AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE
10.1002/2014WR016498

Evaluating snow models with varying process representations for hydrological applications

Jan Magnusson[1], Nander Wever[1], Richard Essery[2], Nora Helbig[1], Adam Winstral[1], and Tobias Jonas[1]

[1]WSL Institute for Snow and Avalanche Research SLF, Davos, Switzerland,
[2]School of GeoSciences, University of Edinburgh, Edinburgh, UK

Key Points:
- Demonstration of informative method for evaluating snow models
- All model types can provide runoff predictions with similar performance
- Multimodel framework is a useful tool for selecting appropriate model structures

Correspondence to:
J. Magnusson,
magnusson@slf.ch

Citation:
Magnusson, J., N. Wever, R. Essery, N. Helbig, A. Winstral, and T. Jonas (2015), Evaluating snow models with varying process representations for hydrological applications, *Water Resour. Res.*, *51*, 2707–2723, doi:10.1002/2014WR016498.

Received 6 OCT 2014
Accepted 2 MAR 2015
Accepted article online 10 MAR 2015
Published online 26 APR 2015

Abstract Much effort has been invested in developing snow models over several decades, resulting in a wide variety of empirical and physically based snow models. For the most part, these models are built on similar principles. The greatest differences are found in how each model parameterizes individual processes (e.g., surface albedo and snow compaction). Parameterization choices naturally span a wide range of complexities. In this study, we evaluate the performance of different snow model parameterizations for hydrological applications using an existing multimodel energy-balance framework and data from two well-instrumented alpine sites with seasonal snow cover. We also include two temperature-index snow models and an intensive, physically based multilayer snow model in our analyses. Our results show that snow mass observations provide useful information for evaluating the ability of a model to predict snowpack runoff, whereas snow depth data alone are not. For snow mass and runoff, the energy-balance models appear transferable between our two study sites, a behavior which is not observed for snow surface temperature predictions due to site-specificity of turbulent heat transfer formulations. Errors in the input and validation data, rather than model formulation, seem to be the greatest factor affecting model performance. The three model types provide similar ability to reproduce daily observed snowpack runoff when appropriate model structures are chosen. Model complexity was not a determinant for predicting daily snowpack mass and runoff reliably. Our study shows the usefulness of the multimodel framework for identifying appropriate models under given constraints such as data availability, properties of interest and computational cost.

1. Introduction

Many snow models have been developed with varying degrees of complexity. The choice of a snow model should depend on the intended application. From the perspective of hydrological forecasting, model requirements differ from those of, for example, avalanche warning and climate modeling. To make well-informed model choices, we need to evaluate models using observational data and model skill measures relevant for the specific purpose. In the present literature, many studies show developments and improvements of snow models including validation against observational data [e.g., *Dutra et al.*, 2010; *Shrestha et al.*, 2010; *Tobin et al.*, 2013; *Vionnet et al.*, 2012]. The purpose of such studies can be to evaluate newly developed models [e.g., *De Michele et al.*, 2013; *Tuteja and Cunnane*, 1999] or introduce more accurate process representations in existing models [e.g., *Wever et al.*, 2014]. However, studies such as those mentioned above often lack comparisons against the range of already existing models making it difficult to judge whether the proposed improved model is best suited for a specific purpose. Additionally, for practical applications, including too many processes or very complex parameterizations may be counterproductive since often the number of undefined parameters increases, which can lead to overfitting and poor predictive capabilities of the model [*Cox et al.*, 2006]. Additionally, the computational time might be increased without much gain in model performance.

Studies evaluating single models often lack an appropriate benchmark and judging acceptable model performance from skill indices is often subjective [*Cox et al.*, 2006; *Seibert*, 2001]. Comparing several snow models against observations reveals their performance relative to each other. Thus, model intercomparison projects naturally include a sort of benchmark for performance if the necessary range of models is included. Some studies compare a large number of snow models with the intention of relating model performance and behavior to differences in model structure and setup [e.g., *Etchevers et al.*, 2004; *Rutter et al.*, 2009; *Slater*

