



Water Resources Research

RESEARCH ARTICLE

10.1029/2018WR024521

Key Points:

- An innovative method of incorporating observations into distributed, multilayer, physics-based snow models has been developed
- Persistent biases were dynamically and efficiently detected leading to improvements in both current-day and future forecast simulations
- The technique dynamically adjusts to changing conditions, conserves mass, and maintains physically consistent model states

Correspondence to:

A. Winstral,
adam.winstral@slf.ch

Citation:

Winstral, A., Magnusson, J., Schirmer, M., & Jonas, T. (2019). The bias-detecting ensemble: A new and efficient technique for dynamically incorporating observations into physics-based, multilayer snow models. *Water Resources Research*, 55, 613–631. <https://doi.org/10.1029/2018WR024521>

Received 4 DEC 2018

Accepted 6 DEC 2018

Accepted article online 14 DEC 2018

Published online 25 JAN 2019

The Bias-Detecting Ensemble: A New and Efficient Technique for Dynamically Incorporating Observations Into Physics-Based, Multilayer Snow Models

A. Winstral¹ J. Magnusson², M. Schirmer¹, and T. Jonas¹

¹WSL Institute for Snow and Avalanche Research SLF, Davos Dorf, Switzerland, ²Norwegian Water Resources and Energy Directorate, Oslo, Norway

Abstract The reliance on distributed energy- and mass-balance snow models as runoff forecasting tools has been increasing. Compared to traditional, conceptual forecasting approaches, these physics-based tools are robust to conditions that deviate from historic norms and offer improved performance in potentially dangerous rain-on-snow events. The physics-based simulations, however, depend on a large suite of accurate forcing data. Current numerical weather prediction products are capable of supplying the full range of required data, but systematic biases are often present. Data assimilation presents a means of compensating for such errors as well as potential snow model errors, yet currently available data assimilation techniques have limited usefulness in these snow models. This study introduced an alternative technique that similarly uses observations to update and improve simulations. As such, it is the first method to incorporate point snow observations in a fully distributed, physics-based, multilayer snow model while conserving mass and maintaining physically consistent layer states that are in accord with observations. At the core of this technique is an ensemble of predetermined perturbations to the model forcings termed the bias-detecting ensemble. Ensemble members were evaluated using observed snow depths to ascertain potential biases at nearly 300 sites across Switzerland. The bias assessments were distributed to 38 independent sites and incorporated into the model. Tests were conducted over three winter seasons using two numerical weather prediction-based products with varying quality. Averaged across the 38 sites and 3 seasons, the Nash-Sutcliffe efficiency score for bias-detecting ensemble-corrected snow depth was 0.98 compared to 0.81 without the bias-detecting ensemble method.

1. Introduction

Distributed energy- and mass-balance snow models are increasingly being applied in operational runoff forecasting (e.g., Hedrick et al., 2018; Magnusson et al., 2017; Painter et al., 2016). The increased dependence on physics-based solutions in these models makes them an attractive alternative to the simpler, conceptual models traditionally relied upon. The conceptual models typically have a historic component (e.g., regressions or parameter fittings) and are based on a smaller set of input variables. These simpler models are capable of producing accurate forecasts in a computationally efficient manner (Anderson, 1973; Franz et al., 2008). However, the reliability of these models comes into question when conditions are encountered that differ from the historical norms that they are based on (Dozier, 2011; Milly et al., 2008). The need for physically-based solutions becomes increasingly important as climate changes affect traditional snowfall and snowmelt patterns (Dozier, 2011; Nayak et al., 2010) and increase the likelihood of extreme events such as rain-on-snow (Freudiger et al., 2014; Musselman et al., 2018; Surfleet & Tullos, 2013). The latter can be particularly difficult to accurately simulate using temperature-based solutions (Kumar et al., 2013).

The classic limitations of the physics-based solutions have been their high computational demands and increased dependence on a wider array of spatially distributed meteorological forcings (Anderson, 1976). Today, increased computer processing capabilities have helped to overcome the former, yet the quality and accuracy of the latter remain a limiting factor. Observations of precipitation, wind, humidity, and radiation are often scarce and rarely capture the scale of known heterogeneities, particularly in mountain regions (Bales et al., 2006; Strachan et al., 2016). Numerical weather prediction models (NWP) are increasingly available, yet even the most highly resolved (e.g., 1-km resolution) are

incapable of fully capturing hydrologically relevant accumulation and melt heterogeneities (Blöschl, 1999; Winstral et al., 2014). Furthermore, NWP forecasts invariably contain errors, some of which are systematic due to forcing data quality, computational limitations, or faulty boundary conditions (Dee, 2005; Schirmer & Jamieson, 2015; Sweeney et al., 2013). Another potential error source in the physics-based solutions is the models themselves. Essery et al. (2013) and Rutter et al. (2009) have shown that largely due to necessary parameterizations of complex and poorly understood physical processes (e.g., turbulent energy fluxes, snow albedo, and internal hydrologic properties), different energy-balance models can exhibit a large spread in modeled outcomes even when forced with the same input data.

Data assimilation (DA) techniques, whereby observations are used to update model simulations, are commonly used to overcome such errors particularly in operational modeling. DA techniques can vary from simple approaches that directly insert observations and force the model to the observed state(s) to more advanced statistical techniques often based on Bayes' theorem that account and correct for errors in the entire modeling chain from forcing data to model parameterizations and physics. These more complex methods can also account for potential errors in the observations used to update the model and to spatially correlate errors to improve model performance at sites without observations (Reichle & Koster, 2003). DA has a rich history of applications in hydrologic (e.g., Clark et al., 2008; Fan et al., 2017; Moradkhani et al., 2012; Rasmussen et al., 2016; Reichle et al., 2002) and atmospheric models (e.g., Auligné et al., 2007; Harris & Kelly, 2001). Many of the more advanced DA approaches are based on derivatives of the Kalman filter (Kalman, 1960), which assumes that model responses are relatively linear. One example of such a development is the ensemble Kalman filter (Evensen, 2003). Observation-based updates for unobserved states and sites are a function of covariances and averaging operations conducted across an ensemble of model outcomes. These approaches apply observations either with or without their potential errors to directly update model states at the end of each evaluation period. As further detailed below, this can lead to physical inconsistencies and subjective modeling choices when these methods are applied to the dynamic layering system present in complex snow models.

Distributed, complex, multilayer snow modeling in mountain environments presents unique challenges to ensemble-based DA approaches. Snow distributions exhibit a high degree of spatial heterogeneity that is the product of nonlinear, interacting processes. Additionally, snow physics models are composed of a dynamically evolving set of snow layers. While hydrologic and atmospheric applications generally have a static number of layers, whose boundaries have fixed positions in space or pressure levels, multilayered snow models have dynamic layering properties whose numbers and thicknesses exhibit spatiotemporal variability reflective of the aforementioned heterogeneities. These layers actively interact with one another and can be added, removed, merged, and split through time. This is especially pronounced in avalanche forecasting models that can have as many as 50 (CROCUS; Brun et al., 1992) to hundreds (SNOWPACK; Bartelt and Lehning, 2002) of layers. Whereas snow hydrology models are relatively simpler—often having maximums of two to four layers—how each boundary layer interacts with the soil below and atmosphere above is governed by internal layer dynamics. For instance, snowpacks with similar bulk properties but dissimilar internal dynamics will have differing responses to the same boundary forcings. Thin and thick snowpacks with the same bulk temperatures will likewise respond differently to identical forcings. These heterogeneities make transference or distribution of bulk or averaged state properties for even commonly available observations such as snow depth (HS) and snow-water-equivalence (SWE) a nontrivial matter. Updating rarely observed states like snow temperatures and liquid water contents that are weakly correlated with HS and SWE can introduce large uncertainties. Since snowpack runoff during potential rain-on-snow flood events (e.g., Rössler et al., 2014; Wayand et al., 2015) is largely controlled by antecedent snow temperatures, liquid water contents, and remaining water holding storages (Trubilowicz & Moore, 2017; Würzer et al., 2016), it is vital that these snow states, which are rarely measured, are accurately rendered in snowmelt runoff simulations.

Nevertheless, DA remains a powerful tool for snow modelers. Liston et al. (1999), Fletcher et al. (2012), and Xu and Shu (2014) used the simple DA method of direct insertion of observation-based spatial products into spatial snow models. In these applications, the snow models were forced to observations that varied from HS to SWE to snow-covered-area (SCA). Optimal interpolation has been used to update snow depths (Brasnett, 1999), solid precipitation fluxes (Magnusson et al., 2014), and SWE (Liston & Hiemstra, 2008); however, this method has static, predetermined error structures (Slater & Clark, 2006) and like direct insertion is only appropriate for updating observed states or fluxes (Magnusson et al., 2014). The ensemble Kalman filter (EnKF)

introduces a means of establishing covariances between states and observations that enable the unobserved state variables to be updated as well. The EnKF can also account for potential errors in the observations and models. The EnKF has been used to incorporate point SWE observations (Slater & Clark, 2006), point HS observations (Magnusson et al., 2014), and satellite retrievals of SCA and SWE (Andreadis & Lettenmaier, 2006) into snow models. Sun et al. (2004) used a one-dimensional (i.e., observations available at every model point), extended Kalman filter to assimilate remotely sensed, passive microwave SWE into a land surface model. The United States National Weather Service assimilates ground-based, airborne, and satellite data into their operational mass- and energy-balance snow model (SNODAS); however, these procedures remain largely unpublished.

Most of the aforementioned snow DA studies used conceptual or single-layer snow models where the need to account for layering complexities and updating unobserved states was limited. Though Slater and Clark (2006) employed a single-layer snow model, it still tracked nine internal states. They noted that in allowing the EnKF to update unobserved states using ensemble averages and covariances, one must set bounds in order to ensure that physical principals are not violated. One modeling necessity was to fix the ratios of the multiple water storage terms in the background (i.e., prior) and updated states. The snow submodel in the Sun et al. (2004) application had up to three snow layers and tracked three state variables: SWE, cold content, and snow depth. They found that estimated correlations from the EnKF between observed SWE and the unobserved states were not well defined due to the dynamically changing snow layers. In a similar fashion to Slater and Clark (2006), they fixed the updated snow temperatures and densities to their background, model-predicted values. These background values were then used in conjunction with the updated SWE to update the unobserved states (i.e., cold content and snow depth). Fixing water storages or snow temperatures to their modeled values may only weakly affect seasonal runoff evaluations, but they can have strong implications on short-term snowmelt forecasts (see Trubilowicz & Moore, 2017; Würzer et al., 2016). The influences of assumptions such as these are magnified when considering increasingly more complex snow models. Yet another drawback common to all of these methods is that mass is not necessarily conserved. In cases where SWE is reduced after assimilation, it cannot be ascertained whether this was due to incorrect precipitation inputs or underestimated melt—an important consideration when linkages to a hydrological model are present.

In evaluating DA techniques compatible with the CROCUS avalanche forecasting model, Charrois et al. (2016) concluded that the dynamic snow layering and nonlinear physics did not lend itself to the assumptions and state-averaging methods at the core of the EnKF. Due to this reasoning, they chose to use a particle filter (PF) to assimilate reflectances in a synthetic modeling experiment. One essential advantage of the PF is that each particle or ensemble member maintains layering and states physically consistent with the snow model. Dechant and Moradkhani (2011) found the PF to be more effective than the EnKF in assimilating brightness temperatures to improve SWE predictions and streamflow modeling using the single-layer SNOW-17 model. They attributed the improved performance of the PF to being robust to the Gaussian error assumptions and linear updating schemes present in the EnKF. Leisenring and Moradkhani (2011) and Magnusson et al. (2017) reached similar conclusions regarding the advantages of the PF in respectively assimilating SWE into SNOW-17 and snow depth into the multilayer JULES Investigation Model (JIM) snow model. The PF, however, has been limited to applications where observations are available; an efficient means of extrapolating these successes to sites lacking observations has yet to be developed. Computational costs associated with PFs can also be considerable due to the large number of particles required.

