Multimodel Estimation of Snow Microwave Emission during CLPX 2003 Using Operational Parameterization of Microphysical Snow Characteristics

RAFAŁ WÓJCIK
Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

KONSTANTINOS ANDREADIS
Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington

MARCO Tedesco
NASA Goddard Space Flight Center, Greenbelt, and Goddard Earth Science and Technology Center, University of Maryland, Baltimore County, Baltimore, Maryland

ERIC WOOD AND TARA TROY
Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

DENNIS LETTENMEIER
Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington

(Manuscript received 27 March 2007, in final form 2 April 2008)

ABSTRACT

Existing forward snow emission models (SEMs) are limited by knowledge of both the temporal and spatial variability of snow microphysical parameters, with grain size being the most difficult to measure or estimate. This is due to the sparseness of in situ data and the lack of simple operational parameterizations for the evolution of snowpack properties. This paper compares snow brightness temperatures predicted by three SEMs using, as inputs, predicted snowpack characteristics from the Variable Infiltration Capacity (VIC) model. The latter is augmented by a new parameterization for the evolution of snow grain morphology and density. The grain size dynamics are described using a crystal growth equation. The three SEMs used in the study are the Land Surface Microwave Emission Model (LSMEM), the Dense Media Radiative Transfer (DMRT) model, and the Microwave Emission Model of Layered Snowpacks (MEMLS). Estimated brightness temperature is validated against the satellite [Advanced Microwave Scanning Radiometer for Earth Observing System (AMSR-E)] data at two sites from the Cold Land Processes Experiment (CLPX), conducted in Colorado in the winter of 2003. In addition, a merged multimodel estimate, based on Bayesian model averaging, is developed and compared to the measured brightness temperatures. The advantages of the Bayesian approach include the increase in the mean prediction accuracy as well as providing a nonparametric estimate of the error distributions for the brightness temperature estimates.

1. Introduction

Improved assimilation of atmospheric sounding data over land from sensors such as the Advanced Microwave Scanning Radiometer for Earth Observing Sys-
theory, the dielectric constant of frozen water is altered relative to that of water in its liquid form, and the effect of snow on the emissivity can be used in algorithms to estimate snow water equivalence (SWE) from spaceborne emissions, typically at 18.7 and 37 GHz. In practice, though, the emissivity of snow depends not only on SWE but also on snowpack microstructure, especially grain size and temperature.

These complications with SWE retrievals have motivated an alternative approach, which focuses on top of the atmosphere (TOA) emissions in the microwave frequencies and attempts to assimilate satellite radiance observations with model predictions rather than retrieving SWE. This approach is particularly relevant for applications such as the retrieval of atmospheric moisture in which the SWE problem is incidental. However, even where there is a motivation to update land surface variables, such as SWE, the assimilation of brightness temperatures (T_b), rather than derived the SWE products, requires knowledge of snow physical properties because they affect the (surface) emissivity. National Centers for Environmental Prediction (NCEP) operational models currently use the Community Radiative Transfer Model (CRTM), which predicts TOA microwave and infrared radiances and brightness temperatures. CRTM is generally much more sophisticated in its representation of atmospheric radiative transfer than of the land surface. The land surface emissivity in CRTM is based on snow depth and surface temperature, from which they use an empirical regression for grain size as inputs into the land emission model (Weng et al. 2001). Whether or not snow is present is determined by the output of the (Noah) land scheme snow depth prediction, which is updated daily using (in operations) a direct insertion approach. The observations are the Air Force Weather Agency (AFWA) global snow depth analysis (SNODEP), which is based on the interpolation of daily station reports (K. Mitchell 2007, personal communication).

The approach we propose to develop in this paper is a part of a broader National Oceanic and Atmospheric Administration–National Aeronautics and Space Administration (NOAA–NASA) Joint Center for Satellite Data Assimilation (JCSDA) sponsored project that is intended to incorporate recent improvements in snow emissions modeling into CRTM. The current approach for estimating snow emissivity in CRTM will be compared with one, or a combination of, forward snow emission models (SEMs) coupled with the Noah land surface scheme. Technically, the objective of the present contribution is twofold: i) to test a new methodology for estimating microphysical snow forcings by a land surface scheme and ii) to evaluate the feasibility of coupling the land surface scheme with SEMs in terms of accuracy of T_b predictions. The first objective is accomplished by improving the existing parameterization of snow physics in the Variable Infiltration Capacity (VIC) model (Liang et al. 1994). The new algorithm accounts for snow metamorphism, compaction from the weight of new snowfall, and an effective internal snowpack compaction. It also calculates the average snow grain size using a crystal growth equation. In the future, this parameterization is intended to be transferred into the operational Noah model to improve CRTM estimates of snow T_b. To accomplish the second objective, we evaluate the performance of three SEMs coupled with the VIC model using AMSR-E satellite observations of T_b (at 18.7, 37, and 89 GHz for both horizontal and vertical polarization) at two sites from the NASA Cold Land Processes Experiment (CLPX) in Colorado during the winter of 2003. The three SEMs are the Microwave Emission Model of Layered Snowpacks (MEMLS) of Mätzler and Wiesmann (1999), a modification of the Land Surface Microwave Emission Model (LSMEM) of Gao et al. (2004), and the Dense Media Radiative Transfer (DMRT) model of Tsang et al. (2000). Additionally, an example of the multimodel T_b estimate, based on Bayesian model averaging, is computed and compared to AMSR-E measurements. It demonstrates, using a bootstrap validation procedure, that the multimodel estimate increases the mean prediction accuracy as well as provides a nonparametric estimate of the error distributions.