et al., 2001]. Such studies show that differences in model efficiency depend on how the models represent, for example, snow albedo, fractional snow cover, and turbulent surface heat exchanges (see *Essery et al*. [2013] and *Rutter et al*. [2009] for detailed summaries of earlier snow model intercomparison projects). However, such comparisons remain difficult to interpret since the models often differ greatly between each other complicating the analysis of individual processes influencing the model performance. Most likely, a more informative approach than comparing different models is to deploy a multimodel framework that uses different representations for individual processes in all possible combinations [e.g., *Clark et al*., 2011; *Essery et al*., 2013]. A multimodel framework allows us to evaluate how different parameterizations for one process influence the simulation performance under an otherwise equal model setup. For example, we can assess how different model parameterizations describing snow albedo influence the ability of a model to reproduce snowpack runoff. However, neither the snow model intercomparison projects nor the multimodel framework studies have yet focused on finding snow models appropriate for hydrological forecasting specifically.

In another branch of studies, the skill of models predicting stream flow has been assessed by including different snow routines [*Franz et al*., 2008; *Kumar et al*., 2013; *Zeinivand and De Smedt*, 2009]. Those studies show varying results, for example, that increasing snow model complexity improves the runoff simulations [*Warscher et al*., 2013] whereas other studies do not find such a relationship [*Lehning et al*., 2006; *Zappa et al*., 2003]. From the perspective of hydrological forecasting, evaluating runoff models appear more relevant than validating the snow routines alone. However, validating the combination of a runoff routing model and snow routine may mask deficiencies in the snow simulations due to compensating mechanisms. Thus, isolated evaluations of snow models provide additional information to studies considering complete hydrological models.

In this study, we evaluate the behavior and performance of point snow models, often referred to as one-dimensional models. In particular, we focus on assessing their usefulness for hydrological applications. For this a purpose, we consider predictions of snow mass and snowpack runoff to be the most important variables. In our evaluation, we include a large range of existing snow models, spanning from simple empirical models [e.g., *Rango and Martinec*, 1995] to intermediate complexity energy-balance models [e.g., *Essery et al*., 2013] to the latest snow-physics models [e.g., *Wever et al*., 2014]. For our analysis, we use high quality input and validation data to examine the highest performance we may expect from the models.

2. Study Site and Data

We use published data sets from two field sites in the European Alps - Weissfluhjoch and Col de Porte - where all necessary meteorological and validation variables have been measured over a long period. The site Weissfluhjoch is situated at high altitude (2540 m) in Switzerland (46.82°N, 9.81°E) and Col de Porte is located at midelevation (1325 m) in France (45.30°N, 5.77°E). The data record for Weissfluhjoch used here spans from 9 October 1997 to 1 July 2010 and for Col De Porte from 18 December 1994 to 4 June 2011. For those periods, both study sites show similar annual precipitation sums of approximately 1800 mm/yr. During the snow-covered period, the average air temperature is lower at Weissfluhjoch (−3.0°C) than Col De Porte (0.2°C). The relative humidity is also lower (69.3%) at the higher elevation site Weissfluhjoch than at the midelevation site Col de Porte (82.3%). On the other hand, wind speeds are greater at Weissfluhjoch (2.4 m/s) than at Col de Porte (1.2 m/s). At Weissfluhjoch, the maximum snow depth per year varies between approximately 180 and 360 cm over the study period and at Col de Porte between approximately 60 and 200 cm. The snow cover lasts longer at Weissfluhjoch (roughly October to July) than at Col de Porte (roughly December to April). The snow cover at Weissfluhjoch typically shows a distinct accumulation period before melting starts, whereas midwinter melt events are common at Col de Porte.

At both sites, the following meteorological variables required for input to the snow models were measured: air temperature, relative humidity, wind speed, precipitation using a heated gauge, incoming longwave and shortwave radiations. In this study, we use the identical model input data as *Wever et al*. [2014]. Details about the technique for partitioning precipitation into rain and snowfall and the method for undercatch correction are described in *Wever et al*. [2014] for Weissfluhjoch and *Morin et al*. [2012] for Col de Porte. The following validation variables have been observed at both sites: snow lysimeter runoff, snow mass (i.e., snow water equivalent), snow depth, and snow surface temperature. In some cases, the same variable has