Given the drawbacks of traditional DA techniques in this modeling realm, this research sought to develop a new method more appropriate for distributed, multilayer, physics-based snow models. The method incorporates the information contained in observations to remove potential biases in the NWP forcing data and snow model constructs. Through adjustments to the model forcings rather than updating model states, mass is conserved, and physical consistency in layering dynamics and modeled states is maintained. The technique bears some resemblance to those of Magnusson et al. (2014) and Vögeli et al. (2016) who also used snow observations to update model forcings. The former, however, employed a conceptual snow model while the latter used only a single lidar retrieval to correct integrated seasonal precipitation forcings to the physics-based SNOWPACK model. In the presented methodology, we use continuously available data from automatic weather stations for independent, daily assessments and updates of both precipitation and energy biases. Station observations in combination with model simulations are used to discriminate snow

accumulation and melt days to obtain the independent bias assessments. There are also some methodological similarities to the PF approach, but rather than a set of random model forcings used to evaluate and adjust posterior states, we apply prescribed perturbations and assess the forcing perturbations that produce the best states.

The ensemble of perturbations given to the snow model is referred to as the bias-detecting ensemble (BDE). The BDE methodology shares some features of the bias-corrected ensemble (BCE) approach for correcting NWP air temperatures and wind speeds introduced by Stensrud and Yussouf (2005), yet there are substantial differences. The bias-corrected ensemble was based on bias correcting multimodel ensembles of NWP data using direct observations of air temperature and wind. The BDE approach is an indirect method of bias-correcting precipitation and energy snow model forcings using observed snow data and evaluation of the predefined ensemble of perturbations.

The BDE procedure addresses a pressing need in the snow modeling community and is appropriate for both operational and research applications. It has the capabilities to dynamically adjust to any changes in the system, whether these are functional changes in the forecasting procedures (e.g., updated processing procedures and changing boundary conditions), seasonal adjustments necessitated by the shifting prominence of energy terms, or episodic events such as albedo-reducing dust storms (e.g., Skiles et al., 2012). The procedure takes into consideration the full modeling system integrating the effects of model forcings and snow model parameterizations correcting for systematic errors in the combined product without specifically determining the error source. In this study we present the BDE methodology, perform BDE evaluations at one set of observational sites, and test the spatiotemporal resiliency of these bias evaluations at a set of independent sites. The sites comprise a broad network of observational data available throughout Switzerland. Two sets of input data with a wide range in accuracy were evaluated to test the capabilities of the BDE approach: (1) 2 years of native, 24-hr lead time forecast data from a 2-km resolution NWP model and (2) 1 year of hindcast reanalysis data from a 1-km resolution NWP product with air temperatures and precipitation forcings further enhanced with additional station observations.

2. Data and Snow Model

2.1. Point Observations

The observational data consisted of data from 334 sites with daily monitored snow depth (HS) across Switzerland (Figure 1). These stations are part of either SLF's (WSL Institute for Snow and Avalanche Research) IMIS (Intercantonal Measurement and Information System) or the MeteoSwiss (Federal Office for Meteorology and Climatology) snow monitoring networks. HS measurements were extracted at a daily time-step and were manually edited where necessary. Implausible data (assessed against neighboring stations at similar elevations) were removed and either replaced with interpolated data from nearby stations or entered as not available. Of these observation sites, 38 also had manual measurements of SWE taken approximately every 2 weeks. Most of the manually measured sites were sampled on a consistent bimonthly basis whereas others were measured in a more sporadic fashion. In the presented analysis, the 296 sites that do not have SWE data are referred to as the HS stations. The BDE bias assessments, whereby observations were incorporated into the modeling framework to evaluate potential precipitation and energy biases in the modeling chain, were conducted using data from these HS stations. The 38 stations that had both automated HS and manual SWE data are referred to as the SWE stations. The SWE stations were treated as an independent set that were not part of the BDE evaluations and were used only for validation purposes. The SWE stations were used to determine whether HS-station bias assessments could improve simulations at these representative “unobserved sites.” The range of elevations of the HS stations is 225–2,950 m and that of the SWE stations is 1,195–2,690 m (Figure 4). Whereas the elevations of the HS stations nicely envelope those of the SWE stations, the HS stations are spatially spread across Switzerland with the SWE stations concentrated in the Alps (Figure 1).

For modeling purposes, the observations were considered error free. The assumption of error-free observations is analogous to direct insertion methods yet contrasts traditional DA methods that include observational errors in the updating framework. The data used here are from established and well-maintained flat field sites where errors should be dominated by instrument limitations. Manufacturer-stated accuracies of the sonic depth sensors are $\pm 1\text{--}2$ cm depending on instrument height and should occur randomly. By

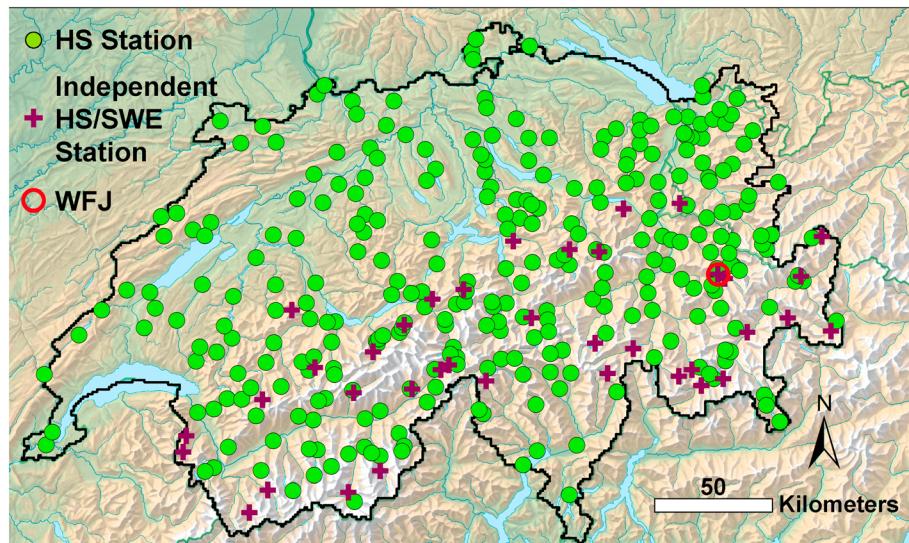


Figure 1. Map of Switzerland with the HS stations where biases were assessed and the snow-water-equivalence (SWE) stations where the bias-detecting ensemble was evaluated. WFJ is the Weissfluhjoch station referenced in the text.

comparison NWP forcing data often contain systematic biases (e.g., Dee, 2005; Schirmer & Jamieson, 2015) that can lead to simulated snow depth (SWE) errors of tens of centimeters (millimeters) (Magnusson et al., 2014, 2017) exhibiting spatial coherency (e.g., Magnusson et al., 2014). The BDE technique is currently set up to assess these larger scale biases under the assumption that observational errors are minimal in comparison. Liston and Hiemstra (2008), Vögeli et al. (2016), Hedrick et al. (2018), and others have shown that incorporating snow data without accounting for observational errors can still greatly improve snow model outcomes. The BDE technique is certainly capable of accounting for observational errors that could further enhance and refine this methodology. However, this is left for future research with the current work serving as an initial proof of concept. The station data used in this application are not representative of the heterogeneity of snow conditions encountered in alpine terrain (Molotch & Bales, 2006), but they are analogous to the conditions represented in the NWP forcing data and those where snow models are typically developed and tested (e.g., Anderson 1976; Bartelt & Lehning, 2002; Vionnet et al., 2012; Essery et al., 2013). Hence, the introduced procedures should accurately reflect biases in these modeling aspects. Capturing the spatially variable patterns present across the landscape is left to other techniques (e.g., Havens et al., 2017; Helbig et al., 2015; Winstral et al., 2009, 2013).

2.2. Forcing Data

The NWP forcing data were Consortium for Small-Scale Modeling (COSMO) products. The COSMO models are nonhydrostatic, limited-area atmospheric prediction models (www.cosmo-model.org). COSMO forecast products are used in operational applications throughout Switzerland by agencies including MeteoSwiss and the Swiss Federal Office for the Environment. The COSMO models provide all of the forcing data required of most energy balance snow models. In WY14 (i.e., the 2013–14 winter season) and WY15, model forcing data consisted of 24-hr lead time forecasts from the COSMO-2 product (horizontal resolution ~2.2 km). These products are representative of what is typically available in forecast situations and are identical to the forcing data applied in Magnusson et al. (2017). A third year, WY17, was also included in the analysis, with the intent of establishing a best-case scenario of model forcings to further test the capabilities of the BDE process. The foundation of the WY17 forcing data were hindcasts taken from the COSMO-1 analysis model (horizontal resolution ~1.1 km), which became fully operational in March 2016. The COSMO-1 forcing data were further enhanced with observation-based precipitation and air temperature reanalysis products detailed below. The COSMO data as well as the RhinesD raster product mentioned below were linearly interpolated to the observation points for modeling purposes. Downscaling of temperature products included a fixed linear lapse rate (0.006 °C/m).

Table 1

Average Monthly Air Temperatures Based on Station Observations Upscaled to a National 1-km Grid

	Nov	Dec	Jan	Feb	Mar	Apr	May	Mean
WY14	0.7	0.1	-0.2	0.0	3.2	6.3	7.8	2.6
WY15	4.6	-0.2	-1.5	-2.9	2.2	5.6	9.5	2.5
WY17	2.0	0.3	-4.9	0.8	4.5	4.5	10.0	2.5

RhiresD gridded daily precipitation products and solid precipitation inferred from observed snow depths were used to enhance the COSMO-1 precipitation product in WY17. RhiresD is a gridded daily precipitation product produced by MeteoSwiss using data from approximately 440 precipitation gauges across Switzerland. The RhiresD gauge network, however, underrepresents elevations above 1,200 m (Frei & Schär, 1998). Daily solid precipitation mass fluxes, inferred from observed snow depth changes (Magnusson et al., 2014) at the HS/SWE stations (Figure 4) complemented the lower elevation RhiresD data and further

enhanced the WY17 precipitation forcings. Daily increases in snow depth were converted to mass fluxes using a density model based on Martinec and Rango (1991). The original Martinec and Rango model, based on snow settling and densification over time, has been modified and recalibrated using data from over 10,000 snow profiles (Jonas et al., 2009). Daily precipitation for WY17 was ultimately determined by passing the RhiresD data through a temperature index snow model that assimilated the observation-inferred solid precipitation data (see Magnusson et al., 2014). The daily-integrated COSMO-1 hourly data were then forced to match the daily analysis fields from the aforementioned DA. To further enhance the air temperature forcings, daily mean air temperature observations from approximately 250 stations across Switzerland were interpolated to the COSMO-1 grid. As with the precipitation forcings, the integrated hourly COSMO-1 data were forced to match the daily observational product. In the WY17 test, where snow depths from both the HS and SWE stations were used to enhance the snow model precipitation and temperature forcings, the sensitivity of the BDE bias evaluations should have greater sensitivity to COSMO radiation fluxes, COSMO winds, and snow model structures.