As discussed above, the motivation of this paper is assessing the assimilating current satellite microwave brightness temperature (from AMSR-E) into microwave emission snow models (often referred to as forward models). The reader should recognize the multitude of challenges in this assimilation, which include the poorly posed problem of predicting microwave emissions as a result of their sensitivity to snow depth, the freeze–thaw cycle that affects snow grain morphology, the development of layering, and snowpack water. The challenge also includes scaling effects from the low-resolution (25 km) AMSR-E pixel resolution and the spatial heterogeneity in vegetation, topography, and snowfall. Therefore, this paper focuses mainly on the most favorable period within CLPX to best assess our ability to assimilate AMSR-E brightness temperatures.

This paper is organized as follows: section 2 provides a description of the new snow module implemented in VIC. The three SEMs used in this study are discussed in section 3. The mathematical formulation of the Bayesian model averaging is given in section 4. The CLPX 2003 measurements and the models’ setup are de-
scribed in section 5. The results are discussed in section 6. The paper closes with the summary and a short discussion of future research directions in section 7.

2. Variable Infiltration Capacity model

a. Basic description

VIC is a macroscale hydrology model that solves the energy and water balance over model grid cells that represent the modeled basin domain (Liang et al. 1994). Each grid cell can have multiple soil layers and be partially covered by different vegetation types in a mosaic-type representation, whereas topography is represented by a maximum of five elevation bands. Soil moisture storage capacity is characterized by a spatial probability distribution, although precipitation can also be spatially nonuniform. Baseflow is calculated as a nonlinear function of the lower soil layer moisture. Moisture and energy fluxes are computed separately for each vegetation class and elevation band within each grid cell and then area-weighted and summed over the grid cell, thus allowing the model to account for subgrid variability in topography, land cover, soil moisture, and precipitation. Streamflow is then simulated by routing subsurface and surface runoff using the method of Lohmann et al. (1998). Snow accumulation and ablation processes are simulated using a two-layer energy and mass balance approach (Cherkauer and Lettenmaier 2003). The surface layer is used to model the energy exchanges between the snowpack and the atmosphere while the lower layer acts as a reservoir for the excess snow mass from the surface layer (Wigmosta et al. 1994). Snowfall can be intercepted by an overstory canopy and then released to the ground snowpack through meltwater drip, mass release, or throughfall. The model accounts for melting/refreezing water within each layer, with water percolation being simulated based on a preset liquid water holding capacity for each layer.

b. Modification of snow module

In this study, a new snow densification algorithm in VIC (see Andreadis et al. 2008) was used to account for the significant effects of snow density and depth on microwave emissivity. Originally, VIC used a constant density of 50 kg/m³ for newly fallen snow and accounted for densification by compacting the snowpack with the weight of new snowfall. Rather than a constant value, we incorporated the algorithm by Hedstrom and Pomeroy (1998) that calculates the density of fresh snow based on air temperature. We also modified the snow densification algorithm by calculating a compaction rate according to Jordan (1991). This model accounts for settling as a result of snow metamorphism, compaction from the weight of new snowfall, and an effective internal snowpack compaction. In addition, a snow crystal growth algorithm was added to the model based on the one used by the snow thermal model SNTHERM (see Jordan 1991). This algorithm is based on the equation that describes growth by sintering in metals and ceramics:

$$d\frac{d}{dt} = a \exp(-b/T),$$ (1)

where $d$ (mm) is the average grain size, $T$ (K) is the temperature, and $a$, $b$ are adjustable parameters. For dry snow in SNTHERM, we considered using a simple function of the form

$$\frac{d}{dt} = \left( \frac{g_s}{d} \right) D_{s,0}(1000/P_s)(T_s/273.15)^6 C_{kT} \left[ \frac{\partial T_s}{\partial z} \right]$$ (2)

where $T_s$ (K) is snow temperature, $P_s$ (hPa) represents the atmospheric pressure, $z$ (m) is snow depth, and $g_s$ (m² kg⁻¹) is an adjustable parameter (here we used the value of $7 \times 10^{-7}$). The value of effective diffusion coefficient for water vapor in snow $D_{s,0}$ at 1000 hPa and 0°C is $0.92 \times 10^{-4}$ (m² s⁻¹), and the value of variation of saturation vapor pressure with temperature relative to phase $C_{kT}$ (kg m⁻² K⁻¹) can be calculated by Eq. (20) in Jordan (1991). Note that $d$ in (2) is expressed in meters. When implementing (2) in VIC, we approximated the absolute vertical thermal gradient $\partial T_s/\partial z$ by $|T_s - T_g|/\Delta z$, where $T_g$ (K) is ground temperature. For wet snow when liquid water content $\theta_l$ within the snowpack exceeds a threshold (0.0001), grain growth rate increases and, as proposed by Jordan (1991), is modeled using the similar growth function to that for dry snow:

$$\frac{d}{dt} = (g_s/d)(\theta_l + 0.005).$$ (3)

Here, $\theta_l$ is set to a maximum of 0.09 if it exceeds that value, and adjustable parameter $g_s$ is taken as $4 \times 10^{-12}$ (m² s⁻¹). The grain size estimated by the model is a depth-weighted average of the grain size of newly fallen snow (taken as 0.1 mm), and the grain size of the existing snowpack is obtained from (2) or (3).

