Remote Sensing of Environment 205 (2018) 329–337

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Validation of the SMAP freeze/thaw product using categorical triple collocation

Haobo Lyu[a,1], Kaighin A. McColl[b,c,1], Xinlu Li[a], Chris Derksen[d], Aaron Berg[e], T. Andrew Black[f], Eugenie Euskirchen[g], Michael Loranty[h], Jouni Pulliainen[i], Kimmo Rautiainen[i], Tracy Rowlandson[e], Alexandre Roy[j,k], Alain Royer[j,k], Alexandre Langlois[j,k], Jilmarie Stephens[f], Hui Lu[a,l,*], Dara Entekhabi[m,n]

[a] Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100084, China
[b] Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA 02138, USA
[c] John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA
[d] Climate Research Division, Environment Canada, Toronto, ON M3H 5T4, Canada
[e] Department of Geography, University of Guelph, 50 Stone Road East, Guelph, Ontario N1G2W1, Canada
[f] Faculty of Land and Food Systems, University of British Columbia, 2357 Main Mall, Vancouver, British Columbia V6T 1Z4, Canada
[g] Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK 99775, USA
[h] Department of Geography, Colgate University, Hamilton, NY, USA
[i] Finnish Meteorological Institute, Arctic Research, P.O. Box 503, FI-00101 Helsinki, Finland
[j] Centre d'Applications et de Recherches en Télédétection, Université de Sherbrooke, Sherbrooke, Québec, Canada
[k] Centre for Northern Studies, Québec, Canada
[l] Joint Center for Global Change Studies, Beijing 100875, China
[m] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[n] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

ARTICLE INFO

Keywords:
SMAP
Triple collocation
Validation
Freeze/thaw

ABSTRACT

The landscape freeze/thaw (FT) state plays an important role in local, regional and global weather and climate, but is difficult to monitor. The Soil Moisture Active Passive (SMAP) satellite mission provides hemispheric estimates of landscape FT state at a spatial resolution of approximately 36^2 km^2. Previous validation studies of SMAP and other satellite FT products have compared satellite retrievals with point estimates obtained from in-situ measurements of air and/or soil temperature. Differences between the two are attributed to errors in the satellite retrieval. However, significant differences can occur between satellite and in-situ estimates solely due to differences in scale between the measurements; these differences can be viewed as 'representativeness errors' in the in-situ product, caused by using a point estimate to represent a large-scale spatial average. Most previous validation studies of landscape FT state have neglected representativeness errors entirely, resulting in conservative estimates of satellite retrieval skill. In this study, we use a variant of triple collocation called 'categorical triple collocation' – a technique that uses model, satellite and in-situ estimates to obtain relative performance rankings of all three products, without neglecting representativeness errors – to validate the SMAP landscape FT product. Performance rankings are obtained for nine sites at northern latitudes. We also investigate differences between using air or soil temperatures to estimate FT state, and between using morning (6 AM) or evening (6 PM) estimates. Overall, at most sites, the SMAP product or in-situ FT measurement is ranked first, and the model FT product is ranked last (although rankings vary across sites). These results suggest SMAP is adding value to model simulations, providing higher-accuracy estimates of landscape FT states compared to models and, in some cases, even in-situ estimates, when representativeness errors are properly accounted for in the validation analysis.

* Corresponding author.
E-mail address: luhui@tsinghua.edu.cn (H. Lu).
[1] Contributed equally.

https://doi.org/10.1016/j.rse.2017.12.007
Received 15 September 2017; Received in revised form 17 November 2017; Accepted 3 December 2017
Available online 11 December 2017

1. Introduction

The landscape freeze/thaw (FT) state at high latitudes is an important land surface state variable with impacts at local, regional and global scales. Locally, the period in which the landscape is thawed bounds the growing season and phenological cycles of vegetation and local ecosystems. Regionally, FT state constrains available soil moisture, which controls surface fluxes of heat and moisture into the atmosphere. In turn, surface fluxes partially control the formation of clouds and modulate regional weather systems (Betts et al., 2001). Since the soil storage has considerable 'memory' of comparatively rapid atmospheric anomalies (Katul et al., 2007; Koster and Suarez, 2001; McColl et al., 2017b, 2017a; Seneviratne et al., 2006), the corresponding anomalies in soil moisture and its FT state can be imprinted on surface fluxes over a much longer period, with implications for land-atmosphere coupling and weather prediction. Finally, at global scales, the long-term thawing of permafrost and associated release of carbon dioxide and methane may have significant implications for future global climate (Grosse et al., 2016; Schuur et al., 2015). It is therefore critical to monitor FT state at high latitudes.

The most viable option for monitoring FT state at continental scales is by using microwave satellite observations. Microwave measurements can be made regardless of solar illumination or the presence of clouds. Both passive and active microwave observations have been used in previous studies to estimate FT state globally. In particular, FT state has been estimated using passive observations from sensors including the Aquarius/SAC-D satellite (Brucker et al., 2014; Le Vine et al., 2007; Roy et al., 2015), the Soil Moisture and Ocean Salinity (SMOS) satellite (Kerr et al., 2010; Rautiainen et al., 2016), Scanning Multichannel Microwave Radiometer (SMMR) and Special Sensor Microwave Imager (SSM/I) (Kim et al., 2011, 2012) and others. Most recently, NASA's Soil Moisture Active Passive (SMAP) mission launched in January 2015 and provides retrievals of landscape FT state (Derksen et al., 2017; Dunbar et al., 2016; Entekhabi et al., 2010).

Satellite observations of landscape FT state must be validated against ground observations prior to use. Error estimates are also required if FT estimates are to be assimilated into land-surface models (Farhadi et al., 2014). Previous validation studies typically directly compare satellite FT retrievals with FT estimates obtained from point-scale in-situ station measurements of air or soil temperature (e.g., Colliander et al., 2012; Derksen et al., 2017; Kim et al., 2011; Podest et al., 2014; Zhang et al., 2003). Differences between the satellite FT retrieval and the in-situ point estimate are attributed to errors in the satellite product. However, differences also occur due to differences in the spatial scale of the satellite and in-situ estimates (Roy et al., 2017). A satellite typically measures a spatial average over an area on the order of 10^2–10^4 km^2, several orders of magnitude larger than the area measured by an in-situ station. The spatial average of landscape FT state over 10^2–10^4 km^2 of a typically heterogeneous land surface may differ substantially from the FT state at a single point within the domain. Therefore, even if a satellite has no measurement error, there will likely still be substantial differences between satellite and in-situ estimates of FT state. These differences can be thought of as errors in the in-situ estimate – often referred to as "representativeness" errors – caused by the in-situ estimate's undesirably small measurement scale. Averaging over multiple in-situ estimates in the domain will reduce representativeness errors, but the difficulties of installing and maintaining in-situ stations at high latitudes typically means relatively few in-situ stations are available for comparison. In addition, the in-situ measurements themselves may be flawed, either due to instrument performance deficiencies near the zero degree threshold, or due to the installation depths of instruments in standard soil monitoring networks (Williamson et al., in press). Overall, this implies that attributing differences between satellite and in-situ estimates of FT state (or any other variable) solely to errors in the satellite product will substantially overestimate the error in the satellite product, since in-situ station errors will be incorrectly classified as satellite measurement errors.