been measured by different methods at the same location. Both sites are equipped with snow lysimeters having a 5 m^2 surface area of the collector, and at Col de Porte, there is also a 1 m^2 lysimeter. For the model evaluation, we mainly use data from the larger lysimeters since those typically provide more reliable recordings than lysimeters with small collectors [*Kattelmann*, 2000]. Data from the smaller collector provided a gauge of runoff variability at Col de Porte. The lysimeter at Weissfluhjoch was malfunctioning during the winters 1999/2000 and 2004/2005 [*Wever et al*., 2014] which were omitted from our analysis. Snow mass was measured manually at both Col de Porte (weekly) and Weissfluhjoch (biweekly). For both sites, we use the automatic measurements of snow depth. At Weissfluhjoch, snow surface temperature was obtained from an infrared sensor and for Col de Porte, this variable was computed from observed outgoing longwave radiation. We only analyzed the goodness of fit between the simulated and observed runoff, and snow surface temperature during periods when any of the models and the observed snow depth exceeded 5 cm. Thus, the snow-free summer months do not influence the goodness-of-fit measures presented below. See *Morin et al*. [2012], *Schmucki et al*. [2014], and *Wever et al*. [2014] for more details about all observations.

Note that the lysimeters used in this study do not include a soil column. Thus, they measure the runoff from the snowpack directly without time delay. In the following, we mostly denote the observed snowpack runoff simply as runoff for convenience.

3. Methods

3.1. Snow Models

We compare the performance and behavior of three different model types that differ in complexity, computation time, and data requirements. In this study, we separate the models into the following three categories:

1. Temperature-index models which are mainly used in hydrological and glaciological applications [e.g., *Hock*, 2003; *Huss et al*., 2008].
2. Energy-balance models which have a wide variety of applications, such as in hydrology, land surface schemes, and weather forecasting models [e.g., *Shrestha et al*., 2010; *Zanotti et al*., 2004].
3. Snow-physics models which are, for example, used for avalanche warning and hydrology [e.g., *Bartelt and Lehning*, 2002; *Lehning et al*., 2006] as well as within snow research [e.g., *Vionnet et al*., 2012].

The two last categories both belong to the family of energy-balance snow models. However, the latter predict the microstructure of individual snowpack layers and give information about, for example, the mechanical stability of the snowpack, whereas the former feature a simplified snowpack layering.

3.1.1. Temperature-Index Models (TI-CDDF/TI-VDDF)

The temperature-index model was run using daily inputs of air temperature and precipitation separated into solid and liquid phases. In this study, we use two different options for this type of model structure. In a first experiment, we use a constant degree-day factor and in a second experiment, we allow this parameter to vary seasonally following *Slater and Clark* [2006]. That the degree-day factor varies during the season is well documented from observations [*Kuusisto*, 1980]. Including this seasonal variability of the melt factor should increase the temperature-index model performance compared to using a constant degree-day factor [*Rango and Martinec*, 1995]. The acronyms TI-CDDF and TI-VDDF are, respectively, used for the temperature-index model using a constant degree-day factor and the model with a seasonally varying degree-day factor. We additionally include a liquid water holding capacity in both models. The model parameters, three for TI-CDDF and four for TI-VDDF, were calibrated for each site individually following the methods presented by *Kokkonen et al*. [2006]. We used the Kling-Gupta efficiency [*Gupta et al*., 2009] as performance measure for the calibration (see section 3.2 for details about this statistic). Parameters are calibrated using data from the first half of the modeling period and evaluated over the second half. Subsequently, the calibration and evaluation periods are swapped to generate a complete, independent model data set for validation. The training data consist of the manual snow mass observations. For Weissfluhjoch, the runtime for this model is approximately 0.01 s/yr on a typical desktop computer.

3.1.2. Energy-Balance Models Represented in a Multimodel Framework (JIM)

JULES investigation model (JIM), a single snow model containing a multimodel framework [*Essery et al*., 2013], was the energy-balance model used to evaluate models of intermediate complexity. The multimodel

Table 1. Summary of Parameterizations Included in JIM[a]