3. The snow emission models

In this section, we provide a brief description of the three SEMs used in this study. For the sake of clarity, the key aspects of these models are listed in Table 1. Before characterizing each individual model, we introduce the inputs required by the models. The inputs are
nearly the same for all models and include snow depth, snow density, snowpack temperature, ground temperature, and surface roughness of the air/snow boundary, except that the LSMEM and DMRT models require the “mean grain size” of snow particles as an input, whereas the MEMLS model requires the “correlation length.” The mean grain size is taken as the radius of spherical ice particles approximating the ice grains in snow. The correlation length is related to snow grain size, shape, and volumetric distribution of snow grains (e.g., Jin 1993). However, this relationship is not straightforward (see, e.g., Pulliainen et al. 1999). To derive a value for the correlation length from the mean grain size, we multiplied the value of the mean grain size by the factor of 0.15 (C. Mätzler 2006, personal communication). This was done so comparisons can be made with results for models using either the mean grain size or a correlation length as inputs.

a. The Land Surface Microwave Emission Model

The LSMEM is a radiative transfer model that can predict the brightness temperature of a surface that can be partially covered with snow and/or vegetation using four different modules that account for emission from vegetation, bare soil, snow, and atmospheric effects (Gao et al. 2004). The snow emission model is based on the Helsinki University of Technology (HUT) model (Pulliainen et al. 1999) that treats the snowpack as a single homogeneous layer. The model describes the emission contribution of a snowpack as a function of snow depth, snow density, snow grain size, snow temperature, temperature at the snow–ground interface, frequency, and incidence angle. In addition to the upward emitted radiation, the model takes into account the contribution emitted downward and reflected upward from the snow/soil boundary, the emission contribution from underlying soil, and the atmospheric radiation reflected from the snow cover. The model also considers the multiple reflections caused by snow/soil and air/snow boundaries. The basic assumption in the HUT snow emission model is that scattering is mostly concentrated in the forward direction. The snow extinction coefficient is calculated from a modified empirical relationship and is a function of grain diameter and frequency (Hallikainen et al. 1987). The dielectric constants of ice and snow (permittivities) can be calculated using different optional models—here we used the model by Hallikainen et al. (1986).

b. The Dense Media Radiative Transfer model

The DMRT model describes the propagating and scattering of particles in a dense medium, which allows
the particles to occupy a fractional volume larger than 10%. In a nontenuous electrically dense medium, the dielectric properties of the particles are significantly different from those of the background medium, and the assumption of independent scattering is no longer valid because there is more than one scatterer within a wavelength distance. Under these conditions, the classical radiative transfer (CRT) theory is not valid. Several methods have been used to derive the DMRT equations including the effective field approximation (EFA), also called Foldy’s approximation (Tsang and Kong 1981); the quasi-crystalline approximation (QCA; Tsang and Kong 1981; Jin 1993); and the QCA with sticky particles (Tsang et al. 2000). In this study, we use the equations of the DMRT derived under the QCA approximation for moderate-size particles (Tsang et al. 2000). The Percus–Yevick equation is used to describe the pair distribution function (Tsang and Kong 1981). The effective propagation constant is computed on the basis of the generalized Lorentz–Lorenz law and the generalized Ewald–Oseen extinction theorem (e.g., Tsang et al. 2000). The extinction coefficient is then calculated from the imaginary part of the effective propagation constant. The formula used for ice permittivity is the same as the one used in the HUT model. The scattering and absorption coefficients are derived from, and the equations of the radiative transfer theory are solved by, the Gaussian quadrature method and the eigenvalues and eigenvectors technique (e.g., Jin 1993). The brightness temperatures are obtained by considering the boundary conditions, which provide the weights of the elements of the base of eigenvectors.

c. The Microwave Emission Model of Layered Snowpack model

In MEMLS (Mätzler and Wiesmann 1999; Wiesmann and Mätzler 1999), the snow cover is thought to be a stack of horizontal layers. Each layer is characterized by a thickness, a correlation length, its density, liquid water content, and temperature. The layer interfaces are assumed to be planar. The sandwich model, based on multiple scattering radiative transfer, is used (Wiesmann et al. 1998) to combine internal scattering and reflections at the interfaces. Internal volume scattering is accounted for by a two-flux model (up- and down-welling streams) derived from a six-flux approach (fluxes in all spatial directions). The absorption and scattering coefficients are functions of the six-flux parameters. The absorption coefficient can be obtained by density, frequency, and temperature, and the scattering coefficient depends on the correlation length, density, and frequency. The MEMLS model is based on the studies carried out by Wiesmann et al. (1998). The measurements of these authors lead to the empirical approach to determine the scattering coefficient of snow in the frequency range of 5–100 GHz and a correlation length range of 0.01–0.3 mm.

4. Bayesian model averaging

Bayesian model averaging (BMA) is a statistical method that infers from an ensemble of competing predictions the probabilistic prediction that possesses more skill and reliability than the original ensemble members (Raftery et al. 2003). BMA has been principally used in generalized linear regression applications. Recently (Raftery et al. 2003, 2005), it has been successfully applied to numerical weather prediction. In this study, we apply BMA to construct multimodel brightness temperature predictions using individual predictions from the three SEMs described in section 3. The BMA scheme is briefly described below.