A general solution to this problem – triple collocation (TC) – was provided by Stoffelen (1998). TC is a technique for validating satellite observations and model estimates using in-situ measurements, without ignoring representativeness errors. It has been widely adopted in validating satellite retrievals of variables including soil moisture (e.g., Draper et al., 2013; Gruber et al., 2016), precipitation (e.g., Alemohammad et al., 2015), ocean surface wind speed (e.g., McColl et al., 2014; Vogelzang et al., 2011), and many others. However, complications arise when applying TC to validating satellite estimates of landscape FT state. TC requires that errors in the satellite, model and in-situ estimates must be uncorrelated with each other, and with the (unknown) true landscape FT state. For a binary variable such as landscape FT state, these assumptions are both strongly violated (McColl et al., 2016), resulting in substantial biases in the TC error estimates. To circumvent this problem, McColl et al. (2016) introduced a variant of TC that can be applied to binary and categorical variables – called categorical triple collocation (CTC) – without resulting in biases in the resulting error estimates. They provided a proof-of-concept demonstration of CTC, using it to validate landscape FT estimates, derived from the NASA/SAC-D Aquarius satellite, Canadian Meteorological Center (CMC) surface analysis, and surface measurements from a network of western Canadian field sites. The results showed that, at most sites, the in-situ observations exhibited the lowest accuracy compared to the model and satellite estimates, demonstrating the influence of representativeness errors on the in-situ observations.

In this study, we apply CTC to the validation of the SMAP FT product over a range of sites at high latitudes. In Section 2, we briefly review CTC, describe the different FT products used in this study, and the different field sites. In Section 3, we present the results of the CTC validation analysis at each of the field sites, compare the results with previous studies, and consider limitations of our analysis. We present conclusions from our analysis in Section 4.

2. Methods

In this section, we review CTC and describe the datasets and field sites used to validate the SMAP FT product.

2.1. Categorical triple collocation (CTC)

We begin by briefly reviewing CTC, as presented in McColl et al. (2016). CTC assumes an error model of the form

$$X_i = T + \varepsilon_i$$

where X_i is an observation from the ith measurement system (model, satellite and in-situ), T is the unknown true value and ε_i is a random error. The observations X_i and truth T are binary variables and can take on values in the set $\{-1, 1\}$ only, i.e.,

$$T = \begin{cases} 1, & \text{if frozen} \\ -1, & \text{otherwise} \end{cases}$$

and similarly for X_i. The errors ε_i are therefore dependent on the value of T (to ensure that X_i remains in the set $\{-1, 1\}$), and can take on values in the set $\{-2, 0, 2\}$ only. In classical TC, it is assumed that the errors ε_i are uncorrelated with each other, and with T. Based on these assumptions, expressions for error metrics can be obtained from the sample covariance matrix $\mathrm{Cov}(X_i, X_j)$, estimated from observations, without assuming any of the three measurement systems are error-free.

However, the errors are dependent on T in the binary case, meaning the errors are correlated with T. Furthermore, since errors from all three measurement systems are correlated with T, they are correlated with each other. Therefore, both key assumptions of classical TC are automatically violated when it is applied to binary variables. We require an alternative to classical TC that does not make this assumption. Rather than assuming errors between measurement systems are uncorrelated

with each other and with T, CTC makes a weaker assumption: that the errors are *conditionally independent*, i.e., $\Pr(\varepsilon_i, \varepsilon_j | T) = \Pr(\varepsilon_i | T) \Pr(\varepsilon_j | T)$, where $\Pr(X|Y)$ refers to the probability of X occurring, conditioned on Y. This weaker assumption is not automatically violated when applied to binary variables. Based on this weaker assumption, McColl et al. (2016) demonstrated that the sample covariance matrix $\mathrm{Cov}(X_i, X_j)$ can be decomposed to obtain performance rankings of the three measurement systems with respect to their balanced accuracies, a performance metric closely related to the simple accuracy metric. More specifically, the balanced accuracy is defined as

$$\pi = \frac{1}{2}(\psi + \eta)$$

where ψ is the measurement sensitivity (varying between zero and one, and is higher when there are fewer false negatives) and η is the measurement specificity (also varying between zero and one, and is higher when there are fewer false positives). In contrast, the simple accuracy (which also varies between zero and one) is the proportion of classifications that are correct. To see the advantage of the balanced accuracy over the simple accuracy, consider the case of a hypothetical satellite retrieval of FT state that always classifies all regions as 'thawed'. This hypothetical retrieval has no physical basis and should receive poor performance estimates. However, in some parts of the world, the landscape is thawed for nearly the entire year; in these regions, the hypothetical FT retrieval will receive a relatively high performance estimate when the simple accuracy metric is used. In this case, the relatively high performance estimate is unjustified, and an unwanted artifact of the simple accuracy metric. In contrast, the balanced accuracy will more heavily penalize the hypothetical FT retrieval when it incorrectly classifies the landscape during the limited time it is frozen. This will result in a lower performance estimate compared to the simple accuracy metric, and better reflects the actual performance of the retrieval.

Full details of the CTC algorithm are given in McColl et al. (2016), but it can be summarized as follows:

1. Estimate the sample 3×3 covariance matrix $Q_{ij} \equiv \mathrm{Cov}(X_i, X_j)$ from the observations X_1, X_2, X_3.
2. Estimate $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} \sqrt{Q_{12}Q_{13}/Q_{23}} \\ \sqrt{Q_{12}Q_{23}/Q_{13}} \\ \sqrt{Q_{23}Q_{13}/Q_{12}} \end{bmatrix}$ from the sample covariance matrix.
3. Sort $\boldsymbol{w}$ in descending order to obtain rankings. For example, if $w_2 > w_1 > w_3$, then measurement system 2 has the best performance with respect to π, followed by measurement system 1, with measurement system 3 exhibiting the poorest performance.

CTC provides relative performance rankings of the three measurement systems with respect to π, but is not able to provide absolute estimates of π for each system. This is the price paid for relaxing two key assumptions of classical TC so that TC can be applied to binary and categorical variables. In addition, like any validation analysis, the CTC performance ranking is subject to sampling error based on sample size. In this study, we obtain sampling error estimates by performing bootstrapping (see Section 2.2.3).

2.2. Data

In this section, the satellite, in-situ, and model FT products used in the analysis are described.

2.2.1. SMAP data

Launched in January 2015, SMAP provided collocated radar (active) and radiometer (passive) L-band observations, with an approximate repeat time of three days, consisting of ascending (6 PM) and descending (6 AM) half-orbits. The radar ceased operation after 11 weeks due to an instrument anomaly, but the radiometer continues to function nominally; we therefore use SMAP radiometer observations in this study. The SMAP radiometer has a resolution of approximately 36^2 km^2 corresponding to the domain over which the radiometer antenna records half of the power it receives. The data are gridded onto a 9^2 km^2 grid with polar projection using Backus-Gilbert optimal interpolation (Chaubell et al., 2016b, 2016a), and converted to a landscape FT state estimate, in a process described below. In this validation study, we use observations spanning the first full annual cycle of SMAP observations (1 April 2015 to 31 March 2016).

The SMAP FT product is acquired from the enhanced Level-3 (L3) product (Dunbar et al., 2017). It provides a daily composite of FT state for land areas above 45°N at 6 AM and 6 PM. The SMAP FT product is derived from the normalized polarization ratio (NPR), defined as

$$NPR = \frac{T_{BV} - T_{BH}}{T_{BV} + T_{BH}}$$

where T_{BV} and T_{BH} are the vertically- and horizontally-polarized brightness temperatures. This ratio is further normalized to obtain

$$FF_{NPR} = \frac{NPR - NPR_{th}}{NPR_{fr} - NPR_{th}}$$

where NPR_{fr} is a reference value of NPR corresponding to a frozen state, and NPR_{th} is a reference value corresponding to a thawed state. The frozen reference value is estimated as the average of the ten lowest NPR values over January and February 2016; the thawed reference is estimated similarly, using the ten highest NPR values over July and August 2015 (Derksen et al., 2017). Finally, the landscape FT estimate is obtained by thresholding FF_{NPR}:

$$X_1^{AM/PM} = \begin{cases} 1, & \text{if } FF_{NPR} \le \Delta(t) \\ -1, & \text{if } FF_{NPR} > \Delta(t) \end{cases}$$

where $\Delta(t)$ is fixed at 0.5 (optimization is possible, but not applied in this version of the product), $X_1 = 1$ means 'frozen' and $X_1 = -1$ means 'thawed', and $X_1^{AM/PM}$ is estimated using AM/PM observations, respectively (Derksen et al., 2017; Rautiainen et al., 2016). The SMAP FT product is estimated separately using only AM brightness temperature observations, and again using only PM observations.