	Snow Compaction [*Essery et al*., 2013, section 4.1]	
0	Physically based compaction rate depending on temperature, density, and overburden of snow	7
1	Empirical compaction rate depending on snow temperature	2
2	Constant snow density	1
	New Snow Density [*Essery et al*., 2013, section 4.2]	
0	Empirical function of air temperature and wind speed	4
1	Empirical function of air temperature	3
2	Constant new snow density	1
	Snow Albedo [*Essery et al*., 2013, section 4.3]	
0	Physically based parameterization depending on snow grain size	10
1	Empirical parameterization depending on snow surface temperature history	5
2	Empirical function of snow temperature	3
	Turbulent Heat Exchange [*Essery et al*., 2013, section 4.4]	
0	Obukhov length parameterization of adjustment for atmospheric stratification	5
1	Richardson number parameterization of adjustment for atmospheric stratification	3
2	Constant exchange coefficient	2
	Snow Cover Fraction [*Essery et al*., 2013, section 4.5]	
0	Empirical function of snow depth and surface roughness	1
1	Empirical tanh function of snow depth	1
2	Empirical linear function of snow depth	1
	Snow Hydraulics [*Essery et al*., 2013, section 4.6]	
0	Bucket model with liquid water capacity proportional to snow density	3
1	Bucket model with liquid water capacity proportional to snow porosity	1
2	Freely draining (i.e., the snowpack does not hold any liquid water)	0
	Thermal Conductivity of Snow [*Essery et al*., 2013, section 4.7]	
0	Empirical quadratic function of snow density	2
1	Empirical power function of snow density	2
2	Constant	1

[a]The numbers to the left indicate the three different options of available parameterizations, and the numbers to the right show the number of parameters used in each parameterization.

framework in JIM presents the user with a variety of methods for representing and modeling snow processes. The process representations vary in complexity. While some are highly parameterized and others not, we will adhere to the nomenclature used in *Essery et al*. [2013] and refer to all model choices as "parameterizations." Parameterization options are available for the following seven processes:

1. *Snow compaction*: The increase in snow density due to metamorphosis and weight of overlying snow.
2. *New snow density*: The influence of meteorological conditions on the size and shape of snowflakes determining the density of newly fallen snow.
3. *Snow albedo*: The variations in reflectivity of snow depending on grain types and the incident angle of shortwave radiation.
4. *Turbulent heat exchange*: The turbulent exchange of heat and moisture between the snow and atmosphere.
5. *Snow cover fraction*: The development of a patchy snow cover during melting influencing the averaged energy-balance for a certain area. This process can be important even in point-scale modeling when snow patches form adjacent to the model point [e.g., *Granger et al*., 2006]. In this study, where lysimeter data with a spatial footprint are used to evaluate point model performance, patchiness extending to the lysimeter evaluation area can additionally affect model evaluations if not accounted for.
6. *Snow hydraulics*: The process for routing liquid water through the snowpack.
7. *Thermal conductivity of snow*: The variations in thermal conductivity with snow density influencing the heat flux through the snowpack.

For each of the seven processes, JIM provides three different options of parameterizations: a so-called physically based option (numbered 0), an empirical option (numbered 1) and one simple option which either neglects the process, represents it using a constant value or uses very simple empirical formulation (numbered 2). The seven different processes and their three options of parameterizations are summarized in Table 1.

JIM iterates through all possible configurations of parameterizations, omitting infeasible combinations resulting in 1701 possible configurations. JIM solves the mass and energy exchanges for three individual snow layers. The surface heat balance equation is solved analytically and the vertical temperature profile in the snowpack is solved by the Crank-Nicolson method. For Weissfluhjoch, the runtime on a typical desktop computer for this model varied between approximately 0.3 and 0.6 s/yr depending on configuration.

3.1.3. Complex Snow-Physics Model (SNOWPACK)

The most complex snow models are an extension of the standard energy-balance models that simulate the internal structure of the snowpack in detail. To include this type of model in our evaluation, we assess the performance of the multilayer SNOWPACK model here. The model provides a detailed physically based process description for the surface energy balance and the heat flow through the snowpack [*Lehning et al*., 2002]; the internal snowpack microstructure, expressed by grain size, grain shape, bond size, sphericity, and dendricity [*Lehning et al*., 2002]; snow settling [*Bartelt and Lehning*, 2002]; and liquid water flow [*Wever et al*., 2014]. When using the recently introduced solver for the Richards equation for snow, as in this study, execution times are approximately 190 s/yr on a typical desktop computer; using the simpler bucket scheme and ignoring soil layers results in shorter calculation times. The model contains many parameterizations of physical processes, for example, new snow density, snow viscosity, and snow metamorphism, mostly determined by field or laboratory experiments. Although over 100 parameters are present in the model, typically only the surface roughness length, which strongly influences turbulent fluxes, needs to be specified. For Weissfluhjoch and Col de Porte, roughness lengths of 0.002 and 0.015 m, respectively, were chosen, identical to *Wever et al*. [2014].