a. Problem statement

Let $y$ be a scalar quantity to be forecast, and let $f_k$ be the forecast of $y$ produced by model $k$. The forecast $f_k$ is then characterized by a conditional probability density function (pdf), $g_k(y|f_k)$, which can be interpreted as the pdf of $y$ conditional on $f_k$, given that $f_k$ is the best forecast in the ensemble. The BMA predictive pdf for the $k$-member ensemble of forecasts can be written as the following mixture model (see Hoeting et al. 1999; Raftery et al. 2003, 2005):

$$p(y|f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k g_k(y|f_k),$$

where $w_k$ is the weight such that $\sum_{k=1}^{K} w_k = 1$. When predicting the brightness temperature, it is reasonable to assume that $g_k(y|f_k) \sim \mathcal{N}(\mu_k, \sigma^2_k)$. Note that although the latter conditionals are Gaussian, (4) is a nonparametric Gaussian mixture model (McLachlan and Peel 2000). The model is fitted to data sample $\chi = \{y_t, f_{1,t}, f_{2,t}, \ldots, f_{K,t}\}_{t=1}^{T}$, where the subscript $t$ denotes time, $f_{k,t}$ is referred to as the $k$th forecast in the ensemble for time $t$, and $y_t$ represents the corresponding verification. An estimate of the parameter vector $\theta = [w_1, \ldots, w_K, \mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K]^T$ is obtained by maximizing the log-likelihood function

$$L(\theta) = \sum_{t=1}^{T} \log \left[ \sum_{k=1}^{K} w_k g_k(y_t|f_{k,t}) \right].$$

Because (6) cannot be solved analytically, we solve it using the expectation–maximization (EM) algorithm.
The EM algorithm (see Dempster et al. 1977; McLachlan and Krishnan 1997) is an iterative procedure that alternates between two steps, the expectation (E) step and the maximization (M) step. It starts with an initial guess, \( \theta_0 \), for the parameter vector \( \theta \). In the E step, the posterior probabilities \( z_{k,t} \) are estimated given the current guess for the parameters. For the BMA in (4), the E step is

\[
\hat{z}_{k,t} = \frac{w_{k}^{(j-1)} g(y_t | \mu_k^{(j-1)}, \sigma_k^{(j-1)})}{\sum_{k=1}^{K} w_{k}^{(j-1)} g(y_t | \mu_k^{(j-1)}, \sigma_k^{(j-1)})},
\]

where the superscript \( j \) refers to the \( j \)th iteration of the EM algorithm, and \( g(y_t | \mu_k^{(j-1)}, \sigma_k^{(j-1)}) \) is a normal density with mean \( \mu_k^{(j-1)} \) and standard deviation \( \sigma_k^{(j-1)} \) evaluated at \( y_t \). The mean is modeled as a function, \( \phi_k \), that depends on \( f_{k,t} \) and a set of parameters, \( \beta_k \). We choose a local model that has linear parameters:

\[
\mu_{k,t} = \phi_k(f_{k,t}) = \sum_{i=1}^{I} \beta_{k,i} \phi_i(f_{k,t}).
\]

The other choice to make is \( \phi_k \), the form of the local model. In the absence of any prior information, a local linear model is a good choice [so \( \phi_1(f_{k,t}) = 1 \) and \( \phi_2(f_{k,t}) = f_{k,t} \)]. For simplicity of further notation, we define the posterior-weighted average of a quantity \( x \) as

\[
\langle x \rangle_k = \frac{\sum_{t=1}^{N} x_t \hat{z}_{k,t}^{(j)}}{\sum_{t=1}^{N} \hat{z}_{k,t}^{(j)}}.
\]

The M step then consists of estimating the \( w_k, \mu_k, \) and \( \sigma_k \) using the current estimates of \( z_{k,t} \) as weights. Thus,

\[
w_{k}^{(j)} = \frac{1}{N} \sum_{t=1}^{N} \hat{z}_{k,t}^{(j)},
\]

\[
a_{i_t} = \langle y \phi_{i_t}(f_k) \rangle_k, \quad i_t = 1, \ldots, I,
\]

\[
B_{i_1, i_2} = \langle \phi_{i_1}(f_k) \phi_{i_2}(f_k) \rangle_k, \quad i_1, i_2 = 1, \ldots, I,
\]

\[
\beta_k^{(j)} = \mathbf{B}^{-1} a,
\]

\[
\mu_k^{(j)} = \phi(f_{k,t}) \beta_k^{(j)}, \quad \text{and}
\]

\[
\sigma_k^{(j)} = \sqrt{\left\langle (y - \mu_k^{(j)})^2 \right\rangle_k}.
\]

The E and M steps are then iterated to convergence, which we defined as changes no greater than some small tolerance \( \varepsilon \) in the log-likelihood. The above algorithm provides a general setting for BMA with Gaussian conditionals. Sometimes, if the data sample \( \chi \) is scarce, some adjustments are needed to keep the number of parameters entering EM, that is \( \text{dim}(\theta) \), low compared to the sample size. A simplification that we incorporated here is that instead of estimating the coefficients \( \beta_k \) iteratively using (13), we fixed their values prior to EM by performing linear regression of \( y_t \) onto each \( f_{k,t} \), as in Raftery et al. (2005).

### c. BMA predictive mean and variance

The predictive mean and variance of (4) can be expressed as (Raftery et al. 2003)

\[
E(y_t | f_{1,t}, \ldots, f_{K,t}) = \sum_{k=1}^{K} w_k (\beta_{0,k} + \beta_{1,k} f_{k,t})
\]

and

\[
\text{VAR}(y_t | f_{1,t}, \ldots, f_{K,t}) = \sum_{k=1}^{K} w_k \beta_{0,k}^2 + \sum_{k=1}^{K} w_k \sigma_k^2.
\]

In essence, the expected BMA prediction is the average of individual predictions weighted by the likelihood that an individual model is correct given the observations. There are several attractive properties to the BMA prediction. First, the BMA prediction receives higher weights from better performing models because the likelihood of a model is essentially a measure of the agreement between the model predictions and the observations. Second, the BMA variance is essentially a second-order uncertainty measure of the BMA prediction. It contains the following two components: i) the between-model variance and ii) the within-model variance, as shown in the first and second terms of the right-hand side of (17). This measure is a better description of predictive uncertainty than that in any nonBMA scheme, which estimates uncertainty based only on the ensemble spread (i.e., only the between-model variance is considered) and, consequently, results in underdispersive predictions (Raftery et al. 2003).