Like all satellite-derived FT estimates, the SMAP retrievals contain contributions from sources beyond the soil FT state, which partially confound the retrieval. These sources include vegetation and snow cover. Rather than attempt to correct for such sources – a difficult task in an environment with incomplete observations – the FT retrievals are typically interpreted as indicative of 'landscape' FT state, which incorporates contributions from snow cover and vegetation into its definition, rather than strict 'soil' FT state. Although there is influence by a wet snow cover (Derksen et al., 2017; Roy et al., 2015), the SMAP L-band observations penetrate deeper into the soil and are expected to be less impacted by snow cover and vegetation compared to retrievals based on higher-frequency microwave observations (such as the satellite Ka-band radiometer measurements used in Kim et al., 2012). Unless otherwise specified, we define FT state as the landscape FT state in this manuscript.

2.2.2. In-situ data

In-situ station observations of surface air temperature (T_a) and near surface soil temperature (T_s; measured at 5-cm depth) are obtained from nine stations spanning Canada, Finland, Eastern Siberia, and Alaska. The sites span a range of landcover types (further details are provided in Table 1). The station air- and soil-temperature observations are translated to landscape FT estimates as follows. Within a given SMAP pixel, multiple in-situ measurements are averaged. The number of in-situ observations averaged at each site is given in Table 1. For station air temperatures, the in-situ FT product is defined according to

Table 1
Information for the study datasets and sites (Derksen et al., 2017). At each site, in-situ observations were averaged across multiple measurement stations.

Site ID	Latitude	Longitude	Region	Vegetation type	Number of stations in average
Kenaston	51.41°N	106.5°W	Saskatchewan, Canada	Croplands	35
BERMS old aspen (OA)	53.63°N	106.2°W	Saskatchewan, Canada	Deciduous Forest	3
BERMS old black spruce (OBS)	53.99°N	105.12°W	Saskatchewan, Canada	Coniferous Forest	2
Sodankyla	67.36°N	26.64°E	Finland	Coniferous Forest	15
Saariselka	68.38°N	27.42°E	Finland	Grasslands	4
Chersky	68.65°N	161.65°E	Eastern Siberia, Russia	Deciduous Needleleaf Forest	6
Imnavait	68.62°N	149.30°W	Alaska, USA	Barren/Sparse	1
Baie-James	53.41°N	75.013°W	Quebec, Canada	Coniferous Forest	2
Cambridge Bay	69.15°N	105.11°W	Northwest Territories, Canada	Barren/Sparse	1

$$X_2^{AM/PM}(T_a) = \begin{cases} 1, & \text{if } T_a \le 0^{\circ}\text{C} \\ -1, & \text{if } T_a > 0^{\circ}\text{C} \end{cases},$$

where T_a refers to the averaged air temperature across in-situ measurement stations within the SMAP pixel. For station soil temperatures, the in-situ FT product is defined according to

$$X_2^{AM/PM}(T_s) = \begin{cases} 1, & \text{if } T_s \le 0^{\circ}\text{C} \\ -1, & \text{if } T_s > 0^{\circ}\text{C} \end{cases},$$

where T_s refers to the averaged soil temperature across in-situ measurement stations within the SMAP pixel. Each in-situ measurement has strengths and weaknesses with respect to sensitivity to actual landscape FT state. While both soil temperatures and air temperatures are correlated with landscape FT state, $X_2^{AM/PM}(T_s)$ has a more direct link to soil FT state, so measurement errors in $X_2^{AM/PM}(T_s)$ are expected to be lower compared to $X_2^{AM/PM}(T_a)$. However, the measurement footprint of $X_2^{AM/PM}(T_s)$ is smaller than that of $X_2^{AM/PM}(T_a)$ – the spatial correlation length of soil temperatures is substantially smaller than that of air temperatures – so $X_2^{AM/PM}(T_a)$ is expected to have smaller representativeness errors when measuring landscape FT state over larger scales relevant to the validation of satellite observations. As discussed further in Section 3.1, uncertainties also arise related to how T_a and T_s capture freeze and thaw transition events. For example, soil temperatures in spring typically remain frozen for a number of days after the onset of snow melt. There is evidence, however, that SMAP radiometer measurements respond to the immediate onset of snow melt, triggering 'thaw' retrievals in the case of wet snow over frozen soil (Derksen et al., 2017). In this scenario, there will be better agreement between SMAP FT retrievals and $X_2^{AM/PM}(T_a)$, even though the near surface soil layer remains frozen.

2.2.3. Model data

The model-derived FT estimate is obtained from 0 to 10 cm surface temperatures (estimated as the average of skin temperature and 10 cm depth soil temperature) from the NASA Global Modeling and Assimilation Office (GMAO) GEOS-5 Nature Run product (Reichle et al., 2016) according to

$$X_3^{AM/PM} = \begin{cases} 1, & \text{if } T_s \le 0^{\circ}\text{C} \\ -1, & \text{if } T_s > 0^{\circ}\text{C} \end{cases},$$

The model spatial resolution is 9^2 km^2.

The three datasets were temporally- and spatially-collocated at each site, using nearest-neighbor sampling for the spatial collocation, and the in-situ air and soil temperatures as the temporal reference with a maximum temporal collocation window of 1 day. CTC was applied to the sample triplets at each site to obtain performance rankings. We performed bootstrapping (Efron and Tibshirani, 1994) using 1000 replicates, to quantify the impact of sampling error on the estimated rankings, as in McColl et al. (2016). Bootstrapping is a commonly-used, non-parametric technique for estimating sampling error. For a given site, with N available sample triplets (call this set S_0), we randomly draw N triplets from those available *with replacement* (call this set S_1). S_0 and S_1 will almost certainly be different: some of the triplets in S_1 will likely be repeated, and therefore S_1 will likely also exclude some triplets from S_0. The process is repeated many times (in our case, 1000 times), generating bootstrapped sets of sample triplets S_1, …, S_{1000}. We perform CTC on each set S_1, …, S_{1000} and obtain 1000 different performance rankings. Differences between estimated performance rankings across sets S_1, …, S_{1000} are attributed to sampling error. If the estimated CTC performance rankings across sets S_1, …, S_{1000} are all identical, then we estimate zero sampling error. On the other hand, if the estimated performance rankings across sets differ substantially, then we estimate large sampling error.

3. Results and discussion

Before presenting the CTC validation results, we aim to better understand the differences between the FT products used in this study. We investigate differences between in-situ products derived from air temperatures ($X_2(T_a)$), and those derived from soil temperatures ($X_2(T_s)$); and investigate differences between estimated landscape FT state using morning (6 AM) or evening (6 PM) measurements. Finally, we use CTC to estimate performance rankings of the satellite (X_1), in-situ (X_2) and model (X_3) FT state estimates across the nine high-latitude study sites.

3.1. Differences between in-situ air and soil temperature FT products

We first examine differences between the in-situ air and soil temperature FT products ($X_2(T_a)$ and $X_2(T_s)$, respectively). Two representative examples are shown in Fig. 1. The site at Sodankyla exhibits significant differences between the seasonal cycles of T_s and T_a. Soil temperatures rarely drop below freezing, resulting in relatively few and relatively short frozen periods (Fig. 1a). In contrast, air temperatures follow a clear seasonal cycle and spend a considerable period of the year below zero, resulting in a much longer estimated frozen period (Fig. 1b). Qualitatively similar patterns are observed (not shown) at Baie James, Chersky, and the two BERMS stations (OA and OBS). In contrast, at the Cambridge Bay site, relatively few differences are observed between soil and air temperatures, and their respective FT products (Fig. 1c and d). Similar behavior is observed at the Imnavait site (not shown).