3.2. Performance Measures

What should be considered the optimal model often changes depending on the choice of evaluation data and performance measure [*Essery et al*., 2013; *Kavetski and Fenicia*, 2011]. We therefore assess the model performance using a suite of different observations and goodness of fit measures relevant for hydrological applications.

3.2.1. Kling-Gupta Model Efficiency

We measure the goodness of fit between simulations and observations using the Kling-Gupta efficiency [*Gupta et al*., 2009]. This statistic was selected because it can be decomposed into a correlation term, a bias term, and a variability term. We compute the KGE-statistic, which has its optimum value at unity, using the modifications introduced by *Kling et al*. [2012]:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\alpha-1)^2} \tag{1}$$

$$\beta = \frac{\mu_s}{\mu_o} \tag{2}$$

$$\alpha = \frac{CV_s}{CV_o} \tag{3}$$

where r is the correlation coefficient, μ is the average, CV is the coefficient of variation, and the subscripts s and o represent simulations and observations, respectively. We compute the KGE for each site separately for the variables snowpack runoff, snow mass, snow depth, and snow surface temperature. In some analyses, we compute the combined performance for both sites by averaging the site specific efficiencies. Additionally, we also analyze the combined performance for several of the above mentioned variables by computing their average efficiency. Thus, we give both sites and the individual variables equal weight although the length of the data records varies slightly between the two locations.

3.2.2. Contingency Tables and Gerrity Skill Score

For flood forecasting, hydrological models should robustly capture flows in relevant categories, particularly during extreme events. Deviations between observed and measured flow that equally influence model efficiency measures as those presented above, do not necessarily bear equal importance to operational water managers. Based on the lysimeter data, we define four categories of flow conditions denoted as dry, low, medium, and high to evaluate situational model performance. We define the separation between the four categories using the 75th, 90th, and 99th percentiles of observed discharge (Figure 1). The limit separating the dry and low flow category was chosen rather high because the lysimeter at Col de Porte exhibits low

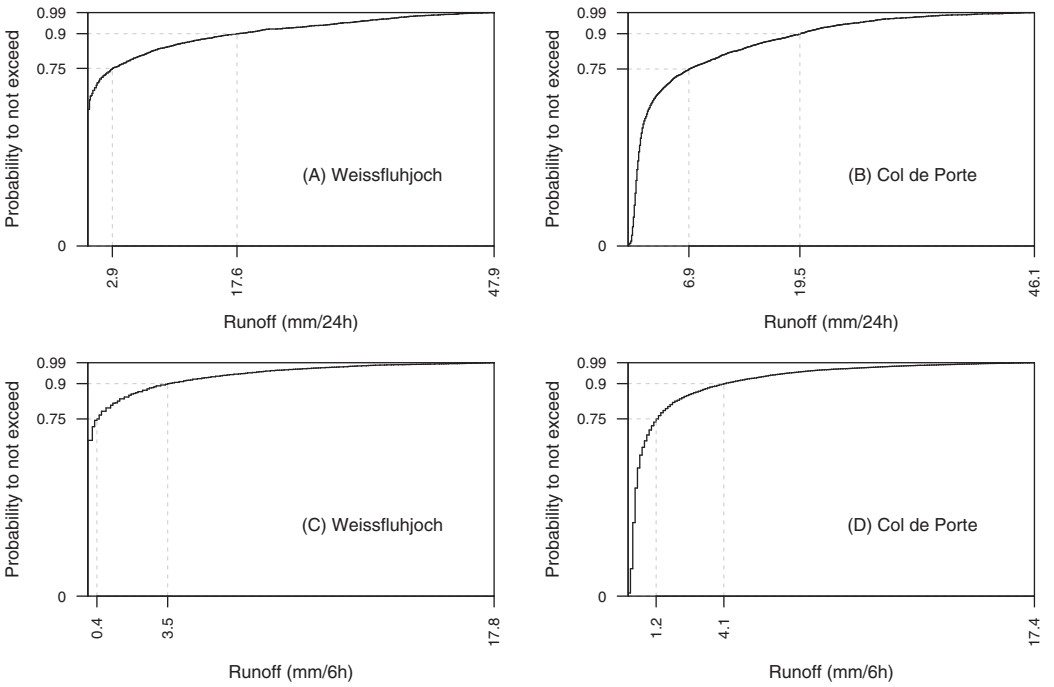

Figure 1. Cumulative distribution functions of (a and b) 24 h and (c and d) 6 h runoff sums observed by the snow lysimeter at the two experimental sites. The dashed lines indicate the 75th and 90th percentiles. The extreme right of the graph shows the 99th percentile of the distribution.