### 5. Data description and model identification

Meteorological measurements were recorded in 10-min intervals at 10 sites within the CLPX small regional study area (SRSA) located in north-central Colorado (39.5°–41°N, 105°–107.5°W) between 20 September 2002 and 1 October 2003. In this study, we used data records from two meteorological towers within the Fraser mesocell study area (MSA), Fraser Alpine (FA)
and Fraser Experimental Forest headquarters (FHQ), from the period of 1 February to 31 May 2003. The dataset included air temperature, atmospheric pressure, relative humidity, wind speed and direction, and short-wave and longwave radiation as well as snow depth [see Feng et al. (2008) for a detailed description of these data]. The observed 10-min values were averaged to 1-h intervals to conform to VIC input requirements. It is important to stress that precipitation was not directly measured at the two CLPX sites. Therefore, for both FA and FHQ sites, we used hourly precipitation data based on 1/8th degree (~12 km) hourly merged gauge–radar precipitation product available from the North American Land Data Assimilation System (NLDAS). Driven by the above mentioned meteorological forcings, VIC was integrated with the time step of 1 h at both FA and FHQ sites. The three SEMs described in section 3 were then forced with the ground temperature, snow temperature, depth, density, and grain size from VIC. The estimates of $T_b$ were obtained for passive microwave frequencies at 18.7, 37, and 89 GHz for both horizontal (H) and vertical (V) polarization except for DMRT, which currently cannot handle a 89-GHz channel. Further, the $T_b$ simulations were restricted only to dry snow conditions. This was because the retrieval of SWE is not possible when liquid water content in snow increases as a consequence of melting. The penetration depth at microwave frequencies when snow is wet is of the order of a few centimeters as a consequence of the absorption coefficient, which limits our capability of deriving SWE/SD (where SD is snow depth) in wet snow conditions from microwave data in general. Moreover, LSMEM and DMRT versions used in this study can only estimate $T_b$ for dry snow. Next, the estimated $T_b$ was compared to the Aqua AMSR-E satellite data at the incidence angle of 55°. These data were gridded to the geographic (latitude–longitude) grids of the CLPX 2003 large regional study area (LRSA) and interpolated from swath space using inverse distance squared resampling (Brodzik 2003). The grid resolution was approximately 1/4° (~25 km), so both FHQ and FA sites are located within the same AMSR-E pixel.

6. Results

a. Comparison of snow depth

Before estimating $T_b$, we compared VIC predictions of snow depth $z$ (in cm) with the available in situ data. The results are shown in Fig. 1. At the FHQ site, the VIC predictions consistently capture the snow accumulation. However, peaks of snow deposition events are oversmoothed and underestimated compared to the measured snow depth. This is mainly a result of the scale mismatch; VIC was forced with a combination of in situ forcings and the NLDAS precipitation, so the estimated snow depth is at the mix of scales as opposed to the point scale in situ measurements. As a result, the timing of early summer snowmelt at the end of May 2003 is delayed. At the FA site, the spatial variability of snow depth is determined by both snowdrift and snowmelt (see Feng et al. 2008). The former process is not accounted for in the current version of VIC. This, together with the scale mismatch, causes a significant overestimation of snow depth. In addition, Fig. 1 shows the fluctuations in VIC-predicted average grain size $d$ (in mm). Because hourly time series of measured grain size are not available for the considered study period, the grain size predictions can only be evaluated qualitatively. Especially for the FHQ site, the crystal growth equation (see section 2b) implemented in VIC consistently describes two processes that govern the snowpack dynamics: i) metamorphism (snow grain coalescence reducing the voids in between, causing grain size to increase) and ii) compaction (reduction in average grain size as a result of new snowfall). Metamor-

1 Andreadis et al. (2007) compared the grain size from VIC with sparse local scale observation site (LSOS) snowpit measurements. The results suggest that, on average, VIC tends to underestimate the in situ measured grain size.
Phenology is shown in the upper panel of Fig. 1 during several snowmelt episodes, and compaction is shown during several snow accumulation episodes.

b. Comparison of brightness temperatures

Brightness temperature predictions at 18.7 and 37 GHz, computed by means of the LSMEM, MEMLS, and DMRT models; and at 89 GHz, by means of the LSMEM and MEMLS models; at both the FHQ and FA sites were first corrected for tree emissions in the analyzed AMSR-E pixel using the following formula:

\[
T_b = fT_{b,\text{tree}} + (1 - f)T_{b,\text{model}},
\]

where \(f\) represents fractional tree cover, \(T_{b,\text{tree}}\) is the brightness temperature of trees and \(T_{b,\text{model}}\) is raw estimate of brightness temperature from a particular SEM. Here, Fraser area \(f = 0.53\), and \(T_{b,\text{tree}}\) was taken

---

**Fig. 2.** AMSR-E measured vs simulated brightness temperatures at 18.7, 37, and 89 GHz (V) obtained with LSMEM, MEMLS, and DMRT models using VIC-predicted forcing data at FHQ and FA sites (February–May 2003).