2.3. Water Year Descriptions

In order to assess relative climatic conditions during the three evaluation years, SWE data taken from a Swiss-wide, 1-km operational product (based on the model framework described by Magnusson et al., 2014) were assessed. Averaged across the entire country, SWE in WY14 and WY15 were very close to the 20-year average (1998–2017; Figure 3). WY17 had well below average wintertime SWE yet had similar melt rates in May and June in comparison to the other years. Though WY14 and WY15 had similar spatial averages, WY15 tended to have less snow in the regions south of the central spine of the Alps and more to the north. Interannual SWE distribution correlation coefficients based on grid cells with snow in either year's 1 April grids were 0.92, 0.82, and 0.96 for WYs14/15, WYs14/17, and WYs15/17, respectively. Despite monthly air temperature differences between the 3 years, mean November thru May differences were less than 0.1 °C (Table 1).

2.4. JIM Snow Model

The physics-based, mass- and energy-balance snow model used in this application consisted of single configurations contained within the multimodel framework of JIM (Essery et al., 2013). JIM was originally designed to test the sensitivity of simulations to process parameterizations and models the snowpack with up to three snow layers. There are a total of 1,701 potential combinations of different process parameterizations available in JIM. The model configuration for WY14 and WY15 was based on an optimization to snow, lysimeter, and radiation observations at the Weissfluhjoch site (2,536 m) in Switzerland (Figure 1), whereas the WY17 model configuration was based on a broader optimization to observations throughout Switzerland. Both model configurations can be found below with a full description of JIM available in Essery et al. (2013)

- Snow Compaction:
 - WY14/15: Physical snow compaction parameterization (Anderson, 1976)
 - WY17: Time-based empirical snow compaction parameterization (Versegheghy, 1991)
- Fresh snow density:
 - All: A function of air temperature (Anderson, 1976)
- Snow Albedo
 - WY14/15: Time-based empirical parameterization with a linear decay rate for cold snow and an exponential decay rate for melting snow (Douville et al., 1995)
 - WY17: Same as above but with decay parameters calibrated to Swiss-wide observations as a function of elevation

- Surface energy exchanges
 - All: parameterization based on Richardson number (Louis, 1979)
- Snow cover fraction
 - All: Since these were all point simulations, not implemented
- Snow hydrology
 - All: Excess water drains immediately from layer when maximum liquid water mass fraction is reached (Anderson, 1976; Boone & Etchevers, 2001)
- Thermal conductivity of snow
 - All: Power function of snow density (Douville et al., 1995)

3. The BDE Method

3.1. Overview

This research was designed to test if snow depth observations can be used to detect potentially systematic NWP and snow model biases to improve snow simulations. The JIM snow model was forced with the data sets described in section 2.2 at the full set of HS stations. Simulated and observed HS were compared on a daily basis at the HS stations to evaluate biases. These daily biases were then back-averaged to establish greater consistency and spatially distributed to the set of independent SWE stations. If the detected biases are representative of persistent biases in the modeling system, they should exhibit spatiotemporal consistencies (e.g., Reichle & Koster, 2003; Shrestha et al., 2013) that can be used to improve simulations at sites where observations are not available. JIM was subsequently run at the SWE stations with the same forcing data and the additional bias information gathered at the HS stations. Analyses and conclusions were based on the impact the estimated biases had on model performance at the SWE stations. Evaluations of 5-day lead time forecasts with and without the bias adjustments were also made. Experiments were conducted for two different input data scenarios: (1) point forcing data down-scaled from the COSMO-2 24-hr forecasts and (2) forcing data down-scaled from the hindcast COSMO-1 reanalysis products further enhanced with observations from a dense network of automated weather stations across Switzerland. The expectation was that the former contains larger errors and biases than the latter and that the latter would provide a much more rigorous test of the presented approach. In this section, the BDE methodology is presented in general terms followed by descriptions of how the derived biases were incorporated into the snow model at the SWE sites. A flow chart depicting all the described steps in the procedure can be found in Figure 2.

At the core of the BDE technique is an ensemble of fixed perturbations to the forcings of the JIM snow model. Potential biases in the snow modeling framework including forcing data and snow model are simplified to mass and energy terms that were independently evaluated at the HS stations. Biases in the mass forcings were evaluated in terms of fixed multipliers to the precipitation forcings. Precipitation multipliers (P) ranged from 0.6 to 1.75 in increments of 0.05. The energy flux perturbations were an additional energy flux added to the energy balance calculations. The energy perturbation (EB) ranged from -75 to 75 W/m^2 in 5 W/m^2 increments. In total there were 744 ensemble members based on all combinations of the precipitation and energy perturbations. Perturbations remained static over 24-hr periods. At the end of each day, the model inputs, outputs, and observations were used to determine if the day would be considered for BDE evaluation and whether the precipitation or energy terms would be assessed (section 3.2). Evaluation day decisions (i.e., P , EB , or neither) were station-specific and mutually exclusive to limit potential equifinality or compensating effects (e.g., P and EB terms interacting to produce similar results). The performance of the ensemble members was evaluated by comparing simulated and observed HS (section 3.3). Model states for simulations at the HS sites were then adjusted to optimize the subsequent day's bias assessments (section 3.4). At the end of each day, the library of daily and back-averaged biases (section 3.5) at the HS stations was updated. The back-averaged biases were then spatially distributed to the SWE sites (section 3.6) for the bias-enhanced simulations and validation of the methodology.

3.2. Daily Criteria for Bias Evaluations

Data assessments at each HS site were made at the end of each day to determine which, if either, of the BDE perturbation terms would be evaluated. A day met the initial criteria for P bias consideration if either the

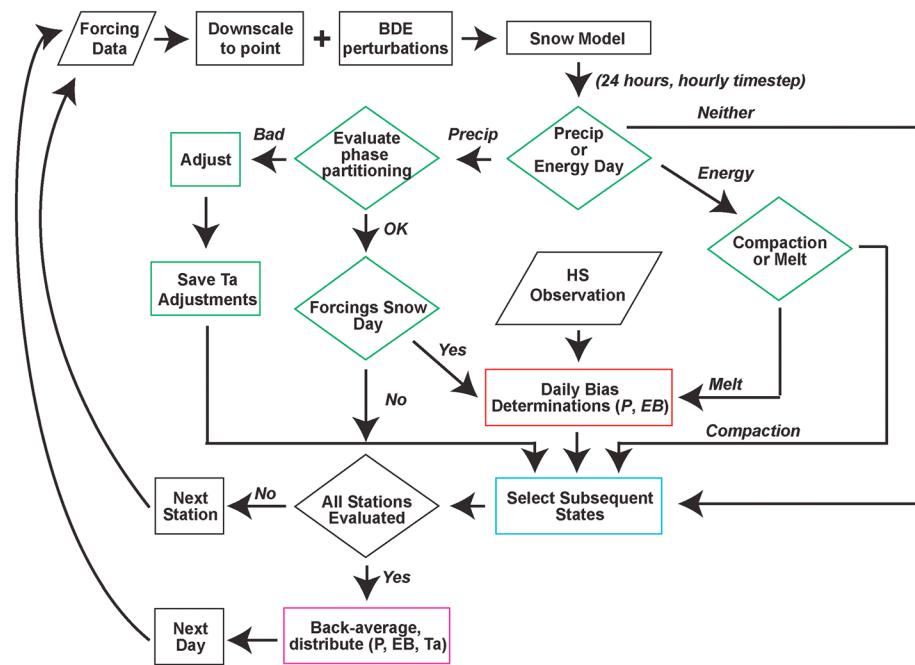


Figure 2. Flow chart depicting the bias-detecting ensemble (BDE) process. Green outlined steps are described in section 3.2, red in section 3.3, light blue in section 3.4, and magenta in sections 3.5 and 3.6.

forcing data or the observations indicated a snow event occurred. On the forcing data side, if daily precipitation forcings multiplied by the back-averaged P bias (section 3.4) at the site was greater than 2.0 mm and the daily air temperature was less than 1 °C, it was considered a snow day. If there was an observed daily increase of 2.0 cm or more, then the day would similarly be considered a potential snow day. If neither of these stipulations were met and there was a decrease in observed HS, the EB biases were considered for evaluation. If the day did not fall into either of these two categories, biases were not estimated at the site. Additional criteria, further detailed below, were also required before a daily bias evaluation was made.

A day would meet the final criteria for P bias evaluation only if the snow-day forcings criteria were met and precipitation phase partitioning errors were not evident. The former was necessary to ensure that there was a sufficient snow mass present in the simulation to be evaluated. The latter consideration ensured that P and EB biases were not applied as a means of compensating for errors attributable to either the temperature forcing data or the simulated phase partitioning. A wrongfully simulated snow day would occur when the forcing data indicated snowfall but observed HS did not increase. A wrongfully simulated rain day would occur when the forcing data did not meet the snow day criteria yet the observational data did (i.e., increase in HS of >2.0 cm). To increase the signal to noise ratio, an additional stipulation for the precipitation phase errors to be triggered was there had to be at least 5.0 mm of simulated precipitation (including the back-averaged P bias multiple). In instances of wrongfully simulated snow (rain) all hourly station air temperatures that day were increased (decreased) in 0.1 °C increments until the simulated snow:rain ratio was 1:3 (3:1 or until the solid precipitation mass with an assumed density of 100 kg/m³ matched the observed increase in HS). The required temperature adjustments were stored for day-specific interpolation to unobserved sites, which in this experiment were the SWE sites. Once simulated precipitation phases reached the aforementioned criteria, the temperature adjustments were removed from the forcing data at the HS stations. Forcing temperatures remained at their original values so that they did not influence future evaluations of the otherwise non-temperature-corrected EB biases.

Compaction events were removed from the BDE evaluations of EB since such events, especially when occurring during cold, midwinter conditions, can lead to gross manipulations of the energy term in order to compensate for differences in simulated and modeled HS. A day was considered a compaction day when there was a decrease in observed HS yet BDE ensemble members that included up to an additional 10 W/m² of energy flux produced less than 1 mm of meltwater leaving the snowpack.

3.3. Daily Bias Determinations

The performance of the ensemble members was based on comparisons between simulated and observed snow depths. Depending on ensemble member performance in the prior timestep, the comparison was based on either bulk snow depths or the daily changes in snow depth (explained in section 3.4). On snow accumulation days, the best performing precipitation perturbations were determined while keeping the energy balance offset fixed at its prior (i.e., back-averaged, explained in section 3.5) forecast value. Likewise on *EB* evaluation days, the precipitation offset was fixed at its prior value and the performance of the *EB* perturbations evaluated. All ensemble members within 1.0 cm HS of the best performing member were considered with the most conservative member (e.g., P closest to 1.0; EB closest to 0) chosen. This helped to ensure that small simulation differences did not produce extreme bias evaluations. Whenever the P or EB terms were not evaluated, daily bias estimations for these terms were not recorded. When biases were evaluated, chosen members as well as the precipitation mass were stored in a library for subsequent backaveraging.