degrees of melt throughout winter (Figures 1b and 1d). By using the defined classes, we can construct contingency tables for interpreting the ability of the models to reproduce the observed flows [*Yossef et al.*, 2012] by flow class. To quantify the overall performance, we compute the Gerrity-Score (GS), which is a multicategory contingency score measuring the correspondence between simulations and observations [*Gerrity*, 1992; *Yossef et al.*, 2012]. This score varies between unity for perfect match and zero indicating no-skill (random or constant forecast). The score rewards hits within less likely categories higher than within probable categories. At the same time, large forecast errors (e.g., predicting low but observing high flow) are punished harder than smaller errors (e.g., predicting medium but observing high flow).

4. Results and Discussion

4.1. Relationship Between Model Complexity and Performance

In this study, we measure model complexity by counting the number of parameters used in the different JIM configurations (see Table 1). Note that, unlike the temperature-index model, none of these parameters were calibrated. In this and the following section 4.2, a combined analysis of data from both sites is presented. The highest KGE increases quicker with model complexity for runoff and snow mass than for snow depth (Figures 2a–2c), and even decreases slightly for snow surface temperature with increasing number of parameters (Figure 2d). For runoff and snow mass, the efficiency appears to reach an upper limit as the number of parameters increases without additional improvement by added complexity. For the most parameter-rich JIM configurations, KGE even decreases slightly for both runoff and snow depth (Figures 2a and 2c). At the same time, the lowest KGE increases with the number of parameters, foremost for runoff and snow mass (Figures 2a and 2b) reducing the spread in model performance throughout the different JIM configurations. The remaining large spread for the parameter-rich configurations indicate that a model with high complexity does not guarantee good performance. However, note also that the number of possible configurations decreases for low or high numbers of parameters. The combined model performance for runoff and snow mass indicates that complex models are not absolutely necessary for hydrological applications requiring daily values of those variables (Figure 2e); models with about 16 parameters seem to suffice to reach optimum model performance. When including snow depth into the combined performance measure only the more parameter-rich models show optimal performance (Figure 2f). In many cases, snow

