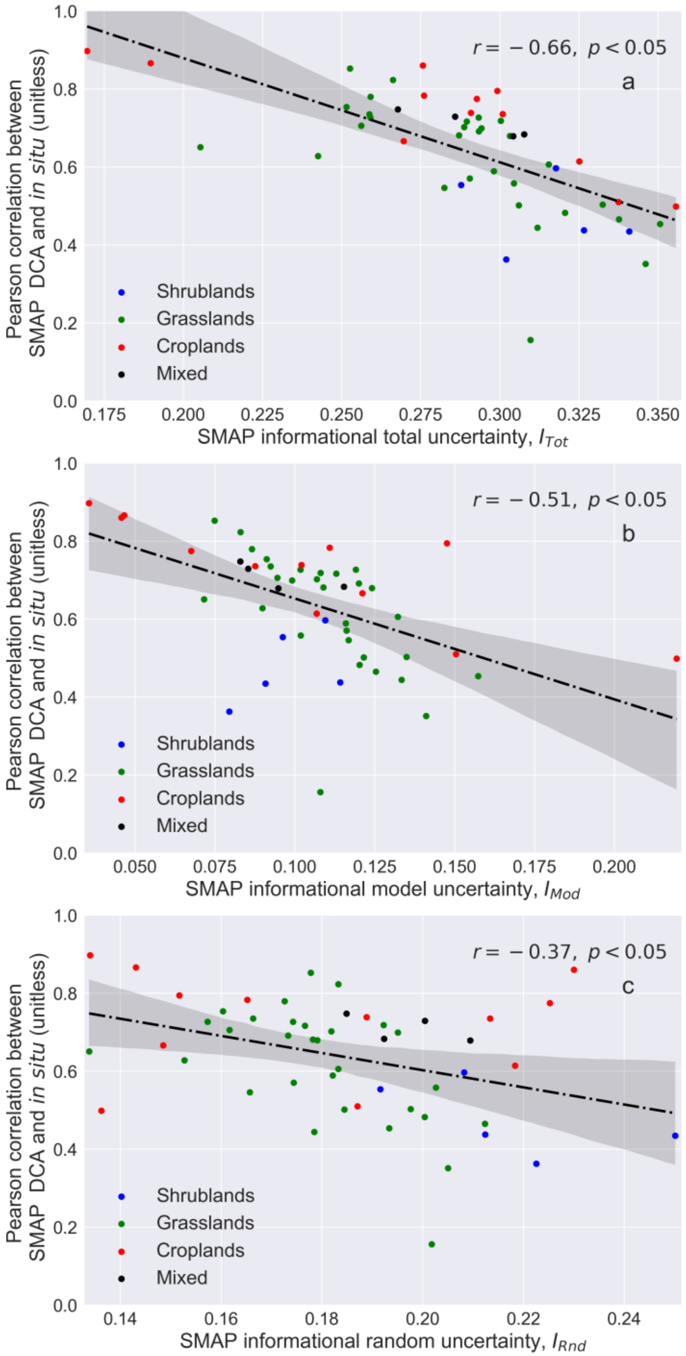


Figure 5 SMAP informational total uncertainty (a), SMAP informational model uncertainty (b) and SMAP informational random uncertainty against Pearson correlation between *in situ* soil moisture and DCA soil moisture. 780