3.4. Select Model States for Subsequent Timestep

On snow accumulation and melt days when BDE members were evaluated as described above, the best-performing BDE member was further evaluated to determine continuing model states (e.g., snow layer temperatures, densities, solid/liquid masses, etc.). This evaluation step also included an assessment of whether BDE evaluations in the subsequent timestep should be based on comparisons of bulk HS or the daily change in HS. If the simulated bulk HS for the best-performing member was within 1.0 cm of the observation, simulated HS was directly matched to the observation. This match was obtained by either adding or subtracting SWE to the simulation using the characteristics of the uppermost snow layer. The snowpack was then relayered using the JIM layering routines. With a direct match between simulated and observed HS heading into the next evaluation day, the evaluation of best-performing members was based on daily fluxes. It is important to note that the induced changes in SWE were only done at the HS stations and only in order to optimally evaluate biases. Proper and accurate estimations of biases at each HS site were the only goal of the procedure at this point. In an operational setting, the biases would be first determined in this manner and then distributed across the modeling domain. The model would then be re-run with the native forcings adjusted with the biases, mimicking what was done at the SWE sites here.

If the current best-performing member was not within 1.0 cm of the observed bulk HS, then it was decided that the BDE perturbations were not sufficient to match the simulation with the observation at this timestep. Hence, the continuing states were those of the current best-performing member, and any HS differences were carried into the next timestep. In this case, bulk HS was used to evaluate the BDE members in the subsequent evaluation allowing the BDE methodology to “catch up” for the prior inadequacies.

In all cases where the BDE members were not evaluated, the continuing states of the model were based on those of the ensemble member with the prior back-averaged P and EB biases at that site. Simulated bulk HS was adjusted to match the observation using the same procedures described above, and daily fluxes were used in the subsequent evaluation.

3.5. Back-averaging

Since the BDE procedure was developed to detect and correct systematic biases, a certain degree of time-stationarity in the biases was expected. Implementing back-averaged values added a degree of stability to the modeling framework. Yussouf and Stensrud (2007) in the development of their bias-corrected ensemble similarly back-averaged systematic biases of air temperature and wind speed in NWP models while testing lookback periods of up to 30 days. They found that a 12-day lookback period yielded good results and that window lengths of up to 30 days yielded little or no improvement. They also employed an exponential weighting to give more weight to more recent observations. For application of the precipitation biases in this study, a maximum 14-day lookback period was used. The 14-day lookback consisted of only the last 14 days when a P bias assessment was made. Rather than a time-weighted average, precipitation biases during this period were mass-weighted to determine the present day's bias. The *EB* lookback period was a much shorter 5 days. This was done to give the *EB* offset greater flexibility to adjust to the variety of processes and their seasonal fluctuations that are subsumed into this term. For example, inadequacies in shortwave radiation or albedo modeling might not become apparent until the spring period, whereas longwave and turbulent

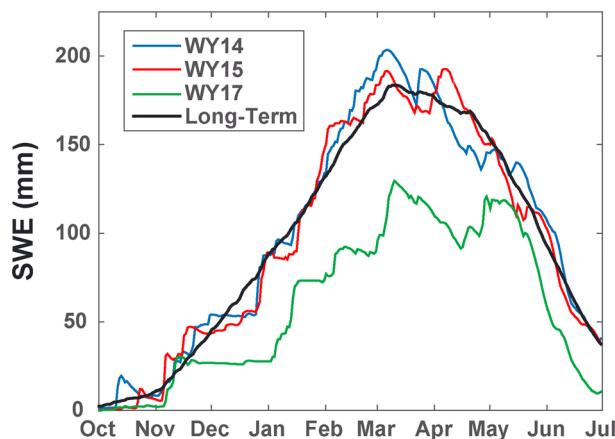


Figure 3. Daily averaged snow-water-equivalence (SWE) across all of Switzerland for the three modeling years compared to the 20-year average. Data taken from WSL-SLF Operational Snow Hydrology Service products.

from 0 (insignificant correlation between elevation and biases; $p > 0.05$) to 60 ($R = 1$). For example, when $R = 1$, a station located 1 km away with an elevation difference of 100 m (distance = 1,000 m + $60 \times \Delta\text{elev}$) would have the same weighting as a station located 7 km away but with no elevation difference. Or if $R = 0.5$, the former station with a 100 m elevation difference would be weighted equally with a station located approximately 4 km distant with no elevation difference.

The same procedure was used for distributing the air temperature adjustments when repartitioning of precipitation phases was necessitated. In this case, however, the adjustments were considered event specific and just the daily values were distributed. Likewise, a minimum of 25 stations requiring temperature adjustments to adjust precipitation phase were needed to trigger the transference of temperature biases to the independent SWE sites.

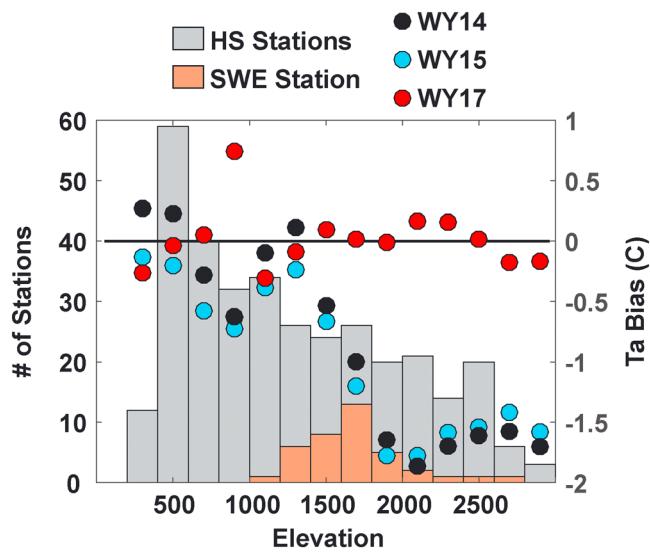


Figure 4. Temperature biases (filled circles) by elevation of the forcing data used in the simulations are depicted on the right y axis. These data are overlaid on the distribution of HS and snow-water-equivalence (SWE) stations by elevation depicted on the left y axis. There were 164 (191) stations available for the temperature bias assessments in WY14 (WY15). The erratic behavior of the temperature biases in the 900-m elevation bin is due to the fact that only one station is located in that bin. By contrast, the 1,700 thru 2,700 m bins all have greater than 10 sites.

errors would be more prone to predominate earlier in the snow season. Back-averaged EB values simply consisted of the mean of available data. If less than three daily bias values were available—for either the P or EB biases—the relevant back-averaged station bias remained at its default value (e.g., $P = 1.0$; $EB = 0.0$), and these biases were not included in the distributions to the SWE stations. If less than 14, but at least 3 daily biases were available for the back-averaging, back-averaging was limited to the available data.

3.6. Distribution to Independent Sites

The back-averaged HS-station P and EB biases were distributed to the independent SWE sites and used to bias-correct same-day forcings at the latter. If less than 25 stations with valid back-averaged biases were available, the HS-station biases were not distributed, and biases at the SWE stations remained at their default values. The spatial distribution of HS-station biases used a basic distance-weighted Gaussian interpolation from the 15 nearest HS stations taking northing, easting, and altitude into account (Magnusson et al., 2014). Weighting of elevation distances varied linearly

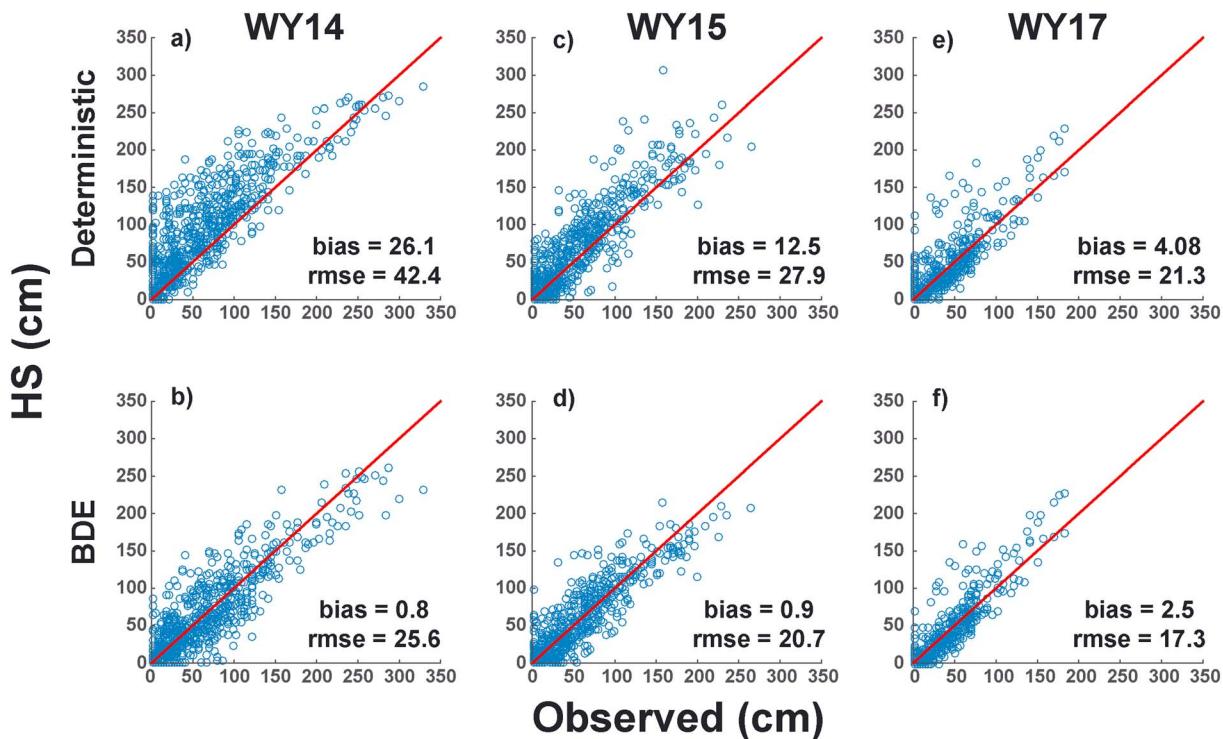


Figure 5. Scatterplots of daily observed and modeled snow depths (HS) for the deterministic and bias-detecting ensemble (BDE) model runs at the snow-water-equivalence stations in WY14, WY15, and WY17. Figures only show 10% of the available data to aid visualization. Bias and root mean square error (rmse) metrics are for the full data set.

Overall, the deterministic runs had stronger biases and larger errors in WY14 than in WY15 (Figures 5a, 5c, 6a, and 6c). (Note that all error metrics in this work were evaluated only when either observed or simulated SWE/HS was greater than zero.) The simulations at the independent SWE stations that benefitted from the HS-station bias evaluations virtually eliminated the strong biases present in both the WY14 and WY15 deterministic runs (Figures 5b, 5d, 6b, and 6d). SWE (HS) root mean square errors (rmse) were reduced by 37% (40%) in WY14 and by 18% (17%) in WY15. The interyear differences in BDE percentage gains are perhaps attributable to the differing performances of the deterministic runs in the 2 years. Performance gains were noticeable in both the accumulation and melt phases with the seasonal split determined by simulated peak SWE in the deterministic model at each station (Figure 7). Though certainly some of the improved performance during the melt season shown in Figure 7 is attributable to accumulation season improvements, rmse differences between the deterministic and BDE runs were even greater during the melt phase.