There are many possible reasons for differences in the seasonal cycles of T_a and T_s. One explanation could be the effects of forest and vegetation cover on snow cover. In the winter, snow insulates the soil, increasing T_s relative to T_a (Bartlett et al., 2004). There is considerable variability in snow cover across climates, biomes and topography. However, in forested areas, reductions in snow accumulation due to canopy interception are often exceeded by reductions in snow ablation due to canopy shading (Varhola et al., 2010). This results in a net increase in snow cover in some forests, compared to equivalent non-forested regions. Therefore, we might expect that at forested or vegetated sites $T_s > T_a$ during the winter, but not at barren or sparsely vegetated sites. This is what we see in the in-situ observations: the sites at which T_a and T_s follow similar seasonal cycles (Cambridge Bay and

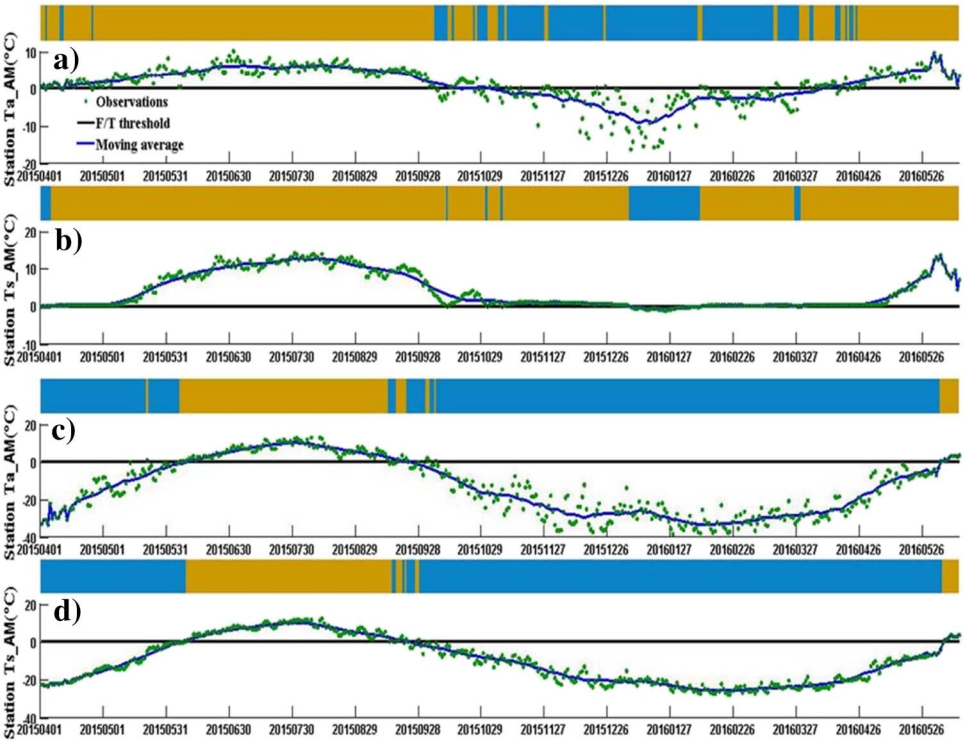

Fig. 1. (a) Time series plots at Sodankyla station of 6 AM in-situ station air temperatures (corresponding to SMAP's descending overpass). Boxes above the time series plot show the inferred FT product $X_2(T_a)$, estimated by thresholding the air temperature time series with respect to the given FT threshold. Light brown, blue and white boxes signify thawed, frozen and missing data, respectively. (b) Same as (a), except showing 6 AM in-situ station soil temperatures (corresponding to SMAP's descending overpass), and corresponding FT product $X_2(T_s)$. (c) Same as (a), except at the Cambridge Bay site. (d) Same as (b), except at the Cambridge Bay site. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

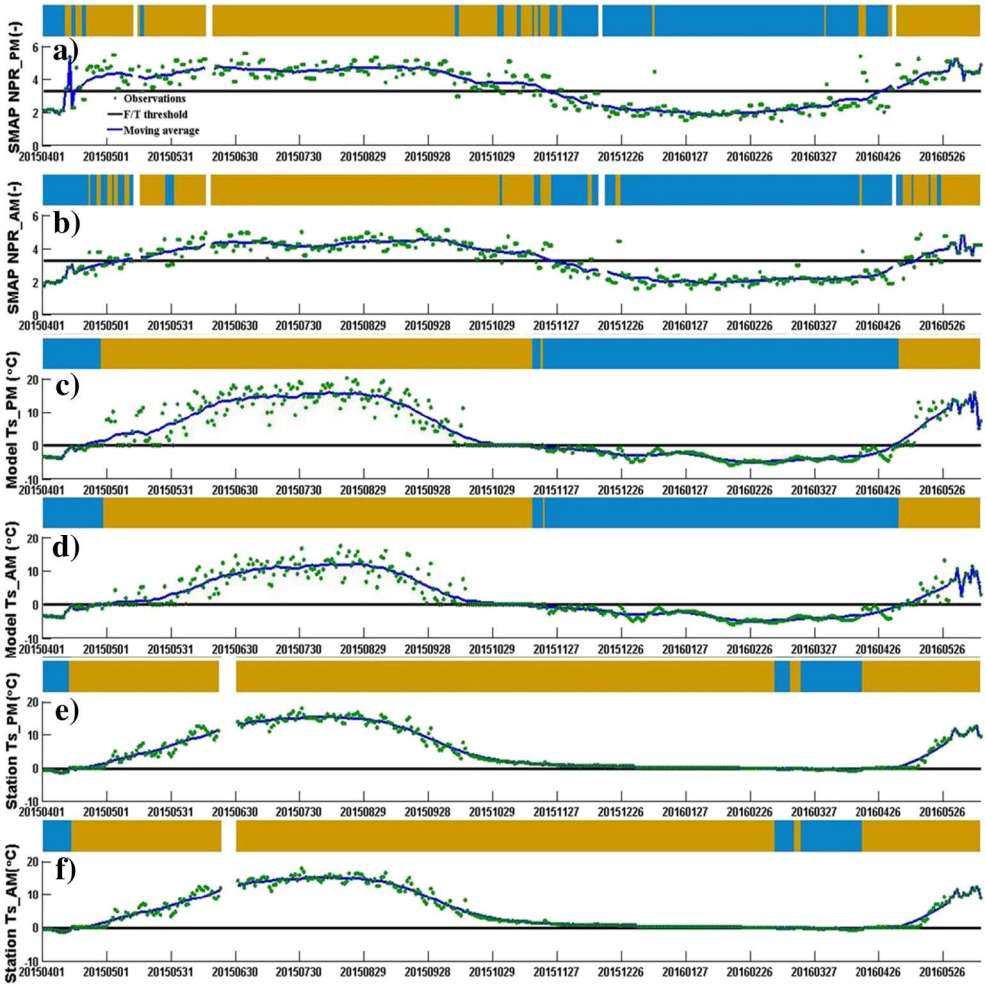

Fig. 2. (a) Time series plots at the Baie-James station of 6 PM (corresponding to SMAP's ascending overpass) *NPR*. Boxes above the time series plot show the inferred FT product X_1^{PM}, estimated by thresholding the *NPR* time series with respect to the given FT threshold (chosen to be equivalent to thresholding FF_{NPR} at a value of 0.5). Light brown, blue and white boxes signify thawed, frozen and missing data, respectively. (b) Same as (a), except using 6 AM *NPR* (corresponding to SMAP's descending overpass) and the inferred FT product X_1^{AM}. (c) Same as (a), except using 6 PM model soil temperatures and the inferred FT product X_3^{PM}. (d) Same as (a), except using 6 AM model soil temperatures and the inferred FT product X_3^{AM}. (e) Same as (a), except using 6 PM in-situ station soil temperatures and the inferred FT product $X_2^{PM}(T_s)$. (f) Same as (a), except using 6 AM in-situ station soil temperatures and the inferred FT product $X_2^{AM}(T_s)$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