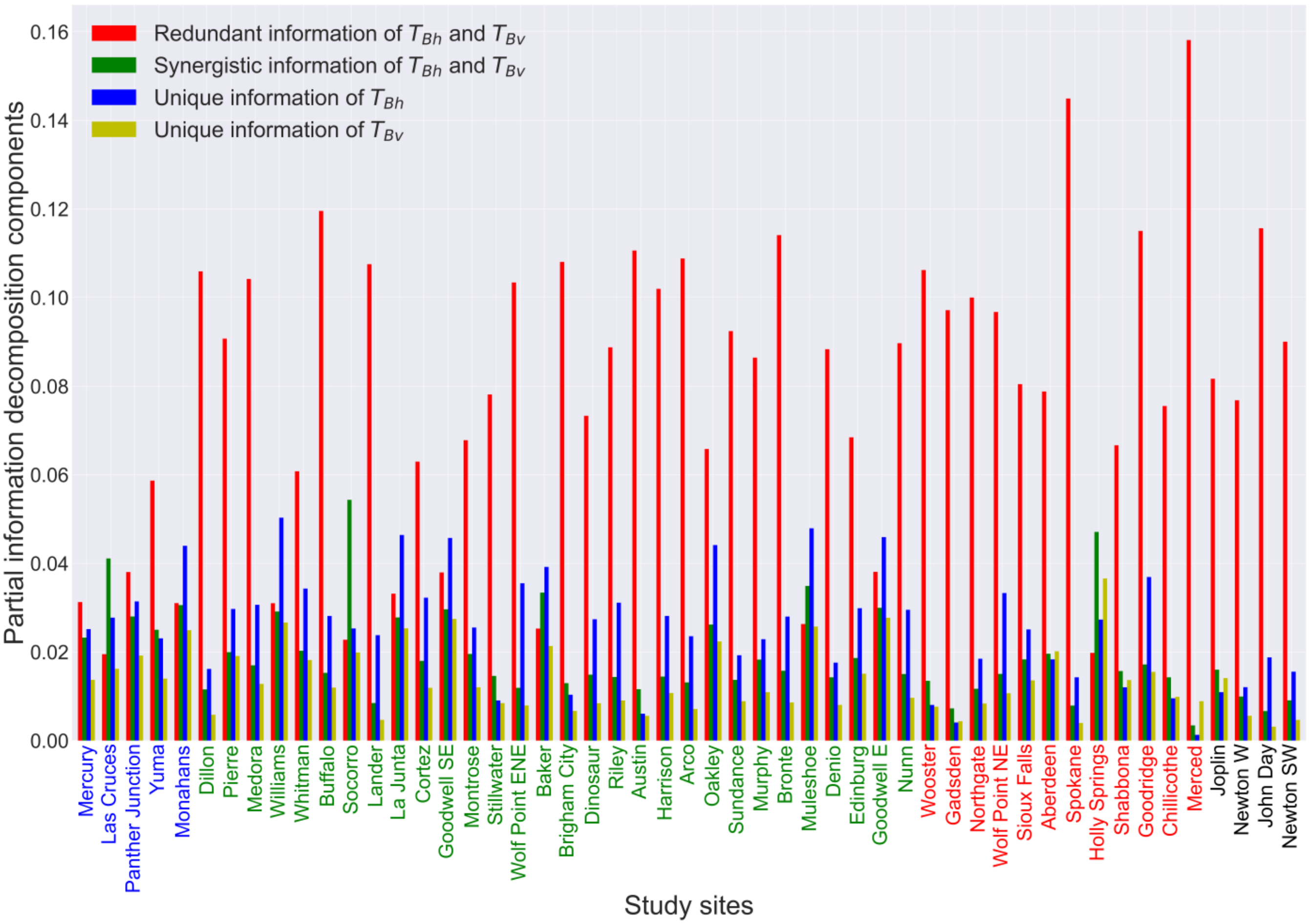
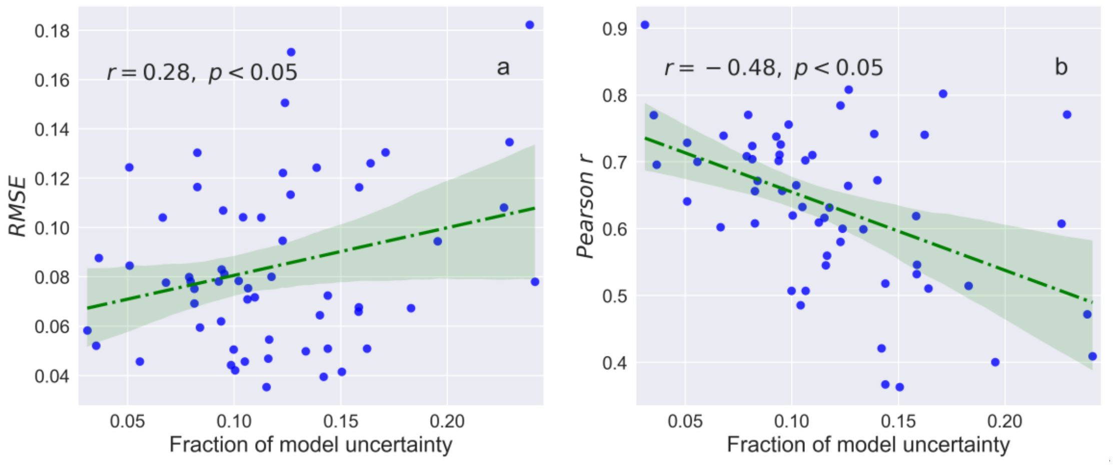