Figure 8 clearly demonstrates some of the strengths and weaknesses of the BDE methodology. The main strength of this approach is the ability to dynamically ascertain and correct broad trends and biases in the combined forcings/model product. In Figures 9a and 9d, the daily averaged observed and simulated HS aggregated over all 38 SWE stations are presented. All statistical metrics for these daily averaged data were vastly improved (Table 1) with Nash-Sutcliffe efficiency scores (Nash & Sutcliffe, 1970) above 0.98 for the BDE simulations in both years. Figures 9b, 9c, 9e, and 9f show results on a station-by-station basis. In both years, the BDE simulations improved performance at a majority of the stations. Similar to the interannual results (Figures 5 and 6; Table 2), relative improvements were greater in WY14. However, in correcting the broad trends apparent in these data, BDE improvements at stations with reasonable deterministic simulations (e.g., lower rmse) or with biases not consistent with the overall trend (e.g., deterministic HS/SWE not greater than observed) were less likely. By contrast, substantial BDE improvements were consistently observed at stations with poor deterministic performance, and these strongly contributed to the overall improvements. For example, in WY15 the overall HS (SWE) bias was reduced from 12.5 to 0.9 cm (51 to -2 mm), yet the absolute HS bias was only reduced at 53% of the stations (note that

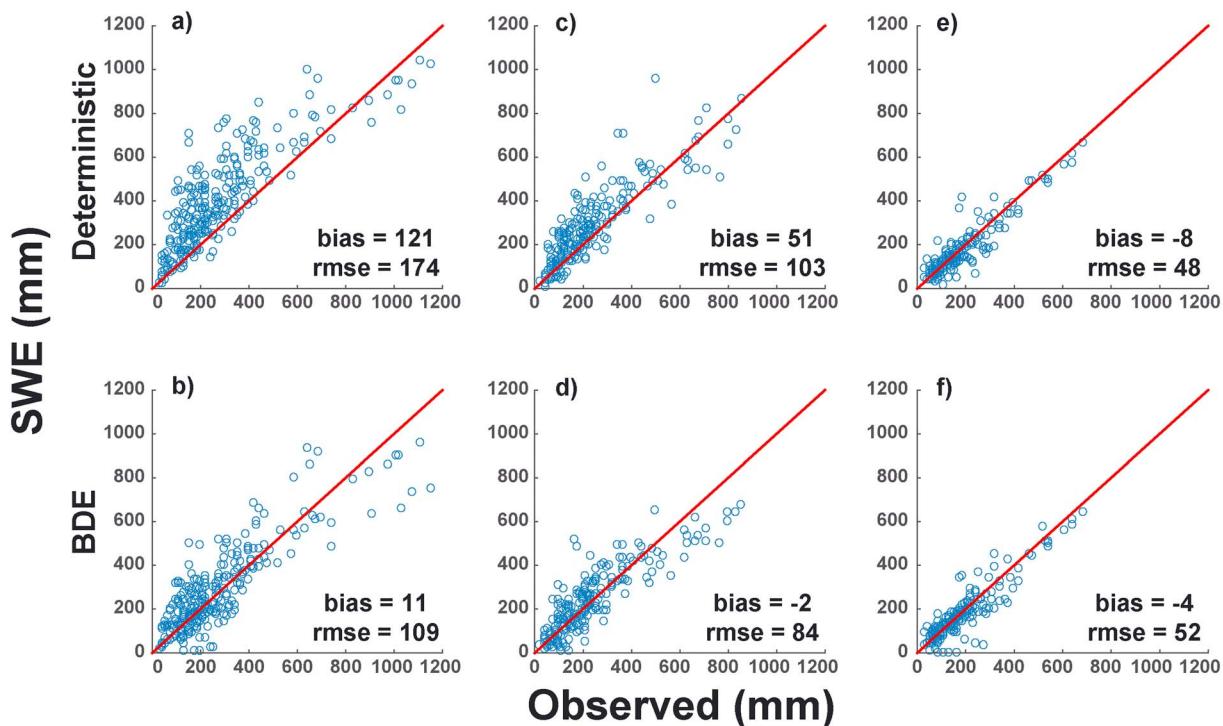


Figure 6. Scatterplots of daily observed and modeled snow-water-equivalent (SWE) for the deterministic and bias-detecting ensemble (BDE) model runs at the SWE stations in WY14, WY15, and WY17. rmse = root mean square error.

these HS values are based on data in Figure 5 and therefore differ slightly from the daily averaged biases in Table 2).

Snow-covered area (SCA) is another key element of spatial snowmelt models as melt volumes are scaled by SCA. Figures 9a and 9b present the spread of meltout day errors, considered as a proxy for spatial SCA estimations, for WY14 and WY15. Meltout days were the first days after the day of peak SWE (determined from the deterministic simulations) with zero observed/simulated HS. Consistent with the prior results, there were large positive errors (i.e., simulated meltout day later than observed) for the deterministic runs and errors much closer to zero for the BDE simulations. Though the range of errors in the two simulations was similar, rmse was reduced in the BDE simulations.

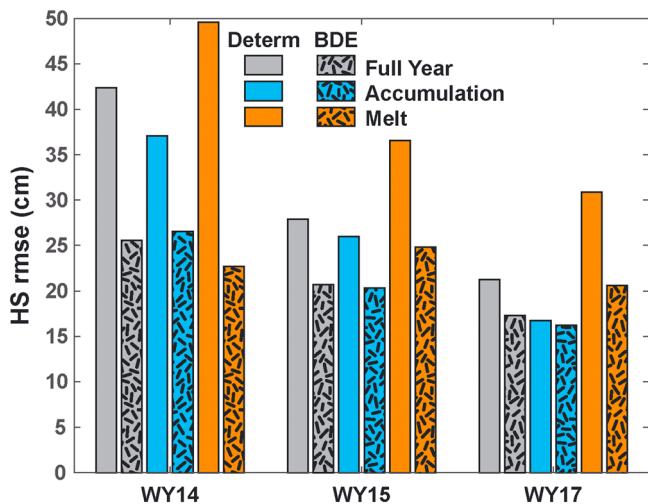


Figure 7. Root mean square errors (rmse) at the snow-water-equivalence (SWE) stations for simulated and observed HS by year and by season. Similar to all modeling, rmse only includes days when either the model or observation had snow present. Accumulation and melt phases were split at each SWE station by time of peak SWE in that station's deterministic simulation. BDE = bias-detecting ensemble.

4.2. WY2017

The WY17 model run began on 01 September 2016 and extended to 30 June 2017. The forcings for WY17 represent a best-case scenario for model simulations. The data consisted of hindcast products combining NWP reanalysis forcings with numerous station observations. Forcing data of this quality are generally not achievable and certainly unavailable for operational forecast scenarios. The quality of this data set is distinctly different from the more typically available WY14/15 data set and was applied here to investigate the limits of the BDE approach. The considerable air temperature biases above 1,400-m elevation present in the WY14 (-1.6°C) and WY15 (-1.5°C) forcing data (Figure 4), which contributed to substantial biases in the deterministic simulations, were, after assimilating observed temperatures, not present in the WY17 forcing data (0.1°C).

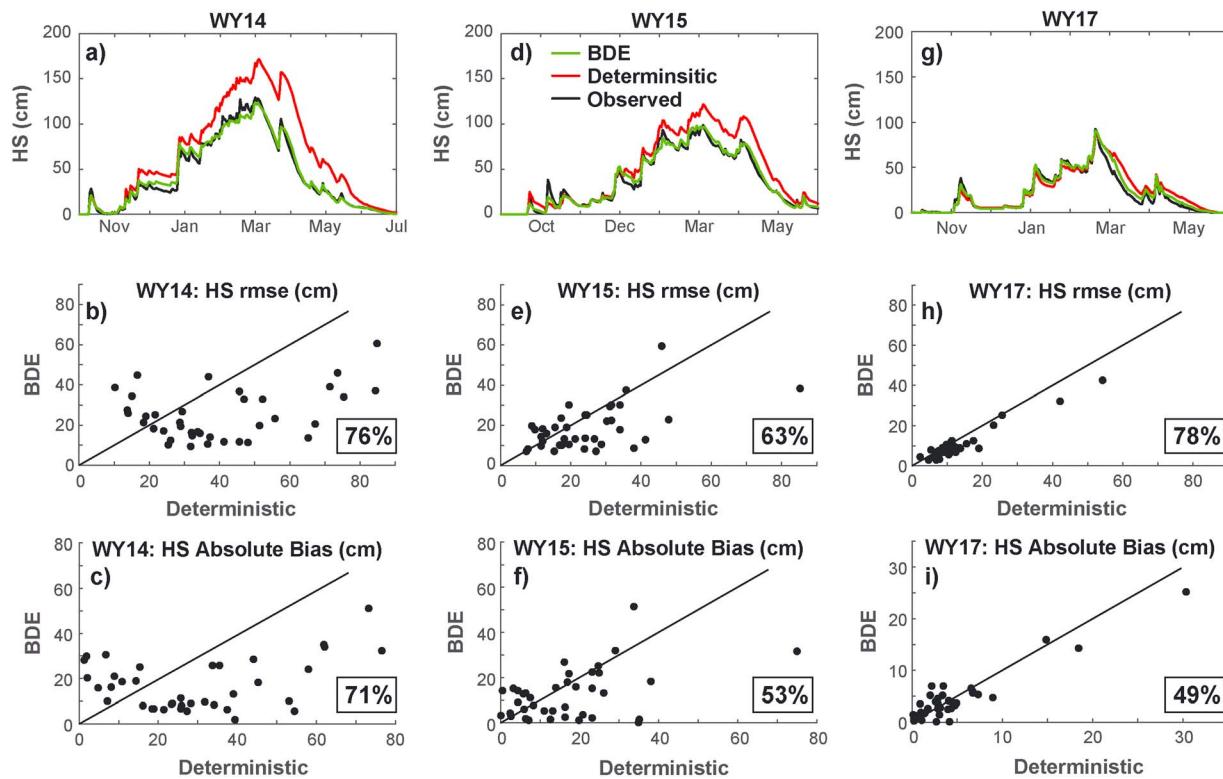


Figure 8. The first row of figures (a, d, g) depict HS for the deterministic and bias-detecting ensemble (BDE) simulations along with observations averaged across all the independent snow-water-equivalence (SWE) stations. The second and third rows plot error metrics for the two simulations with circles representing results from individual SWE stations. The straight lines are one-to-one plots. Circles plotting below the one-to-one line indicate stations where the BDE improved simulations. The percentages indicate the percentage of stations with BDE improvements. rmse = root mean square error.

The deterministic runs had very low HS and SWE biases (Figures 5e and 6e). Moreover, the HS and SWE rmse metrics for this run were lower than those in the BDE simulations in WY14 and WY15. There was a slight underestimation of HS during the accumulation season and a stronger overestimation during the melt phase (Figure 8g). The aggregated deterministic data showed a slight positive bias in HS yet a negative bias in SWE (Figures 5f, 6f). This contrast could be due to the fact that while 34% of the HS observations occurred during the melt phase when HS overestimation errors became noticeable, just 19% of the SWE measurements did. Errors in the density component of the snow model, which the HS-based BDE corrections presented here cannot account for, could also be affecting this disparity. Corresponding with the improved temperature forcings, there were only 2 days with positive temperature adjustments of precipitation phase and two with negative adjustments.

The aggregated bias metrics for HS and SWE were slightly better for the BDE run (Figures 5f and 6f vs. Figures 5e and 6e). Although rmse was reduced for HS, it slightly increased for SWE in the BDE simulation. The increase in SWE rmse could be a product of the aforementioned sampling disparities or density components of the snow model. However, it may also be indicative that the BDE method, given this best-case scenario of model forcings, is approaching one of the current limitations of this technique—the lack of a component for handling random errors. When random errors are of the same magnitude or greater than the detected systematic biases, any advantages provided by the current BDE approach become uncertain.