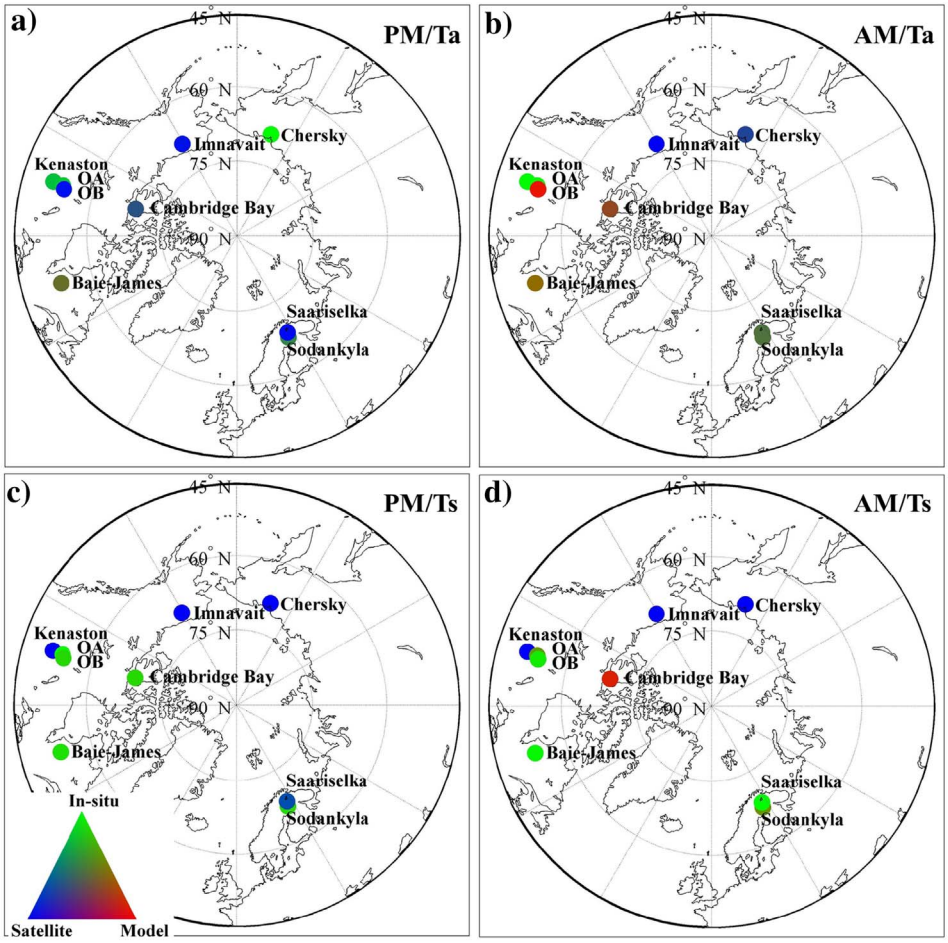

Fig. 3. Maps of the measurement system ranked first at each site using CTC with four different sets of measurements: (a) X_1^{PM} (i.e., based on SMAP PM observations), $X_2^{PM}(T_a)$ (i.e., based on in-situ PM air temperature observations) and X_3^{PM} (i.e., based on PM model output); (b) X_1^{AM}, $X_2^{AM}(T_a)$ and X_3^{AM}; (c) X_1^{PM}, $X_2^{PM}(T_s)$ and X_3^{PM}; (d) X_1^{AM}, $X_2^{AM}(T_s)$ and X_3^{AM}. Since the top-ranked measurement system may vary across the 1000 bootstrap replicates at any given site, we calculate the proportion of bootstrap replicates ranking the satellite, in-situ and model FT estimates first, and map these to a red–green–blue color space, respectively. For example, if the satellite is ranked the highest in all 1000 bootstrapped performance rankings at a site, it is colored blue; if the satellite is ranked first in 50% of the bootstrapped rankings, and the model is ranked first in the other 50%, it is colored purple. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Imnavait) are both sparsely vegetated, whereas $T_s > T_a$ during the winter at the other (forested or relatively well-vegetated) sites. If this explanation is correct, $X_2(T_s)$ clearly provides a more accurate estimate of soil FT state compared to $X_2(T_a)$.

However, alternative explanations exist that yield the opposite conclusion. While $X_2(T_s)$ is the ultimate measure of soil FT state at a single point, over the scale of a satellite footprint, representativeness errors in $X_2(T_s)$ may be substantial. This is because the FT state at a point is not necessarily an accurate estimate of the FT state averaged over a larger area (the quantity observed by satellites), particularly for heterogeneous variables like landscape FT state. While T_a has a weaker physical relation to soil FT state compared to T_s, it has a longer correlation length, and therefore integrates information over a larger footprint that more closely matches satellite observations. It may, therefore, be less affected by representativeness errors compared to T_s, implying that $X_2(T_a)$ may provide a more accurate estimate of landscape FT state compared to $X_2(T_s)$.

These are only two plausible hypotheses for observed differences in the seasonal cycles of T_a and T_s. There are insufficient data to definitively choose one over the other; we simply highlight these differences and stress their importance in interpreting the validation results to follow (and in other satellite FT validation studies).

3.2. Differences between AM and PM FT products

We next consider differences between FT products estimated with morning (6 AM) and evening (6 PM) observations. For the in-situ products, the greatest differences occur when using air temperatures compared to soil temperatures, with air temperatures considerably lower at 6 AM compared to 6 PM. This results in a significantly extended frozen period when estimating an in-situ FT product using 6 AM air temperatures rather than 6 PM temperatures (not shown). The difference is much less pronounced when using soil temperatures instead (Fig. 2e and f show representative examples from the Baie-James site), probably due to the soil's substantial thermal inertia and snowpack thermal insulation.

For satellite FT products, the differences are less pronounced, although the frozen period tends to be longer when using AM (descending) observations compared to PM (ascending). The most pronounced differences occur in the spring transition (Fig. 2a and b show representative examples from the Baie-James site). Derksen et al. (2017) found better agreement between SMAP retrievals and ground observations for PM observations, compared to AM. This is attributed to ephemeral refreezing events that can occur overnight, which are not observed by SMAP or soil temperature products, and are only seen in air temperature products, which can respond to short-term forcing.

Finally, since the model FT product is based on soil temperature, the differences between using AM and PM temperatures are relatively small (Fig. 2c and d).