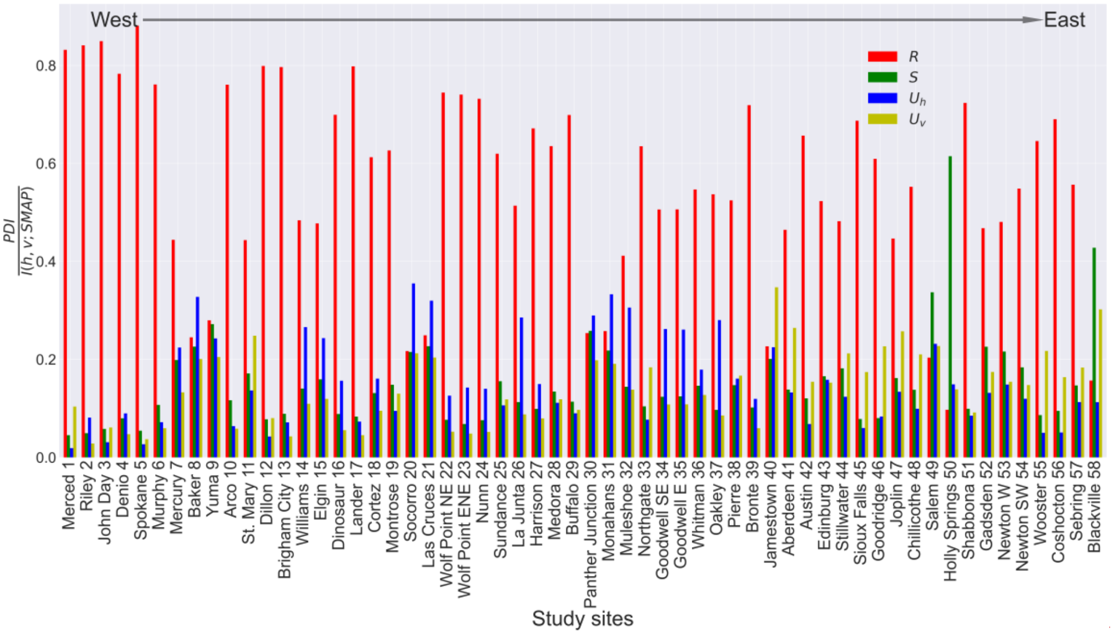