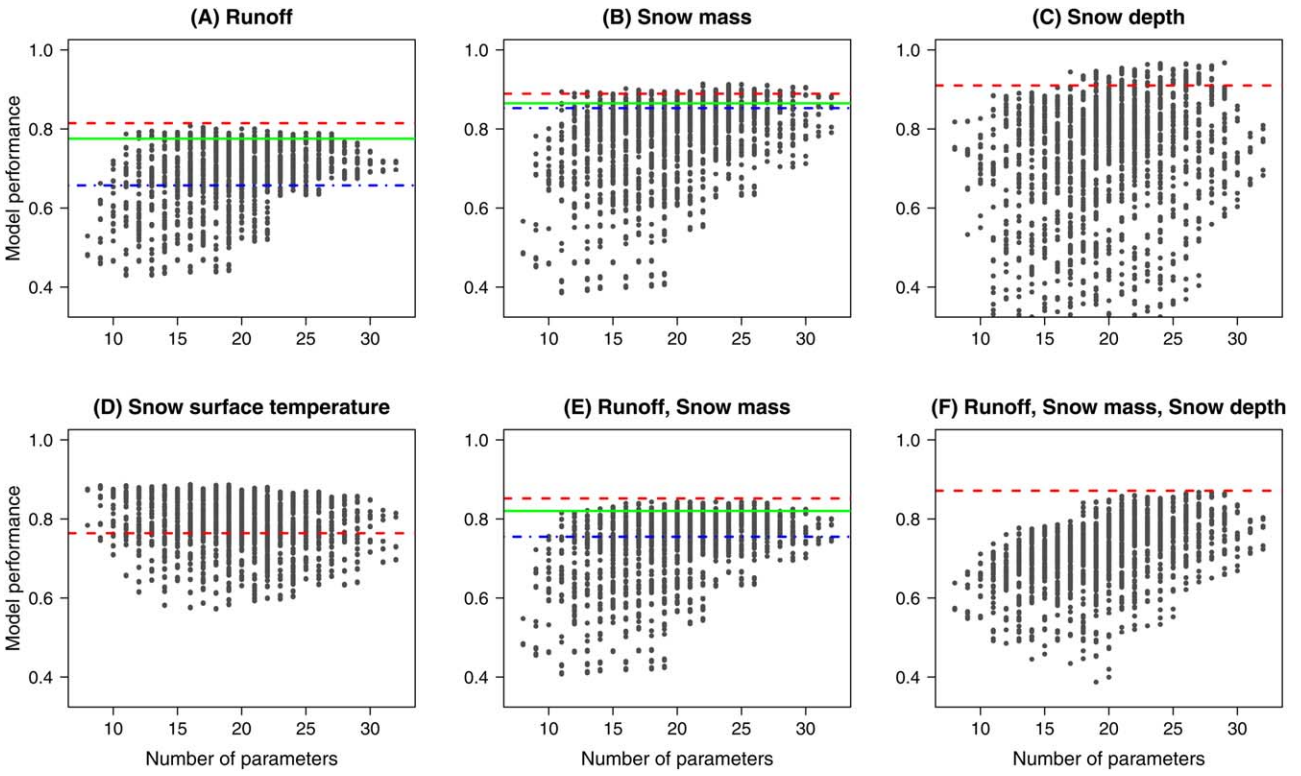

Figure 2. Model performance, measured by the Kling-Gupta efficiency (equation (1)), for daily values and averaged for both sites plotted against number of parameters for the complete set of JIM configurations (gray dots). SNOWPACK (red dashed line) and TI-VDDF (green solid line) achieve results with performance similar to the best JIM combinations. However, TI-CDDF (blue dash-dotted line) shows lower performance for runoff than the JIM configurations with high model efficiency. SNOWPACK and the temperature-index models are shown as horizontal lines for clarity since their number of parameters differs largely from JIM.

depth is used for assimilation into models [*Brown et al*., 2003; *Liu et al*., 2013]. In such applications, models require an appropriate level of complexity to accurately depict snow density in order to obtain reliable estimates of snow mass from measurements of snow depth.

For most variables, SNOWPACK and the best JIM configurations reach similar performance (Figure 2) even though they differ greatly in their description of internal snowpack structure (see section 3.1). Otherwise, both models are very similar, particularly how they compute the energy transfer between the atmosphere and the snowpack. Therefore, other factors than the model formulation of the internal snowpack structure appear to limit model performance. Likely maximum model performance is limited by errors in input and evaluation data as well as deficiencies in the formulation of the surface heat exchanges.

Both temperature-index models were calibrated using only snow mass observations and provide results for this variable with only slightly lower efficiency than the physically based models (Figure 2b). For simulating runoff, however, the temperature-index models differ in performance with only TI-VDDF approaching the performance of the best JIM configurations and SNOWPACK (Figure 2a). Thus, including seasonal variations in the melt parameter increases the model performance of the degree-day method as already suggested by *Rango and Martinec* [1995]. Nevertheless, even though it is known that a constant melt parameter deteriorates the model efficiency for runoff many studies still apply this method [e.g., *Walter et al*., 2005; *Warscher et al*., 2013]. Thus, for a high performance, an appropriate use of a model seems more important than high model complexity.

To summarize, for daily values of runoff and snow mass, all model types can provide simulations with high model performance except for the simplest temperature-index approach which displays lower efficiency for runoff (Figures 2a, 2b, and 2e). The number of parameters and the run-time, on the other hand, differ largely between the model types (Table 2). The complex snow-physics model requires much longer computation time than the two other types of models and includes many more parameters. The temperature-index model runs faster than JIM and SNOWPACK, and relies on fewer input variables but requires site-specific calibration.