3.3. CTC performance rankings

Given the known differences in FT products estimated using AM and PM air temperatures and soil temperatures, we now present the CTC performance rankings of the different products at each site (Figs. 3–5). There is considerable variation between sites and products. The results also vary depending on the choice of products used in the analyses, and the timing of the observations (AM or PM). Imnavait is the only site where the same FT product (in this case, the satellite product) is confidently ranked first for all combinations of analyses (Fig. 3). The

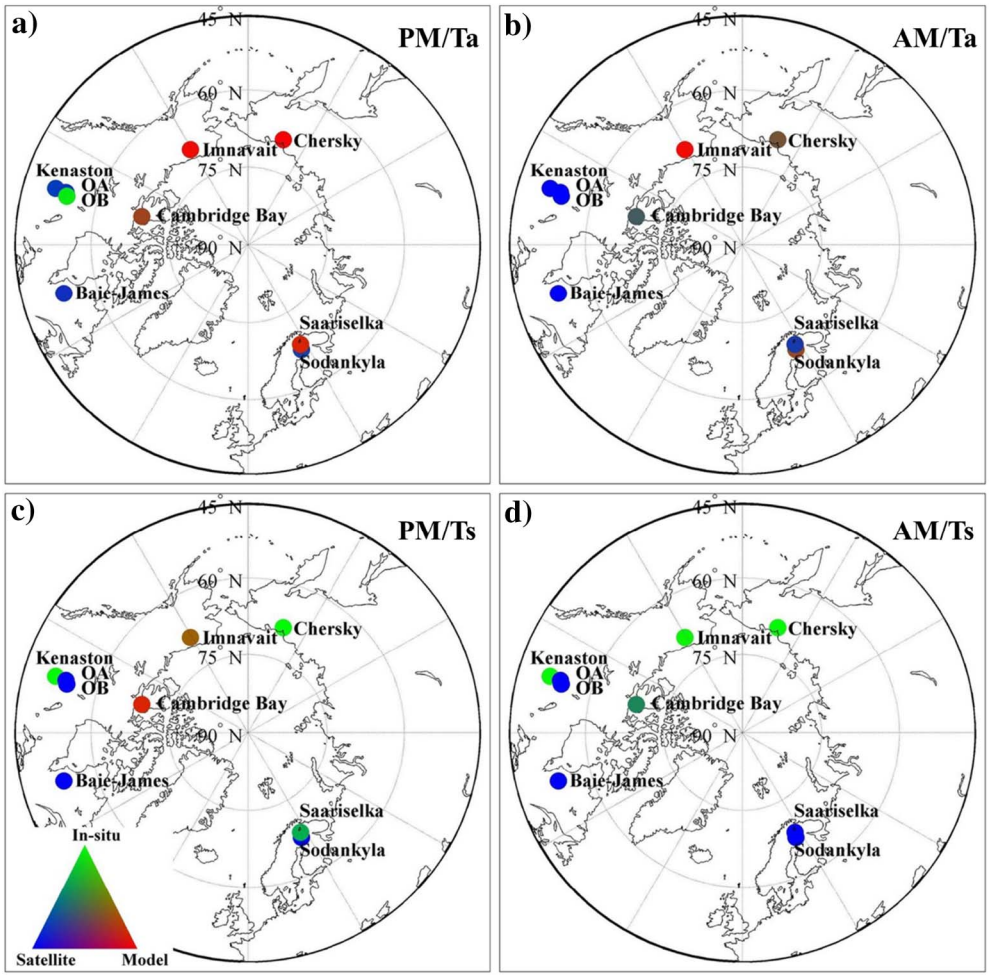

Fig. 4. Similar to Fig. 3 for the measurement system ranked second at each site using CTC.

variability is expected given the differences in FT products outlined in Sections 3.1 and 3.2. Furthermore, some of the observed variability in performance rankings is likely attributable to differences in the number of in-situ measurement stations present at each site (Table 1). All else being equal, a site with more measurement stations would be expected to have lower representativeness errors, resulting in a potentially higher performance ranking for the in-situ product.

Despite this variability, some broad conclusions can be drawn. First, either the satellite or in-situ product is ranked first in most cases (Fig. 3). The only exceptions to this are the BERMS OBS site when air temperatures and AM observations are used (Fig. 3b), and the Cambridge Bay site when soil temperatures and AM observations are used (Fig. 3d). Second, the model FT product is most likely to be ranked last, particularly for the case where soil temperatures and AM observations are used to derive FT products (Fig. 5d). However, this result is more dependent on choices made in defining the FT product (for example, in Fig. 5a, the model is not clearly ranked last at the majority of sites) and therefore more uncertain.

These results are in contrast to a previous study, which found in-situ observations frequently displayed the poorest performance (McColl et al., 2016). However, in that study, in-situ observations were being used to validate a satellite product based on Aquarius observations at a resolution of $\sim 100^2$ km^2, compared with a resolution of 36^2 km^2 in this study. Furthermore, a single in-situ measurement station was used to validate each Aquarius observation in McColl et al. (2016); in contrast, in this study, at most sites, multiple in-situ measurement stations are averaged before comparing with SMAP observations. Therefore, the representativeness errors present in the in-situ observations in McColl et al. (2016) are expected to be significantly larger. The higher spatial resolution of the SMAP observations, and the use of averaged multiple in-situ measurements, allows for a more reasonable comparison between the satellite and in-situ measurements.

Our analysis is subject to several limitations. First, the number of field sites in this study is relatively small, due to the difficulty of maintaining field sites at high latitudes. This problem is compounded by the fact that FT state is heterogeneous both horizontally (e.g., due to variations in land cover and topography) and vertically (e.g., due to variations in soil, snow and vegetation). This is a continuing challenge for all satellite FT validation studies. A sustained effort to establish and maintain sites under difficult operating conditions is required. Second, we have used one year of observations, a relatively small sample size, limited by the availability of SMAP observations. We have performed bootstrapping to estimate effects of sample size on the analysis; in most cases, bootstrapping shows that we are able to cleanly identify performance rankings in spite of the low sample size, with some exceptions. This problem will be mitigated as the satellite record grows with time. Third, for CTC to provide unbiased performance rankings, we require that errors in the satellite, model, and in-situ FT products are conditionally independent of one another; this assumption may not always hold. However, while assumptions of zero error cross-correlation are not always valid in other domains (for instance, using triple collocation to estimate errors in soil moisture products (Yilmaz and Crow, 2014)), the assumption of conditional independence required by CTC is considerably weaker than the assumptions of classical triple collocation. We therefore expect this assumption to be violated to a lesser degree compared with the assumptions of classical triple collocation.

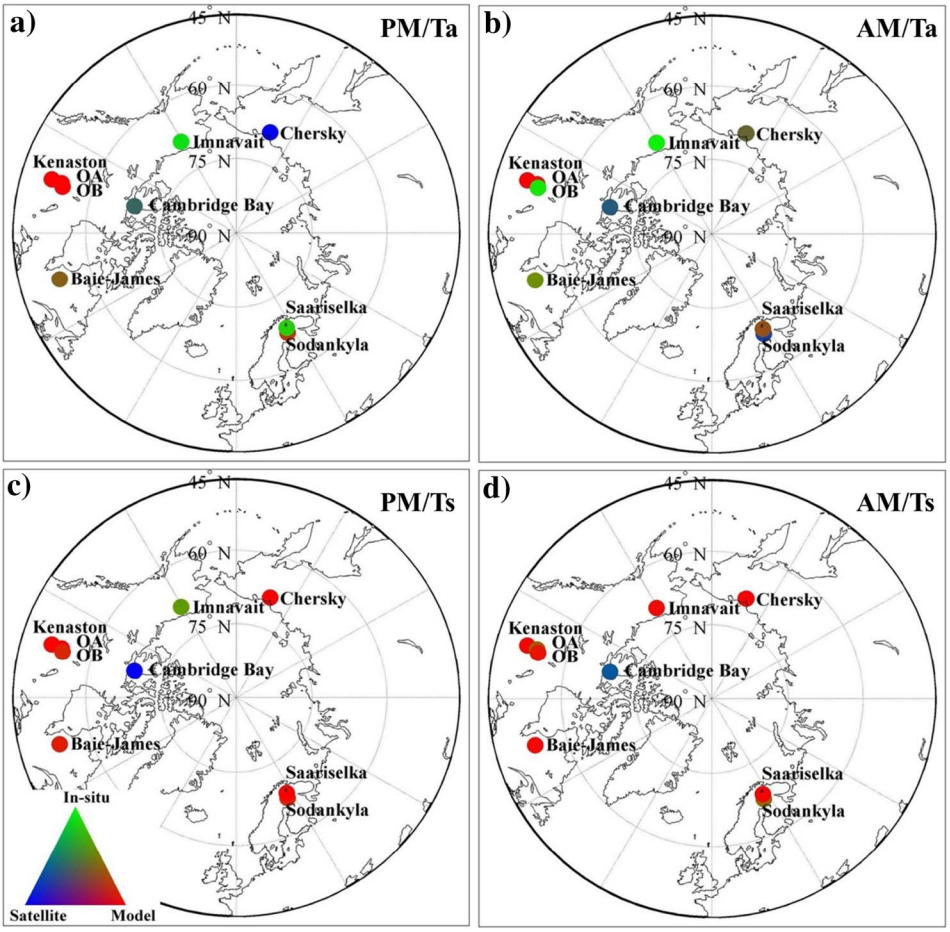

Fig. 5. Similar to Fig. 3 for the measurement system ranked third at each site using CTC.