Figure 6 Partial information decomposition components between horizontally (T_{Bh}) and vertically (T_{Bv}) polarized brightness temperature and DCA soil moisture. The colored labels of the horizontal axis represent different landcover of the study sites (blue: Shrublands, green: Grasslands, red: Croplands, black: Mixed).



~~**Figure 4 Fraction of MDCA model uncertainty against RMSE of MDCA soil moisture and *in situ* soil moisture (a) and fraction of MDCA model uncertainty against pearson *r* of MDCA soil moisture and *in situ* soil moisture (b).**~~



~~**Figure 5 The normalized partial information decomposition components between T_{Bh}, T_{Bv} and MDCA soil moisture.**~~

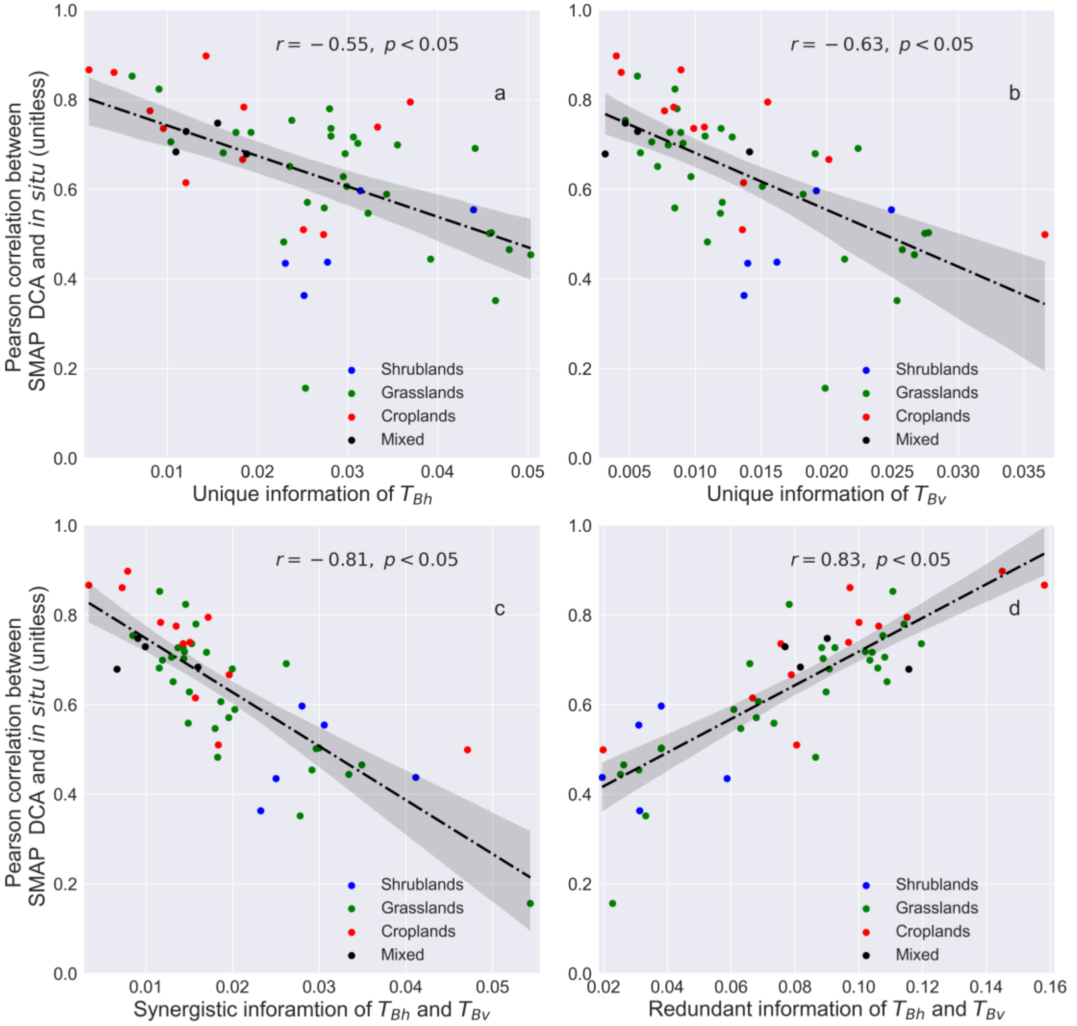