Seasonal assessments of HS rmse showed slight improvements for the BDE simulation during the accumulation phase (3%), with stronger improvements during the melt season (33%; Figure 7). The obvious reduction in model bias present in Figure 8g in conjunction with the improved melt-season HS metrics indicate that the degraded rmse performance for SWE in WY17 was due to either the limited SWE observations or the density components of the snow model and not due to the relative magnitudes of systematic and random errors.

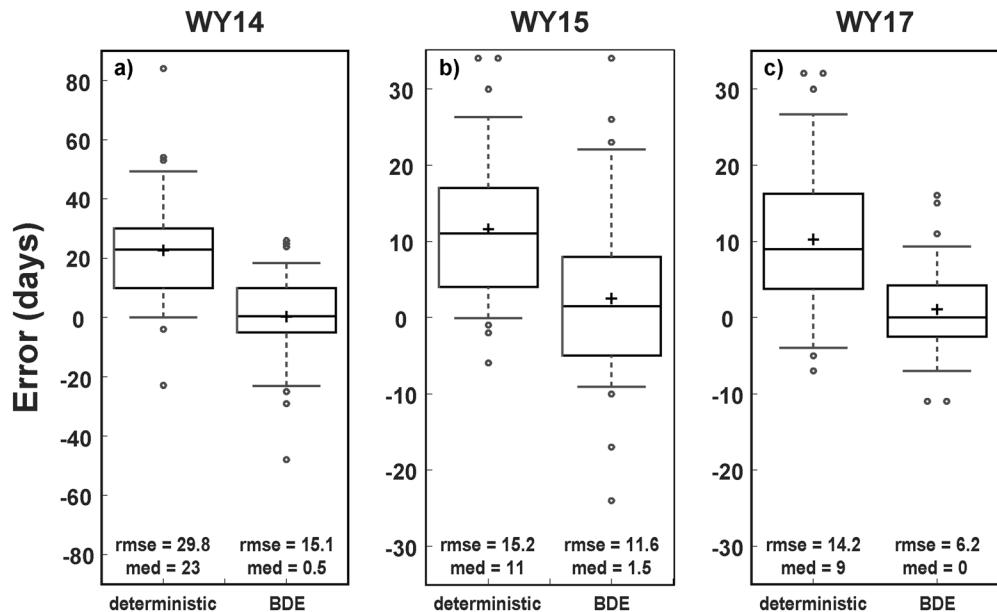


Figure 9. Box and whiskers plots for meltout day errors at the snow-water-equivalence (SWE) stations. Boxes represent the interquartile range, within box lines the median, crosses the mean, and whiskers extend to 9–91 percentiles. The root mean square errors (rmse) and error medians (med; days) are also listed. Median and mean errors were strongly positive in all years prior to the bias-detecting ensemble (BDE) application and near zero afterward.

The station-by-station comparisons showed that the BDE simulations decreased HS rmse at 78% of the stations (Figure 8h)—a similar percentage as in WY14. Reductions in absolute biases, however, only occurred at 49% of the stations (Figure 8i). One potential explanation for these differences again points to the fact that most of the BDE improvements occurred during the shorter melt phase (Figures 7 and 8g). Whereas both overall rmse (30.9 to 20.6 cm) and absolute bias (15.6 to 5.0 cm) were substantially reduced during the melt phase, during the longer accumulation season, rmse slightly decreased (16.7 to 16.2 cm), but mean absolute bias slightly increased (0.1 to 1.8 cm).

In this instance, where BDE improvements occurred predominantly in the melt phase, the obvious inference is that errors were present somewhere in the solar radiation routines. This could be due to underestimations of the model forcings or modeled albedos or a combination of both. However, as regards the solar radiation inputs, Bellaire et al. (2017) found that COSMO-1 actually overestimated solar radiation in their single observation point analysis. Recall that substantial errors during the melt phase were also present in WY14 and WY15. An example of potential snow model based errors would be errors in the albedo decay formulation. Such errors could be related to model parameterizations or to environmental effects such as synoptic dust events not accounted for in the modeling framework.

Table 2

Statistical Metrics Between Daily Averages of Observed and Modeled HS From all Snow-Water-Equivalence (SWE) Stations for the Deterministic and Bias-Detecting Ensemble (BDE) Simulations Shown in Figures 8a, 8d, and 8g

Metrics	WY14		WY15		WY17	
	Determ	BDE	Determ	BDE	Determ	BDE
rmse (cm)	29.0	4.4	15.7	4.1	8.0	4.1
Bias (cm)	22.1	0.7	10.8	0.8	2.8	1.6
N-S	0.71	0.99	0.84	0.98	0.87	0.97

Note. N-S is Nash-Sutcliffe model efficiency coefficient (Nash & Sutcliffe, 1970). rmse = root mean square error.

Errors in simulated meltout days at the SWE stations displayed consistent improvements after the BDE application (Figure 9c). As was the case in WY14 and WY15, centroid error metrics for WY17 were substantially reduced and very close to zero, while rmse was also reduced after application of the BDE. In contrast to the earlier years, the range of meltout day errors was also reduced.

Using the WY17 data, the temporal stability of the estimated biases was also tested by extending the application of the biases into future forecasts. In this test, both the deterministic and BDE simulations at the SWE stations were initialized each day with bulk snow depth matched to the start-of-day observations. The same hindcast COSMO-1 products used in the BDE evaluations were used as the “forecast” products in

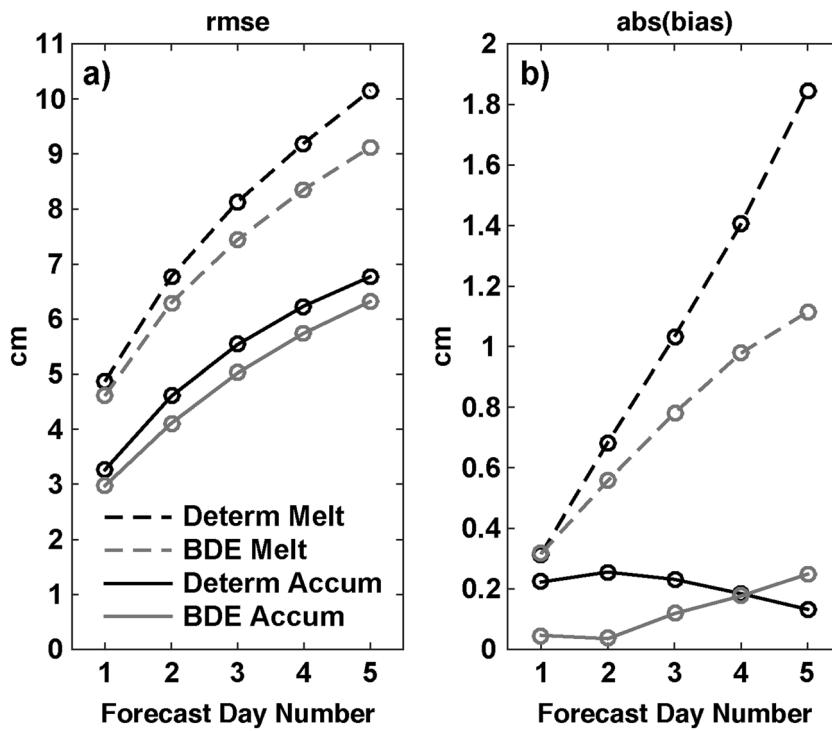


Figure 10. Root mean square error (rmse) and bias (error) metrics by forecast day for the deterministic and bias-detecting ensemble (BDE) HS simulations at the snow-water-equivalence (SWE) stations. The BDE methodology had increasing forecast skill through the 5-day test period during the melt phase. During the accumulation period, rmse improvements remained consistent through the 5 days with bias reductions persisting for up to 4 days.

these forecast simulations. Model runs extended for a 5-day period with BDE adjustments fixed at their start time values. HS error metrics were assessed at the end of each day. The 5-day forecasts were done for each day in the WY17 simulation up until 25 June (5 days prior to the end of the model run). The rmse for simulated HS using the BDE-corrected forcings were reduced in the accumulation and melt seasons and rmse improvements extended throughout the 5-day period tested (Figure 10a). While differences in the two model runs in terms of rmse remained consistent during the accumulation season, the BDE simulation offered increased forecasting skill with each subsequent forecast day during the melt season. A similarly behaving, but stronger, pattern was evident in the bias assessments (Figure 10b). Absolute bias or error reduction consistently improved with each advancing forecast day during the melt phase and was reduced for up to 4 days during the accumulation season.

5. Discussion

The introduced BDE methodology was shown to capably and efficiently improve snow model performance given two very different sets of forcing data. While one set consisted of modern, high-resolution NWP forecast data, the other was a best-case scenario comprised of hindcast reanalysis NWP data enhanced with air temperature and precipitation data from a dense observation network. The presented methodology aptly removed the prevalent systematic biases displayed in the deterministic simulations. In some cases though, particularly where systematic biases were not exhibited at stations, performance gains using the BDE were uncertain. In this section, the benefits and limitations of the presented approach are evaluated.

Issues of scale and representativeness will always be present when mixing spatial and point data. In this application, as is often the case, gridded NWP products and snow simulations were assessed using point observations. There were some pervasive trends in these simulations: cold temperature biases in WY14 and WY15 and a slower than observed melt in WY17. Given the number of stations exhibiting these

trends and consequences (Figures 4, 5a, 5c, 6a, 5c, 8a, 8d, and 8g), it can be confidently stated that these trends were prevalent in either the forcing data or model constructs. However, not all stations exhibited these trends. The question remains of how representative were these “outlying” stations of local trends. If they are indeed representative, then perhaps the basic bias-interpolation scheme (section 3.5) employed here should be improved upon to better represent these local effects. If they are not representative, then one can be satisfied with the acknowledged BDE errors at these stations. A further challenge, not investigated here, is how to upscale the point BDE assessments for spatial applications. Snow observations, such as the ones used in this study to assess biases, are often made in wind-sheltered and nonforested locations. Due to these characteristics, these sites are often not representative of other nearby terrain (e.g., Molotch & Bales, 2006). This is an important consideration when upscaling these point-based bias observations to spatial applications.

Since the premise of this application was based on differences between HS simulations and observations serving as a proxy for mass fluxes, an obvious dependency was the accuracy of modeled densities. Whereas a SWE-based approach, such as one using the North American SNOTEL SWE data, would be more straightforward, HS-based snow modeling applications are becoming more common due to the lower costs associated with station-based installations and the increasing availability of lidar products (e.g., Painter et al., 2016). The P bias evaluation in HS-based applications is primarily affected by the accuracy of simulated new snow densities and secondarily by the compaction routines. The evaluation of the EB bias is influenced solely by the latter and its temporally integrated effect on the accuracy of the simulated bulk density. In any BDE applications based on observed HS data, inaccuracies in modeled fluxes due to inaccurately modeled densities should not outweigh flux inaccuracies produced by model forcings and non-density-related model constructs.

This application benefitted from data taken from a dense HS observation network resolved at a daily time-step. When bias trends are widespread, as was seemingly the case in all 3 years presented here, similar BDE benefits should be expected even with a lower density observation network. However, the relative power of detecting these trends, the ability to differentiate local and regional trends, and questions of representativeness become more prominent as the density of observations decreases. The daily resolved data provided a straightforward means of independently assessing precipitation and energy influences on the snowpack. Independent assessments become more complicated when data acquisition times span longer periods (e.g., ASO lidar data (Painter et al., 2016)). Nonpublished work using the ASO data though has shown that objective statistical measures can provide a means of evaluating BDE members in these situations and improving simulations as well.