4. Conclusions

Satellite observations of FT state are essential for monitoring and predicting the response of high-latitude environments to a changing climate. In this study, we validated the performance of the SMAP FT product through comparison with in-situ air and soil temperatures, and with model soil temperatures. In a departure from most other FT validation studies, we did not assume that representativeness errors – implicit errors in in-situ products when using the in-situ point estimate to represent a large-scale average – were zero, a common but often substantially incorrect assumption. To avoid this assumption, we used CTC (a variant of triple collocation specifically designed for use with binary and categorical variables) to estimate performance rankings for the model, satellite, and in-situ FT products. While there was considerable variation across sites, we observed that, at most sites, either the SMAP or in-situ FT product had the highest performance ranking. In contrast, at most sites, the model product had the poorest performance ranking. These results suggest that SMAP FT estimates are adding value compared to model FT estimates. They also suggest that comparing in-situ FT estimates with SMAP FT retrievals (36^2 km^2) is more justified than comparisons with coarser-resolution satellite FT products (such as $\sim 100^2$ km^2 Aquarius retrievals) used in previous studies, since higher resolution FT products resolve more of the FT spatial variability.

Acknowledgements

K.A.M. is funded by a Ziff Environmental Fellowship from the Harvard University Center for the Environment. H. Lyu, H. Lu and X.L. are funded by the National Basic Research Program of China (2015CB953703) and the National Natural Science Foundation of China (91537210 and 41371328). The authors acknowledge funding from the MIT Greater China Fund for Innovation project “Global monitoring of the water cycle using new microwave space-borne instruments” and from the SMAP mission. The computation for this work was supported by the Tsinghua National Laboratory for Information Science and Technology. Work at the Chersky sites by M.L. was supported by National Science Foundation award PLR-1304464. The Canadian Space Agency supported work at Kenaston by A.B. and T.R., and work at the Cambridge Bay and Baie-James sites by A. Royer, A. Roy and A.L.

References

Alemohammad, S.H., McColl, K.A., Konings, A.G., Entekhabi, D., Stoffelen, A., 2015. Characterization of precipitation product errors across the United States using multiplicative triple collocation. Hydrol. Earth Syst. Sci. 19, 3489–3503. http://dx.doi.org/10.5194/hess-19-3489-2015.

Bartlett, M., Chapman, D.S., Harris, R.N., 2004. Snow and the ground temperature record of climate change. J. Geophys. Res. 109. http://dx.doi.org/10.1029/2004JF000224.

Betts, A.K., Viterbo, P., Beljaars, A.C.M., van den Hurk, B.J.J.M., 2001. Impact of BOREAS on the ECMWF forecast model. J. Geophys. Res. Atmos. 106, 33593–33604. http://dx.doi.org/10.1029/2001JD900056.

Brucker, L., Dinnat, E.P., Koenig, L.S., 2014. Weekly gridded Aquarius L-band radiometer/scatterometer observations and salinity retrievals over the polar regions – part 1: product description. Cryosphere 8, 905–913. http://dx.doi.org/10.5194/tc-8-905-2014.

Chaubell, M.J., Chan, S.K., Dunbar, R.S., Peng, J., Yueh, S., 2016a. SMAP Enhanced L1C Radiometer Half-orbit 9 km EASE-grid Brightness Temperatures, Version 1. NASA National Snow and Ice Data Center Distributed Active Archive Center, Boulder, Colorado, USA.

Chaubell, M.J., Yueh, S., Entekhabi, D., Peng, J., 2016b. Resolution enhancement of SMAP radiometer data using the Backus Gilbert optimum interpolation technique. In: IEEE Int. Geosci. Remote Sens. Symp, pp. 284–287.

Colliander, A., McDonald, K., Zimmermann, R., Schroeder, R., Kimball, J.S., Njoku, E.G., 2012. Application of QuikSCAT backscatter to SMAP validation planning: freeze/thaw state over ALECTRA sites in Alaska from 2000 to 2007. IEEE Trans. Geosci. Remote Sens. 50, 461–468. http://dx.doi.org/10.1109/TGRS.2011.2174368.

Derksen, C., Xu, X., Scott Dunbar, R., Colliander, A., Kim, Y., Kimball, J.S., Black, T.A., Euskirchen, E., Langlois, A., Loranty, M.M., Marsh, P., Rautiainen, K., Roy, A., Royer, A., Stephens, J., 2017. Retrieving landscape freeze/thaw state from soil moisture active passive (SMAP) radar and radiometer measurements. Remote Sens. Environ. 194, 48–62. http://dx.doi.org/10.1016/j.rse.2017.03.007.

Draper, C., Reichle, R., de Jeu, R., Naeimi, V., Parinussa, R., Wagner, W., 2013. Estimating root mean square errors in remotely sensed soil moisture over continental scale domains. Remote Sens. Environ. 137, 288–298. http://dx.doi.org/10.1016/j.rse.2013.06.013.

Dunbar, S., Xu, X., Colliander, A., Derksen, C., Kimball, J.S., Kim, Y., 2016. Algorithm Theoretical Basis Document (ATBD): SMAP Level 3 Radiometer Freeze/Thaw Data Products (L3_FT_P and L3_FT_P_E). (No. Technical Document JPL D-56288).

Dunbar, R.S., Xu, X., Colliander, A., Derksen, C., Kimball, J.S., Kim, Y., 2017. SMAP Enhanced L3 Radiometer Northern Hemisphere Daily 9-km EASE-Grid Freeze/Thaw State, Version 1. NASA National Snow and Ice Data Center Distributed Active Archive Center, Boulder, Colorado, USA.

Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC Press.

Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., Kimball, J., Piepmeier, J.R., Koster, R.D., Martin, N., McDonald, K.C., Moghaddam, M., Moran, S., Reichle, R., Shi, J.-C., Spencer, M.W., Thurman, S.W., Tsang, L., Van Zyl, J., 2010. The soil moisture active passive (SMAP) mission. Proc. IEEE 98, 704–716. http://dx.doi.org/10.1109/JPROC.2010.2043918.

Farhadi, L., Reichle, R.H., De Lannoy, G.J.M., Kimball, J.S., 2014. Assimilation of freeze/thaw observations into the NASA catchment land surface model. J. Hydrometeorol. http://dx.doi.org/10.1175/JHM-D-14-0065.1.

Grosse, G., Goetz, S., McGuire, A.D., Romanovsky, V.E., Schuur, E.A.G., 2016. Changing permafrost in a warming world and feedbacks to the earth system. Environ. Res. Lett. 11, 40201. http://dx.doi.org/10.1088/1748-9326/11/4/040201.

Gruber, A., Su, C.-H., Zwieback, S., Crow, W., Dorigo, W., Wagner, W., 2016. Recent advances in (soil moisture) triple collocation analysis. In: Int. J. Appl. Earth Obs. Geoinformation. Advances in the Validation and Application of Remotely Sensed Soil Moisture - Part 1 45, Part B. pp. 200–211. http://dx.doi.org/10.1016/j.jag.2015.09.002.