**Fig. 3.** VIC predictions of average temperature of snowpack at FHQ (black line) and FA (gray line) sites (February–May 2003).
as the mean of the distribution of brightness temperatures for trees (see Tedesco et al. 2005 for details), that is, 268 K at 18.7 GHz and 271 K at 37 GHz. At 89 GHz, the value of $T_{b,\text{tree}}$ was assumed to be 272 K. Figure 2 shows $T_b$ estimates for FHQ site together with the corresponding AMSR-E measurements (ascending and descending overflights are grouped together).

It is clear that for 18.7 and 37 GHz, all three SEMs overestimate AMSR-E brightness temperature in the beginning of the study season [1 February–13 March 2003; time index 0–1000 (hour)]. Although there are many factors that can contribute to overestimation, one factor is that the value of $T_{b,\text{tree}}$ might be too high, which could occur if the trees were snow covered. Another problem, especially in the case of 37 GHz $T_b$, is that the grain size estimates by VIC might be too low.

On the other hand, the overestimation is not present at 89 GHz (both horizontal and vertical). This can be explained considering that, in general, values of $T_b$ at 89 GHz are strongly influenced by the diurnal freeze–thaw cycles at the surface of snowpack. The surface is usually cooler and dryer than deeper layers of snowpack. Note that in Fig. 2, the variability in both estimated and measured $T_b$ decreases with the decrease of the frequency. Additionally, 37- and 89-GHz channels are generally more sensitive to the attenuating influence of the snowpack than the 18.7-GHz channel, which is influenced by the soil underlying the snowpack, especially for thin snowpacks. For these two channels, LSMEM estimates of $T_b$ exhibit large fluctuations compared to the other two models. This can be attributed to the sensitivity of LSMEM predictions of $T_b$ to the variability in the av-
c. BMA and multimodel brightness temperature prediction

The differences between observed and modeled snow $T_b$ can be due to the following: i) the misrepresentation of snow properties (especially gain size) by the models at spatial scales is equivalent to the AMSR-E surface footprint size, ii) the AMSR-E $T_b$ is actually a combination of $T_b$ contributions from the snow, ground, vegetation cover, and atmosphere, and iii) instrumental and input errors. Individual SEMs capture only some of these uncertainties and then probably only partially, depending on the frequency channel used and the adequacy of parameterization of snow emission physics. Accordingly, by trying to select the “best” SEM out the multimodel ensemble, some sources of uncertainty captured by the remaining SEMs might be ignored, thus underestimating the total uncertainty in $T_b$ estimates. To tackle this problem, we constructed multimodel AMSR-E $T_b$ predictions using the BMA technique described in section 4. Given the difficulties at FA described in the previous section, the BMA analysis is performed only for the FHQ site.

1) Bootstrap validation for BMA

The simplest approach for assessing the performance of the BMA predictions is to define an error measure, split the entire data sample deterministically into a training set and an independent testing set, fit the BMA using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set. Using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set. Using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set. Using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set. Using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set. Using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set. Using the training set, and estimate the error on the testing set. Because the total number of available AMSR-E measurements in our study period was only 193, this approach is problematic because the results would highly depend on the (deterministic) way the data would have been split. To resolve the problem, we used the following bootstrap validation procedure. By randomly sampling without replacement from the original data, we generated $B = 30$ independent replicates of the training data (145 points, which is 75% of the total sample size) and the testing data (48 points, which is 25% of the total sample size). Then, we refitted the BMA scheme to each of the training replicates and estimated the error on the testing set.
the bivariate set of estimates of the BMA predictive mean in (16) and associated AMSR-E verifications in the $b$th bootstrap testing set, then the estimate of the average bootstrap prediction error for the BMA scheme is

$$\text{Err} = \frac{1}{B} \sum_{b=1}^{B} e(Y^{*b}),$$

where $N_{\text{cal}}$ denotes the number of points in the $b$th testing set, and $e(\cdot)$ is the loss function for measuring
errors. A typical choice of the loss function is the root mean squared error (RMSE),
\[
\varepsilon(y^{*b}) = \sqrt{\frac{1}{N_{\text{val}}} \sum_{r_{\text{val}}=1}^{N_{\text{val}}} (y_{r_{\text{val}}}^{*b} - \hat{y}_{r_{\text{val}}}^{*b})^2}.
\] (20)

In addition, we calculated the mean absolute error (MAE),
\[
\varepsilon(y^{*b}) = \frac{1}{N_{\text{val}}} \sum_{r_{\text{val}}=1}^{N_{\text{val}}} |y_{r_{\text{val}}}^{*b} - \hat{y}_{r_{\text{val}}}^{*b}|,
\] (21)

which is similar to the RMSE but is less sensitive to large forecast errors. For small or limited datasets, the use of MAE is sometimes preferred.

2) BMA RESULTS

To initialize the EM fitting procedure, for each frequency channel and for each polarization, the variances of the conditional Gaussian pdfs on the right-hand side of (4) were taken as variances of individual SEMs \(T_b\) estimates, respectively. Figure 5 shows an example of the fitted conditionals and predictive BMA pdf of 37-GHz (vertical) \(T_b\) at FHQ site at 0800 UTC on 20 March 2003. Note that these estimates were obtained by running the EM on one of the replicates of the bootstrap training set. There was a disagreement among the raw ensemble member predictions: two of them (LSMEM and MEMLS) were around 260 K, whereas the other one (DMRT) was around 270 K.