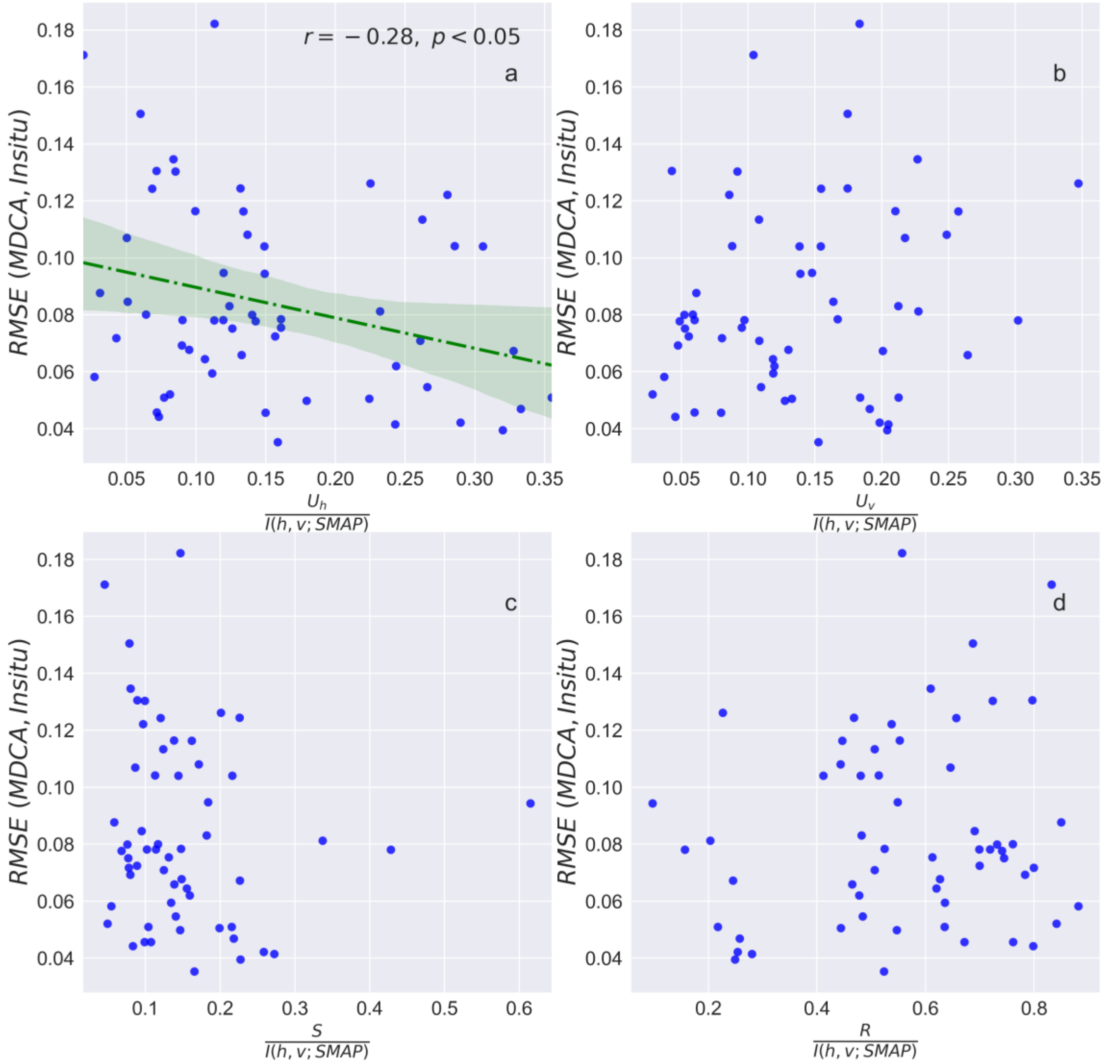
Figure 7 Partial information decomposition components between horizontally (T_{Bh}) and vertically (T_{Bv}) polarized brightness temperature against Pearson correlation coefficient between *in situ* and DCA soil moisture.

Landcover	Informational random uncertainty, I_{Rnd} (and its % of I_{Tot})	Informational model Uncertainty, I_{Mod} (and its % of I_{Tot})	Informational total uncertainty, I_{Tot} (and its % of H_{CN}(*in situ*))
Shrublands	0.22 (69%)	0.10 (31%)	0.32 (88%)
Grasslands	0.20 (62%)	0.09 (38%)	0.29 (83%)
Croplands	0.18 (65%)	0.10 (35%)	0.28 (79%)
Mixed	0.20 (68%)	0.09 (32%)	0.29 (81%)
Overall	0.18 (64%)	0.11 (36%)	0.29 (82%)