Katul, G.G., Porporato, A., Daly, E., Oishi, A.C., Kim, H.-S., Stoy, P.C., Juang, J.-Y., Siqueira, M.B., 2007. On the spectrum of soil moisture from hourly to interannual scales. Water Resour. Res. 43. http://dx.doi.org/10.1029/2006WR005356.

Kerr, Y.H., Waldteufel, P., Wigneron, J.P., Delwart, S., Cabot, F., Boutin, J., Escorihuela, M.J., Font, J., Reul, N., Gruhier, C., Juglea, S.E., Drinkwater, M.R., Hahne, A., Martin-Neira, M., Mecklenburg, S., 2010. The SMOS mission: new tool for monitoring key elements of the global water cycle. Proc. IEEE 98, 666–687. http://dx.doi.org/10.1109/JPROC.2010.2043032.

Kim, Y., Kimball, J.S., McDonald, K.C., Glassy, J., 2011. Developing a global data record of daily landscape freeze/thaw status using satellite passive microwave remote sensing. IEEE Trans. Geosci. Remote Sens. 49, 949–960. http://dx.doi.org/10.1109/TGRS.2010.2070515.

Kim, Y., Kimball, J., Zhang, K., McDonald, K., 2012. Satellite detection of increasing northern hemisphere non-frozen seasons from 1979 to 2008: implications for regional vegetation growth. Remote Sens. Environ. 121, 472–487.

Koster, R.D., Suarez, M.J., 2001. Soil moisture memory in climate models. J. Hydrometeorol. 2, 558–570. http://dx.doi.org/10.1175/1525-7541(2001)002<0558:SMMICM>2.0.CO;2.

Le Vine, D.M., Lagerloef, G.S.E., Colomb, F.R., Yueh, S.H., Pellerano, F.A., 2007. Aquarius: an instrument to monitor sea surface salinity from space. IEEE Trans. Geosci. Remote Sens. 45, 2040–2050. http://dx.doi.org/10.1109/TGRS.2007.898092.

McColl, K.A., Vogelzang, J., Konings, A.G., Entekhabi, D., Piles, M., Stoffelen, A., 2014. Extended triple collocation: estimating errors and correlation coefficients with respect to an unknown target. Geophys. Res. Lett. 41, 2014GL061322. http://dx.doi.org/10.1002/2014GL061322.

McColl, K.A., Roy, A., Derksen, C., Konings, A.G., Alemohammed, S.H., Entekhabi, D., 2016. Triple collocation for binary and categorical variables: application to validating landscape freeze/thaw retrievals. Remote Sens. Environ. 176, 31–42. http://dx.doi.org/10.1016/j.rse.2016.01.010.

McColl, K.A., Alemohammad, S.H., Akbar, R., Konings, A.G., Yueh, S., Entekhabi, D., 2017a. The global distribution and dynamics of surface soil moisture. Nat. Geosci. 10 (2), 100–107. http://dx.doi.org/10.1038/ngeo2868. (February).

McColl, K.A., Wang, W., Peng, B., Akbar, R., Short Gianotti, D.J., Lu, H., Pan, M., Entekhabi, D., 2017b. Global characterization of surface soil moisture drydowns. Geophys. Res. Lett., 2017GL072819. http://dx.doi.org/10.1002/2017GL072819.

Podest, E., McDonald, K.C., Kimball, J.S., 2014. Multisensor microwave sensitivity to freeze/thaw dynamics across a complex boreal landscape. IEEE Trans. Geosci. Remote Sens. 52, 6818–6828. http://dx.doi.org/10.1109/TGRS.2014.2303635.

Rautiainen, K., Parkkinen, T., Lemmetyinen, J., Schwank, M., Wiesmann, A., Ikonen, J., Derksen, C., Davydov, S., Davydova, A., Boike, J., Langer, M., Drusch, M., Pulliainen, J., 2016. SMOS prototype algorithm for detecting autumn soil freezing. Remote Sens. Environ. 180, 346–360. (Special Issue: ESA's Soil Moisture and Ocean Salinity Mission - Achievements and Applications). https://doi.org/10.1016/j.rse.2016.01.012.

Reichle, R., De Lannoy, G., Koster, R.D., Crow, W.T., Kimball, J.S., 2016. SMAP L4 9 km EASE-grid Surface and Root Zone Soil Moisture Geophysical Data (No. Version 2). NASA National Snow and Ice Data Center Distributed Active Archive Center, Boulder, Colorado, USA.

Roy, A., Royer, A., Derksen, C., Brucker, L., Langlois, A., Mialon, A., Kerr, Y.H., 2015. Evaluation of spaceborne L-band radiometer measurements for terrestrial freeze/thaw retrievals in Canada. IEEE J. Sel. Top. Appl Earth Obs. Remote Sens. 8, 4442–4459. http://dx.doi.org/10.1109/JSTARS.2015.2476358.

Roy, A., Toose, P., Derksen, C., Rowlandson, T., Berg, A., Lemmetyinen, J., Royer, A., Tetlock, E., Helgason, W., Sonnentag, O., 2017. Spatial variability of L-band brightness temperature during freeze/thaw events over a prairie environment. Remote Sens. 9, 894. http://dx.doi.org/10.3390/rs9090894.

Schuur, E.A.G., McGuire, A.D., Schädel, C., Grosse, G., Harden, J.W., Hayes, D.J., Hugelius, G., Koven, C.D., Kuhry, P., Lawrence, D.M., Natali, S.M., Olefeldt, D., Romanovsky, V.E., Schaefer, K., Turetsky, M.R., Treat, C.C., Vonk, J.E., 2015. Climate change and the permafrost carbon feedback. Nature 520, 171–179. http://dx.doi.org/10.1038/nature14338.

Seneviratne, S.I., Koster, R.D., Guo, Z., Dirmeyer, P.A., Kowalczyk, E., Lawrence, D., Liu, P., Mocko, D., Lu, C.-H., Oleson, K.W., Verseghy, D., 2006. Soil moisture memory in AGCM simulations: analysis of global land–atmosphere coupling experiment (GLACE) data. J. Hydrometeorol. 7, 1090–1112. http://dx.doi.org/10.1175/JHM533.1.

Stoffelen, A., 1998. Toward the true near-surface wind speed: error modeling and calibration using triple collocation. J. Geophys. Res. 103, 7755–7766.

Varhola, A., Coops, N.C., Weiler, M., Moore, R.D., 2010. Forest canopy effects on snow accumulation and ablation: an integrative review of empirical results. J. Hydrol. 392, 219–233. http://dx.doi.org/10.1016/j.jhydrol.2010.08.009.

Vogelzang, J., Stoffelen, A., Verhoef, A., Figa-Saldaña, J., 2011. On the quality of high-resolution scatterometer winds. J. Geophys. Res. 116. http://dx.doi.org/10.1029/2010JC006640.

Williamson, M., Adams, J.R., Berg, A.A., Derksen, C., Toose, P., Walker, A., 2017. Plot-scale assessment of soil freeze/thaw detection and variability with impedance probes: implications for remote sensing validation networks. Hydrol. Res. http://dx.doi.org/10.2166/nh.2017.183. (March 3).

Yilmaz, M.T., Crow, W.T., 2014. Evaluation of assumptions in soil moisture triple collocation analysis. J. Hydrometeorol. http://dx.doi.org/10.1175/JHM-D-13-0158.1.

Zhang, T., Armstrong, R.L., Smith, J., 2003. Investigation of the near-surface soil freeze-thaw cycle in the contiguous United States: algorithm development and validation. J. Geophys. Res. Atmos. 108, 8860. http://dx.doi.org/10.1029/2003JD003530.