Compared to typical state-based DA methods, the forcing-based BDE method conserves mass and maintains consistency between internal snow states. Mass conservation is certainly an important consideration when snow simulations are linked to hydrologic models or mass-balance assessments. State-consistency refers to the fact that internal model states (e.g., layering, temperature, density, and grain size) remain consistent with each other and the simulation environment. Maintaining state-consistency in complex snow models—with the exception of certain cases with application-specific objectives independent of internal structure—is vital (Magnusson et al., 2017). By contrast, state-based assimilation methods oftentimes are not adept at mass conservation (e.g., unable to ascertain if overestimated SWE was a consequence of overestimated snowfall or underestimated melt) and either ignore or apply a priori rules to handle state inconsistencies (e.g., observed melt from a frozen snowpack). Certainly, there were numerous a priori decisions made also in this application (e.g., look-back periods and daily bias criteria). However, none of these decisions violated or challenged accepted snowpack physics. While it was outside the breadth of this study to assess model sensitivities to these decisions, it is our belief that the current BDE methodology is based on sound principles, and further sensitivity analysis will only contribute to additionally enhancing the technique.

It was also shown that the BDE method of assessing biases can improve forecasting capabilities. In the WY17 forecasting test performed here, strong temperature biases were not present in the forcings. The forecasting analysis was limited to WY17 since WY14 and WY15 had substantial air temperature biases that were not comprehensively investigated in the current work. In order to optimize forecasting capabilities, temperature biases should be thoroughly and, if possible, directly assessed. Methods such as those

introduced by Yussouf and Stensrud (2007) could aptly fill this need. Temperature corrections would be compatible with the current BDE methodology as long as they are performed prior to the *P* and *EB* evaluations.

Two limitations of the current BDE approach introduced here are an inability to discriminate process-based energy balance biases and the lack of a random error component. Whereas longwave and shortwave energy fluxes have seasonal tendencies in terms of prominence that the back-averaging technique might reasonably account for, expectations are that more stochastic processes that exhibit strong day-to-day or sub-daily fluctuations such as turbulent energy exchanges would not be handled as well. Future work to relate the *EB* term to either observed or simulated meteorology will be investigated. Coupling the current BDE approach with a random error component will likewise be tested. Yet another current shortcoming that plagues any model updating method based on snow data is the scarcity and quality of forest snow observations. Accurate below-canopy snow observations are required to ascertain modeling errors related to forest-snow processes.

6. Conclusions

An innovative algorithm for updating model simulations with observations appropriate for complex snow models, termed the BDE approach, has been developed. To the best of our knowledge, this is the first algorithm capable of incorporating point-scale observations in a fully distributed, multilayer, physics-based snow model while maintaining physically consistent states and conserving mass. The method is based on a predetermined set of perturbations input to the snow model. This set of prescribed model perturbations is referred to as the BDE. Evaluations of the performance of the ensemble members against observed snow states, in this case snow depth, allow an assessment of biases in the modeling chain. Over the course of three winters that encapsulated a wide range in the quality of forcing data, the capabilities of this technique were demonstrated. When the forcing data contained a persistent bias, it was also clearly evident in the deterministic model outcomes. The BDE method aptly detected biases and consequent model runs were, on average, largely bias-free. Other performance metrics were also improved. In the case where the forcing data were of the highest quality possible, deterministic model simulations were improved compared to the former case, yet a cold bias was evident during the melt phase. Though the current BDE method could not discriminate the reason for the melt phase bias, the method did aptly correct for it.

Updating model states using techniques such as the Kalman filter does not necessarily conserve mass, and their integration with complex, multilayer snow models is not straightforward (e.g., Andreadis & Lettenmaier, 2006; Durand et al., 2009). Forcing-based DA techniques such as is possible with PFs have been shown to be compatible with complex snow models (e.g., Charrois et al., 2016; Magnusson et al., 2017). However, the capabilities of these latter techniques for improving simulations at sites without observations, as would typically be required in distributed snow modeling, have yet to be demonstrated. The BDE method, however, conserves mass, maintains physically consistent snow states in a complex snow model, and in this study, was shown to improve model simulations at unobserved sites as well as in a forecast situation. Whereas the aforementioned, traditional modeling techniques intrinsically account for random errors but often encounter difficulties when large systematic errors are present (Dee, 2005; Magnusson et al., 2014), the BDE technique as currently structured does not include a random error component and directly assesses systematic biases. Future work will look to merge these techniques to further improve snow modeling efforts.

Acknowledgments

COSMO data and much of the SWE data were provided by the Federal Office of Meteorology and Climatology (MeteoSwiss). This study was partly funded by the Swiss Federal Office for the Environment (FOEN). Use of the COSMO model used to create the weather forcings in this study is available free of charge for research applications through COSMO. The station meteorological and snow data used in this research are available at the Envidat data portal (www.envidat.ch).

References

- Anderson, E. A. (1973). National Weather Service River Forecast System—snow accumulation and ablation model, NOAA Technical Memorandum: NWS Hydro-17, US National Weather Service.
- Anderson, E. A. (1976). A point energy and mass balance model of a snow cover. *Rep NOAA Technical Report NWS 19*, 150 pp, National Oceanic and Atmospheric Administration: Washington, DC.
- Andreadis, K. M., & Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources*, 29(6), 872–886. <https://doi.org/10.1016/j.advwatres.2005.08.004>
- Auligné, T., McNally, A. P., & Dee, D. P. (2007). Adaptive bias correction for satellite data in a numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society*, 133(624), 631–642. <https://doi.org/10.1002/qj.56>
- Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R., & Dozier, J. (2006). Mountain hydrology of the western United States. *Water Resources Research*, 42, W08432. <https://doi.org/10.1029/2005WR004387>
- Bartelt, P., & Lehning, M. (2002). A physical SNOWPACK model for the Swiss avalanche warning: Part I: Numerical model. *Cold Regions Science and Technology*, 35(3), 123–145. [https://doi.org/10.1016/S0165-232X\(02\)00074-5](https://doi.org/10.1016/S0165-232X(02)00074-5)

- Bellaire, S., van Herwijnen, A., Mitterer, C., & Schweizer, J. (2017). On forecasting wet-snow avalanche activity using simulated snow cover data. *Cold Regions Science and Technology*, 144, 28–38. <https://doi.org/10.1016/j.coldregions.2017.09.013>
- Blöschl, G. (1999). Scaling issues in snow hydrology. *Hydrological Processes*, 13(14–15), 2149–2175. [https://doi.org/10.1002/\(SICI\)1099-1085\(199910\)13:14/15<2149::AID-HYP847>3.0.CO;2-8](https://doi.org/10.1002/(SICI)1099-1085(199910)13:14/15<2149::AID-HYP847>3.0.CO;2-8)
- Boone, A., & Etchevers, P. (2001). An intercomparison of three snow schemes of varying complexity coupled to the same land surface model: Local-scale evaluation at an alpine site. *Journal of Hydrometeorology*, 2(4), 374–394. [https://doi.org/10.1175/1525-7541\(2001\)002<0374:AIOTSS>2.0.CO;2](https://doi.org/10.1175/1525-7541(2001)002<0374:AIOTSS>2.0.CO;2)
- Brasnett, B. (1999). A global analysis of snow depth for numerical weather prediction. *Journal of Applied Meteorology*, 38(6), 726–740. [https://doi.org/10.1175/1520-0450\(1999\)038<0726:AGAOSD>2.0.CO;2](https://doi.org/10.1175/1520-0450(1999)038<0726:AGAOSD>2.0.CO;2)
- Brun, E., David, P., Sudul, M., & Brunot, G. (1992). A numerical model to simulate snow-cover stratigraphy for operational avalanche forecasting. *Journal of Glaciology*, 38(128), 13–22. <https://doi.org/10.1017/S002214300009552>
- Charrois, L., Cosme, E., Dumont, M., Lafaysse, M., Morin, S., Libois, Q., & Picard, G. (2016). On the assimilation of optical reflectances and snow depth observations into a detailed snowpack model. *The Cryosphere*, 10(3), 1021–1038. <https://doi.org/10.5194/tc-10-1021-2016>
- Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., et al. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, 31(10), 1309–1324. <https://doi.org/10.1016/j.advwatres.2008.06.005>
- Dechant, C., & Moradkhani, H. (2011). Radiance data assimilation for operational snow and streamflow forecasting. *Advances in Water Resources*, 34(3), 351–364. <https://doi.org/10.1016/j.advwatres.2010.12.009>
- Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 131(613), 3323–3343. <https://doi.org/10.1256/qj.05.137>
- Douville, H., Royer, J.-F., & Mahfouf, J.-F. (1995). A new snow parameterization for the Météo-France climate model. *Climate Dynamics*, 12(1), 21–35.
- Dozier, J. (2011). Mountain hydrology, snow color, and the fourth paradigm. *Eos, Transactions American Geophysical Union*, 92(43), 373–374.
- Durand, M., Kim, E. J., & Margulis, S. A. (2009). Radiance assimilation shows promise for snowpack characterization. *Geophysical Research Letters*, 36, L02503. <https://doi.org/10.1029/2008GL035214>
- Essery, R., Morin, S., Lejeune, Y., & Ménard, C. B. (2013). A comparison of 1701 snow models using observations from an alpine site. *Advances in Water Resources*, 55, 131–148. <https://doi.org/10.1016/j.advwatres.2012.07.013>
- Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4), 343–367. <https://doi.org/10.1007/s10236-003-0036-9>
- Fan, Y. R., Huang, G. H., Baetz, B. W., Li, Y. P., Huang, K., Chen, X., & Gao, M. (2017). Development of integrated approaches for hydrological data assimilation through combination of ensemble Kalman filter and particle filter methods. *Journal of Hydrology*, 550, 412–426. <https://doi.org/10.1016/j.jhydrol.2017.05.010>
- Fletcher, S. J., Liston, G. E., Hiemstra, C. A., & Miller, S. D. (2012). Assimilating MODIS and AMSR-E snow observations in a snow evolution model. *Journal of Hydrometeorology*, 13(5), 1475–1492. <https://doi.org/10.1175/JHM-D-11-082.1>
- Franz, K. J., Hogue, T. S., & Sorooshian, S. (2008). Operational snow modeling: Addressing the challenges of an energy balance model for National Weather Service forecasts. *Journal of Hydrology*, 360(1–4), 48–66. <https://doi.org/10.1016/j.jhydrol.2008.07.013>
- Frei, C., & Schär, C. (1998). A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology*, 18(8), 873–900. [https://doi.org/10.1002/\(SICI\)1097-0088\(19980630\)18:8<873::AID-JOC255>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9)
- Freudiger, D., Kohn, I., Stahl, K., & Weiler, M. (2014). Large-scale analysis of changing frequencies of rain-on-snow events with flood-generation potential. *Hydrology and Earth System Sciences*, 18(7), 2695–2709. <https://doi.org/10.5194/hess-18-2695-2014>
- Harris, B. A., & Kelly, G. (2001). A satellite radiance-bias correction scheme for data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 127(574), 1453–1468. <https://doi.org/10.1002/qj.49712757418>
- Havens, S., Marks, D., Kormos, P., & Hedrick, A. (2017). Spatial modeling for resources framework (SMRF): A modular framework for developing spatial forcing data for snow modeling in mountain basins. *Computers & Geosciences*, 109, 295–304. <https://doi.org/10.1016/j.cageo.2017.08.016>
- Hedrick, A. R., Marks, D., Havens, S., Robertson, M., Johnson, M., Sandusky, M., et al. (2018). Direct insertion of NASA airborne snow observatory-derived snow depth time series into the iSnobal energy balance snow model. *Water Resources Research*, 10, 8045–8063.
- Helbig, N., van Herwijnen, A., Magnusson, J., & Jonas, T. (2015). Fractional snow-covered area parameterization over complex topography. *Hydrology and Earth System Sciences*, 19(3), 1339–1351. <https://doi.org/10.5194/hess-19-1339-2015>
- Jonas, T., Marty, C., & Magnusson, J. (2009). Estimating the snow water equivalent from snow depth measurements in the Swiss Alps. *Journal of Hydrology*, 378(1–2), 161–167. <https://doi.org/10.1016/j.jhydrol.2009.09.021>
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 35–45. <https://doi.org/10.1115/1.3662552>
- Kumar, M., Marks, D., Dozier, J., Reba, M. L., & Winstral, A. (2013). Evaluation of distributed hydrologic impacts of temperature-index and energy-based snow models. *Advances in Water Resources*, 56, 77–89. <https://doi.org/10.1016/j.advwatres.2013.03.006>
- Leisenring, M., & Moradkhani, H. (2011). Snow water equivalent prediction using Bayesian data assimilation methods. *Stoch Environ Res Risk Assess*, 25(2), 253–270. <https://doi.org/10.1007/s00477-010-0445-5>
- Liston, G. E., & Hiemstra, C. A. (2008). A simple data assimilation system for complex snow distributions (SnowAssim). *Journal of Hydrometeorology*, 9(5), 989–1004. <https://doi.org/10.1175/2008JHM871.1>
- Liston, G. E., Pielke, R. A., & Greene, E. M. (1999). Improving first-order snow-related deficiencies in a regional climate model. *Journal of Geophysical Research*, 104, 19,559–19,567. <https://doi.org/10.1029/1999JD900055>
- Louis, J.-F. (1979). A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology*, 17(2), 187–202. <https://doi.org/10.1007/BF00117978>
- Magnusson, J., Gustafsson, D., Hüslar, F., & Jonas, T. (2014). Assimilation of point SWE data into a distributed snow cover model comparing two contrasting methods. *Water Resources Research*, 50, 7816–7835. <https://doi.org/10.1002/2014WR015302>
- Magnusson, J., Winstral, A., Stordal, A. S., Essery, R., & Jonas, T. (2017). Improving physically based snow simulations by assimilating snow depths using the particle filter. *Water Resources Research*, 53, 1125–1143. <https://doi.org/10.1002/2016WR019092>
- Martinec, J., & Rango, A. (1991). Indirect evaluation of snow reserves in mountain basins. In H. Bergmann, H. Lang, W. Frey, D. Issler, & B. Salm (Eds.), *Proceedings: Snow, hydrology and forest in high alpine areas*, (pp. 111–120). Vienna: IAHS Publication.
- Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P., & Stouffer, R. J. (2008). Stationarity is dead: Whither water management? *Science*, 319(5863), 573–574. <https://doi.org/10.1126/science.1151915>