This difference of 10 K is quite large. After estimating the ensemble member conditional means (hereafter the bias-corrected \(T_b\) estimates), the raw SEMs predictions were shifted toward verifying the AMSR-E observation (cf. raw SEMs predictions with the means of the Gaussians in Fig. 5). The latter turned out to be outside the raw ensemble range. The resulting BMA predictive mean slightly overestimates AMSR-E \(T_b\). However, it is much more accurate than the raw SEMs estimates. Because each MEMLS and DMRT bias-corrected \(T_b\) underestimates the verifying observation and LSMEM \(T_b\) overestimates it, the BMA predictive pdf is positively skewed. The values of the weights of the model-conditional components of this pdf were 0.6 for LSMEM, 0.36 for DMRT, and 0.04 for MEMLS. The small weight for the MEMLS component is due to the high correlation (0.87) between LSMEM and MEMLS \(T_b\) predictions. This implies that if the LSMEM prediction is known, the additional information from the MEMLS prediction is much less than it would be if the two predictions were uncorrelated. Figure 6 shows the time evolution of one bootstrap realization of BMA predictive pdf, its mean, and the standard deviation envelopes at the FHQ site for all analyzed frequencies and polarizations. It is clear that the variability of the AMSR-E data is well described by the BMA standard deviation envelopes. Moreover, the BMA predictive mean approximates AMSR-E \(T_b\) better than raw SEMs predictions in Fig. 2. This is confirmed in Table 2, which

<table>
<thead>
<tr>
<th></th>
<th>18.7 GHz (H)</th>
<th></th>
<th>18.7 GHz (V)</th>
<th></th>
<th>37 GHz (H)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Testing</td>
<td>Training</td>
<td>Testing</td>
<td>Training</td>
</tr>
<tr>
<td></td>
<td>RMSE MAE</td>
<td>RMSE MAE</td>
<td>RMSE MAE</td>
<td>RMSE MAE</td>
<td>RMSE MAE</td>
</tr>
<tr>
<td>BMA</td>
<td>8.63 7.14</td>
<td>9.03 7.47</td>
<td>6.44 5.11</td>
<td>6.78 5.44</td>
<td>11.29 9.37</td>
</tr>
</tbody>
</table>

2) TABLE 2. Comparison of mean performance statistics RMSE and MAE (K) for BMA and raw SEMs predictions of \(T_b\). The means are taken over 30 bootstrap realizations of training and testing data.

2) TABLE 3. Comparison of standard deviations (K) of the RMSE and MAE for BMA and raw SEMs predictions of \(T_b\). The estimates are based on 30 bootstrap realizations of training and testing data.
compares the average bootstrap values of $\text{RMSE}$ and $\text{MAE}$ of BMA predictive mean with raw SEMs predictions. The values are stratified according to frequency and polarization and refer to the training and testing set averages separately. In general, the values of both $\text{RMSE}$ and $\text{MAE}$ for the BMA predictive mean are lower than those for the raw individual SEMs predictions. The BMA scores are better for vertical polarization than those for horizontal polarization for all frequencies. Moreover, as expected theoretically, $\text{MAE}$ estimates are slightly smaller than $\text{RMSE}$ estimates. This indicates the sensitivity of the latter statistic to outliers. The highest improvements in the prediction accuracy are found for the 37-GHz (V) channel. Here, the differences between the corresponding estimates of $\text{RMSE}$ and $\text{MAE}$ for the BMA predictive mean and the best performing models (LSMEM and MEMLS) on the testing set are about 3.5 and 3 K, respectively. The best prediction accuracy of the BMA scheme is found for 18.7-GHz (V) channel. The testing set estimates of $\text{RMSE}$ and $\text{MAE}$ are as low as 6.78 and 5.44 K, respectively. This is due to the relatively small variability in the corresponding AMSR-E measurements. Conversely, for 89 GHz in which the variability is large, the errors in the BMA predictions are large too. With regard to the errors of raw SEMs predictions, the best performing model is LSMEM, with slightly better $\text{RMSE}$ and $\text{MAE}$ scores than MEMLS and much better scores than DMRT for 18.7 (H), 37 (H), and 37-GHz (V) frequency, respectively. Another important feature of the results in Table 2 is the reasonable agreement between the mean error estimates for the training set samples and the testing set samples. This indicates that on average, the BMA model is robust against overfitting. To further examine the robustness of the $\text{RMSE}$ and $\text{MAE}$, we estimated their standard deviations. These results are listed in Table 3. Clearly, the dispersion around the mean value in the bootstrap samples of $\text{RMSE}$ and $\text{MAE}$ for the training set is much smaller than in those for the testing set. This is because the size of the testing set is only 25% of the total number of AMSR-E measurements, which in turn makes the bootstrap error estimates on the testing set more vulnerable to sampling effects, to the presence of outliers, and to the nonstationarity in AMSR-E data as a result of the snowmelt event at the end of the study season. On the other hand, the standard deviations of both $\text{RMSE}$ and $\text{MAE}$ of the BMA scheme never exceed $\pm 1$ K, which confirms the robustness of the results in Table 2.

### 7. Summary and outlook

In this paper, brightness temperatures of snow at 18.7 and 37 GHz were simulated by three SEMs (LSMEM,

<table>
<thead>
<tr>
<th>37 GHz (V)</th>
<th>89 GHz (H)</th>
<th>89 GHz (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Testing</td>
<td>Training</td>
</tr>
<tr>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>14.31</td>
<td>11.97</td>
<td>14.37</td>
</tr>
<tr>
<td>15.98</td>
<td>13.57</td>
<td>16.03</td>
</tr>
</tbody>
</table>

* Not available.