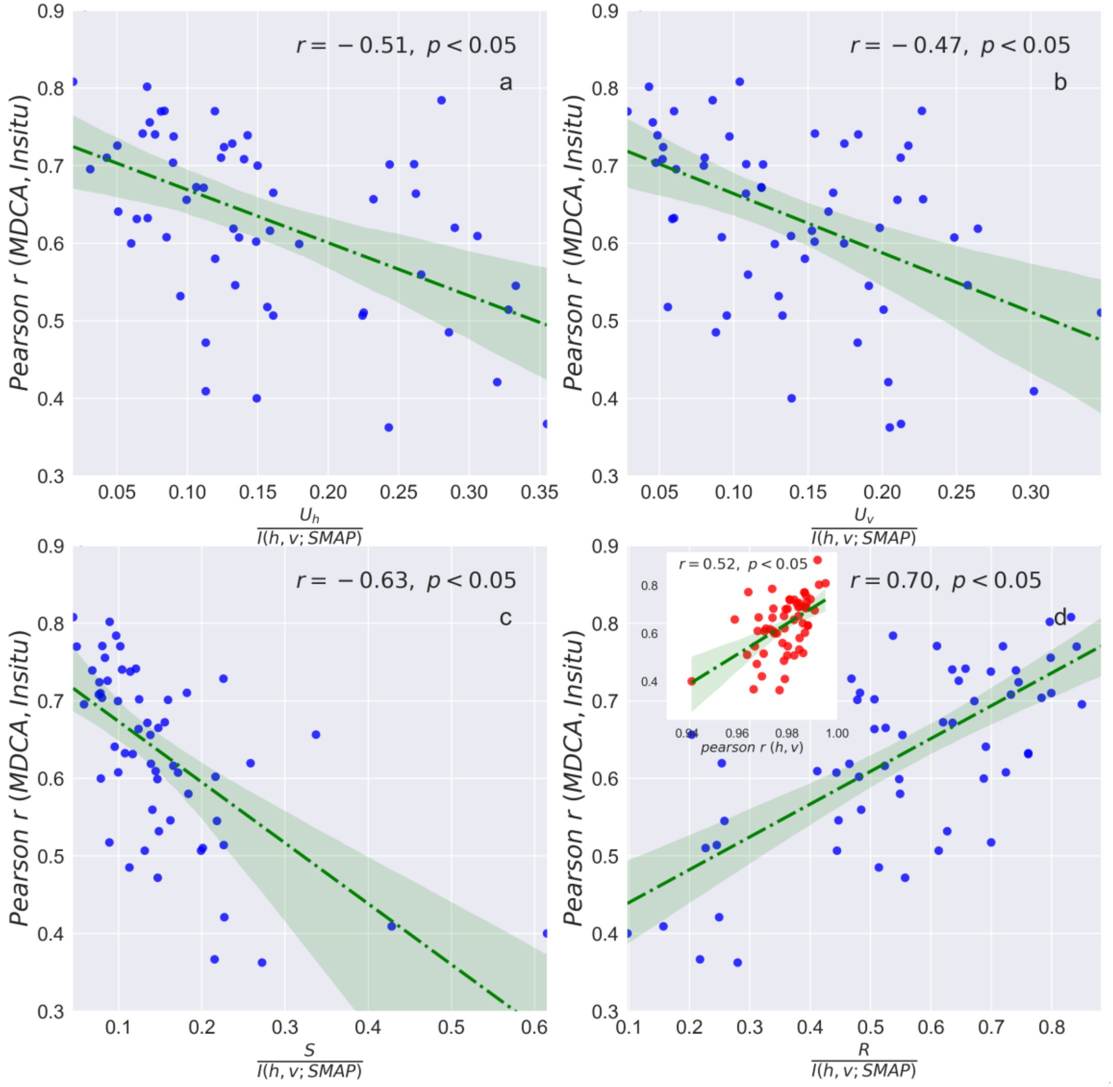
Table 1 The amount of informational uncertainties in percentage. The values in the table are the average of each landcover. The values in "Overall" is the average of all the sites.

Landcover	Unique information of T_{Bh} (U_h) (and its % I(T_{Bh}, T_{Bv}; DCA))	Unique information of T_{Bv} (U_v) (and its % I(T_{Bh}, T_{Bv}; DCA))	Synergistic information of T_{Bh} and T_{Bv} (S) (and its % I(T_{Bh}, T_{Bv}; DCA))	Redundant information of T_{Bh} and T_{Bv} (R) (and its % I(T_{Bh}, T_{Bv}; DCA))	Mutual information (I(T_{Bh}, T_{Bv}; DCA))
Shrublands	0.03 (27%)	0.017(15%)	0.03 (26%)	0.036 (32%)	0.113
Grasslands	0.029 (21%)	0.014 (10%)	0.02 (14%)	0.077 (55%)	0.14
Croplands	0.017 (12%)	0.013 (9%)	0.016 (12%)	0.095 (67%)	0.141
Mixed	0.014 (12%)	0.007 (6%)	0.01 (8%)	0.091(74%)	0.122
Overall	0.026 (19%)	0.013 (10%)	0.019 (14%)	0.08 (57%)	0.137

Table 2 The partial information decomposition components. The values in the table are the average of each landcover. The values in “Overall” is the average of all the sites.



~~**Figure 6 Normalized partial information decomposition components against RMSE of MDCA and *in situ* soil moisture.**~~



~~**Figure 7 Normalized partial information decomposition components against pearson *r* of MDCA and *in situ* soil moisure.**~~