- Molotch, N. P., & Bales, R. C. (2006). SNOTEL representativeness in the Rio Grande headwaters on the basis of physiographics and remotely sensed snow cover persistence. *Hydrological Processes*, 20(4), 723–739. <https://doi.org/10.1002/hyp.6128>
- Moradkhani, H., DeChant, C. M., & Sorooshian, S. (2012). Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. *Water Resources Research*, 48, W12520. <https://doi.org/10.1029/2012WR012144>
- Musselman, K. N., Lehner, F., Ikeda, K., Clark, M. P., Prein, A. F., Liu, C., et al. (2018). Projected increases and shifts in rain-on-snow flood risk over western North America. *Nature Climate Change*, 8(9), 808–812. <https://doi.org/10.1038/s41558-018-0236-4>
- Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, 10(3), 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)
- Nayak, A., Marks, D. G., Chandler, D., & Seyfried, M. S. (2010). Long-term snow, climate and streamflow trends from the Reynolds Creek Experimental Watershed, Owyhee Mountains, Idaho, United States. *Water Resources Research*, 46, W06519. <https://doi.org/10.1029/2008WR007525>
- Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., et al. (2016). The airborne snow observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sensing of Environment*, 184, 139–152. <https://doi.org/10.1016/j.rse.2016.06.018>
- Rasmussen, J., Madsen, H., Jensen, K. H., & Refsgaard, J. C. (2016). Data assimilation in integrated hydrological modelling in the presence of observation bias. *Hydrology and Earth System Sciences*, 20(5), 2103–2118. <https://doi.org/10.5194/hess-20-2103-2016>
- Reichle, R. H., & Koster, R. D. (2003). Assessing the impact of horizontal error correlations in background fields on soil moisture estimation. *Journal of Hydrometeorology*, 4(6), 1229–1242. [https://doi.org/10.1175/1525-7541\(2003\)004<1229:ATIOHE>2.0.CO;2](https://doi.org/10.1175/1525-7541(2003)004<1229:ATIOHE>2.0.CO;2)
- Reichle, R. H., McLaughlin, D. B., & Entekhabi, D. (2002). Hydrologic data assimilation with the ensemble Kalman filter. *Monthly Weather Review*, 130(1), 103–114. [https://doi.org/10.1175/1520-0493\(2002\)130<0103:HDAWWE>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<0103:HDAWWE>2.0.CO;2)
- Rössler, O., Froidevaux, P., Börst, U., Rickli, R., Martius, O., & Weingartner, R. (2014). Retrospective analysis of a nonforecasted rain-on-snow flood in the Alps—A matter of model limitations or unpredictable nature? *Hydrology and Earth System Sciences*, 18(6), 2265–2285. <https://doi.org/10.5194/hess-18-2265-2014>
- Rutter, N., Essery, R., Pomeroy, J., Altimir, N., Andreadis, K., Baker, I., et al. (2009). Evaluation of forest snow processes models (SnowMIP2). *Journal of Geophysical Research*, 114, D06111. <https://doi.org/10.1029/2008JD011063>
- Schirmer, M., & Jamieson, B. (2015). Verification of analysed and forecasted winter precipitation in complex terrain. *The Cryosphere*, 9(2), 587–601. <https://doi.org/10.5194/tc-9-587-2015>
- Shrestha, D. L., Robertson, D. E., Wang, Q. J., Pagano, T. C., & Hapuarachchi, H. A. P. (2013). Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. *Hydrology and Earth System Sciences*, 17(5), 1913–1931. <https://doi.org/10.5194/hess-17-1913-2013>
- Skiles, S. M., Painter, T. H., Deems, J. S., Bryant, A. C., & Landry, C. C. (2012). Dust radiative forcing in snow of the Upper Colorado River Basin: 2. Interannual variability in radiative forcing and snowmelt rates. *Water Resources Research*, 48, W07522. <https://doi.org/10.1029/2012WR011986>
- Slater, A. G., & Clark, M. P. (2006). Snow data assimilation via an ensemble Kalman filter. *Journal of Hydrometeorology*, 7(3), 478–493. <https://doi.org/10.1175/JHM505.1>
- Stensrud, D. J., & Yussouf, N. (2005). Bias-corrected short-range ensemble forecasts of near surface variables. *Meteorological Applications*, 12(03), 217–230. <https://doi.org/10.1017/S135048270500174X>
- Strachan, S., Kelsey, E. P., Brown, R. F., Dascalu, S., Harris, F., Kent, G., et al. (2016). Filling the data gaps in mountain climate observatories through advanced technology, refined instrument siting, and a focus on gradients. *Mountain Research and Development*, 36(4), 518–527.
- Sun, C., Walker, J. P., & Houser, P. R. (2004). A methodology for snow data assimilation in a land surface model. *Journal of Geophysical Research*, 109, D08108. <https://doi.org/10.1029/2003JD003765>
- Surfleet, C. G., & Tullos, D. (2013). Variability in effect of climate change on rain-on-snow peak flow events in a temperate climate. *Journal of Hydrology*, 479, 24–34. <https://doi.org/10.1016/j.jhydrol.2012.11.021>
- Sweeney, C. P., Lynch, P., & Nolan, P. (2013). Reducing errors of wind speed forecasts by an optimal combination of post-processing methods. *Meteorological Applications*, 20(1), 32–40. <https://doi.org/10.1002/met.294>
- Trubilowicz, J. W., & Moore, R. D. (2017). Quantifying the role of the snowpack in generating water available for run-off during rain-on-snow events from snow pillow records. *Hydrological Processes*, 31(23), 4136–4150. <https://doi.org/10.1002/hyp.11310>
- Verseghy, D. L. (1991). Class—A Canadian land surface scheme for GCMS. I. Soil model. *International Journal of Climatology*, 11(2), 111–133.
- Vionnet, V., Brun, E., Morin, S., Boone, A., Faroux, S., Le Moigne, P., et al. (2012). The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2. *Geoscientific Model Development*, 5(3), 19.
- Vögeli, C., Lehning, M., Wever, N., & Bavay, M. (2016). Scaling precipitation input to spatially distributed hydrological models by measured snow distribution. *Frontiers in Earth Science*, 4(108). <https://doi.org/10.3389/feart.2016.00108>
- Wayand, N. E., Lundquist, J. D., & Clark, M. P. (2015). Modeling the influence of hypsometry, vegetation, and storm energy on snowmelt contributions to basins during rain-on-snow floods. *Water Resources Research*, 51, 8551–8569. <https://doi.org/10.1002/2014WR016576>
- Winstral, A., Marks, D., & Gurney, R. (2009). An efficient method for distributing wind speeds over heterogeneous terrain. *Hydrological Processes*, 23(17), 2526–2535. <https://doi.org/10.1002/hyp.7141>
- Winstral, A., Marks, D., & Gurney, R. (2013). Simulating wind-affected snow accumulations at catchment to basin scales. *Advances in Water Resources*, 55, 64–79. <https://doi.org/10.1016/j.advwatres.2012.08.011>
- Winstral, A., Marks, D., & Gurney, R. (2014). Assessing the sensitivities of a distributed snow model to forcing data resolution. *Journal of Hydrometeorology*, 15(4), 1366–1383. <https://doi.org/10.1175/JHM-D-13-0169.1>
- Würzer, S., Jonas, T., Wever, N., & Lehning, M. (2016). Influence of initial snowpack properties on runoff formation during rain-on-snow events. *Journal of Hydrometeorology*, 17(6), 1801–1815. <https://doi.org/10.1175/JHM-D-15-0181.1>
- Xu, J., & Shu, H. (2014). Assimilating MODIS-based albedo and snow cover fraction into the Common Land Model to improve snow depth simulation with direct insertion and deterministic ensemble Kalman filter methods. *Journal of Geophysical Research: Atmospheres*, 119, 10,684–10,701. <https://doi.org/10.1002/2014JD022012>
- Yussouf, N., & Stensrud, D. J. (2007). Bias-corrected short-range ensemble forecasts of near-surface variables during the 2005/06 cool season. *Weather and Forecasting*, 22(6), 1274–1286. <https://doi.org/10.1175/2007WAF2007002.1>