<table>
<thead>
<tr>
<th>37 GHz (V)</th>
<th>89 GHz (H)</th>
<th>89 GHz (V)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>Testing</td>
<td>Training</td>
</tr>
<tr>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>0.27</td>
<td>0.27</td>
<td>0.77</td>
</tr>
<tr>
<td>0.27</td>
<td>0.27</td>
<td>0.84</td>
</tr>
<tr>
<td>0.27</td>
<td>0.27</td>
<td>0.88</td>
</tr>
<tr>
<td>0.38</td>
<td>0.38</td>
<td>1.20</td>
</tr>
</tbody>
</table>

* Not available.
MEMLS, and DMRT) and at 89 GHz by two SEMs (LSMEM and MEMLS) coupled with the VIC land surface scheme using data from the FHQ and FA sites collected during the CLPX 2003 experiment. The objective of this study was to assess the feasibility of the coupling in terms of the quality of \( T_b \) predictions. Because we simulated hourly time series of \( T_b \) for the period of February–May 2003, the natural way to validate our simulations was to compare them to satellite AMSR-E measurements. All the analyzed SEMs had common inputs predicted by VIC. The latter was driven by a combination of in situ and NLDAS forcings, which introduced a scale discrepancy between the SEMs predictions and the AMSR-E observations. In the beginning of the study season at 18.7 and 37 GHz, all the models overestimated the AMSR-E \( T_b \) at the FHQ site. This may have been caused either by trees covered by snow in the AMSR-E pixel or by an underestimation of the average snow grain size. Also, the signature of early summer snowmelt event in the AMSR-E measurements was not well captured by the SEMs at all analyzed frequencies. The reason for that is because in our investigation, we only considered dry snow conditions. Thus, \( T_b \) was underestimated compared to the \( T_b \) signature of wet snow that would be measured by AMSR-E. In general, except for the 89-GHz channel, the variability range of AMSR-E measurements of \( T_b \) was not well captured by SEMs. In particular, MEMLS and DMRT produced oversmoothed \( T_b \) estimates. LSMEM predictions, on the other hand, turned out to be sensitive to the average snow temperature fluctuations. This was particularly pronounced at 37 and 89 GHz. The penetration depth of the latter frequency channel is small, so effective fluctuations in snowpack surface temperature govern the dynamics of \( T_b \). Our simulations were less successful at the FA site because the spatial variability of snow depth was highly influenced by snowdrift. This phenomenon is not accounted for in the current version of VIC, so the simulation of the microphysical snow properties at FA was of poor quality. As a result, the simulated \( T_b \) exhibited some artifacts as explained in section 6a.

Apart from considering the individual SEMs predictions, we proposed a new multimodel procedure for estimating the AMSR-E brightness temperature of snow based on an ensemble of SEMs (LSMEM, MEMLS, and DMRT) coupled with the VIC land surface scheme. The procedure is based on BMA and offers not only an adequate nonparametric description of the \( T_b \) predictive pdf, but it also improved the \( T_b \) prediction accuracy compared to individual SEMs results. This is because, to some extent, BMA reduces the effects of scaling, model error sources, atmospheric contribution to radiative transfer, and sparseness of AMSR-E measurements—problems that are notorious when comparing satellite snow \( T_b \) with SEMs predictions. Using a simple bootstrap validation procedure, we have shown that the BMA predictive mean outperformed the predictions of individual SEMs in terms of RMSE and MAE scores, particularly for 37 GHz. The results for 18 GHz (H) were also encouraging. Another interesting feature of BMA results was its low variability (less than 1 K) around the bootstrap estimates of the aforementioned error measures, which confirms the robustness of the presented results. Despite the increased precision of BMA estimates of \( T_b \) compared to SEMs predictions, our preliminary results have shown that the errors of BMA were still significant.

There are a number of issues that need further investigation to improve the BMA \( T_b \) estimates to make them useful in hydrometeorological practice. i) Additional research is needed to better understand the influence of atmospheric moisture profiles on \( T_b \) estimates from SEMs. An attractive option here would be to implement the ensemble of SEMs as a microwave surface module of CRTM and run it in the coupled mode with atmospheric absorption and atmospheric scatter modules. ii) It is critical to further understand the influence of the mean grain size (or correlation length in the case of MEMLS) on the SEMs performance. Because the measurements of this parameter are not usually collected systematically, perhaps a better option would be to compute an “effective” grain size by simply calibrating this parameter to produce \( T_b \) estimates that match AMSR-E \( T_b \) as closely as possible. Note that this effective value would automatically account for any potential scaling mismatch. Some preliminary results of the parameter calibration approach for LSMEM—applied for soil moisture retrievals though, not for snow \( T_b \) estimation—are given in Pan et al. (2006). And, (iii) the final issue concerns the use of BMA predictive pdfs in data assimilation. These pdfs could, for example, be applied to construct the nonparametric likelihood operators in particle filters, offering enhanced quality of the updates of variables that describe snowpack evolution in land surface models.

**Acknowledgments.** This research was possible through support from the NOAA Joint Center for Satellite Data Assimilation (resulting in the Development of Improved Forward Models for the Retrieval of Snow Properties using EOS-era Satellites proposal) to Princeton University (Agreement NA04NES4400002) and to the University of Washington (Agreement NA04NES4400003). This support is gratefully acknowledged.
REFERENCES