Table 2. Simulation Performance for Daily Values of Runoff and Snow Mass for the Different Model Types Including Number of Parameters and Model Run-Time[a]

Model	KGE for Runoff	KGE for Snow Mass	KGE for Runoff and Snow Mass	Number of Parameters	Approximate Run-Time (s)
TI-CDDF	0.66	0.85	0.76	3	0.01
TI-VDDF	0.78	0.86	0.82	4	0.01
JIM	0.81	0.87	0.84	16	0.3
SNOWPACK	0.81	0.89	0.85	>100	190

[a]The JIM configuration with the best performance for snowpack runoff. The number of parameters in SNOWPACK was only counted for the seven processes listed in Table 1.

Since all model types seem capable of providing daily runoff and snow mass simulations with similar quality, model choices will depend on data availability, computer resources, and run time constraints.

4.2. Relationship in Performance Between Variables

Many hydrological applications in cold regions require snowpack runoff predictions without any need for forecasting other variables such as snow depth. However, most studies evaluating snow models use such auxiliary data [*Carrera et al*., 2010; *De Michele et al*., 2013; *Schmucki et al*., 2014] which may not be a good indicator of snowpack runoff and therefore not valuable for selecting models for hydrological applications. We find that model configurations which reproduce snow mass with high KGE also tend to produce accurate runoff simulations (Figure 3a). However, even when a model shows high agreement with the snow mass observations, there is still considerable variation in runoff performance. Furthermore, the KGE values for runoff correlate weakly with both the KGE values for snow depth and snow surface temperature (Figures 3b and 3c). This result agrees with earlier studies showing that high model performance for one variable can be achieved even though other variables may not be adequately represented [*Bloschl et al*., 1991]. In particular, a wide range of JIM configurations reproduce daily snowpack runoff with high KGE, but show large variations in performance for capturing snow depth (Figure 3b). It seems as if the choice of snow density model does not largely influence the ability of the models to reproduce the daily runoff observations. Thus, snow depth observations alone are not a critical measure for testing the reliability of snow models for hydrological applications requiring daily time steps. The same conclusion also appears valid for snow surface temperature observations.

From the analysis above, we find that model configurations which reproduce snow mass with high efficiency also tend to perform well for runoff. However, a considerable variation in runoff performance still remains for JIM configurations which capture snow mass with high efficiency (Figure 3a). For a subset of model configurations which reproduce snow mass accurately (the 20% of JIM configurations with highest KGE for snow mass), three different process representations show systematic influences on the runoff model performance (Figure 4). First, the more complex albedo formulations (options 0 and 1) provide runoff

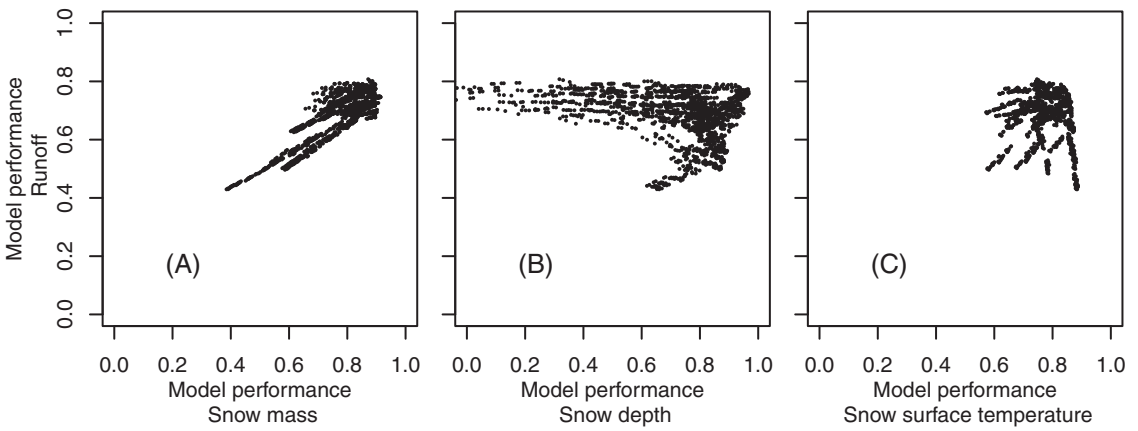

Figure 3. Relationships between model performance, given by the KGE-statistic, for daily runoff against snow mass, snow depth, and snow surface temperature, respectively. The model efficiency was computed for each site separately and afterward averaged between the sites for each JIM configuration.

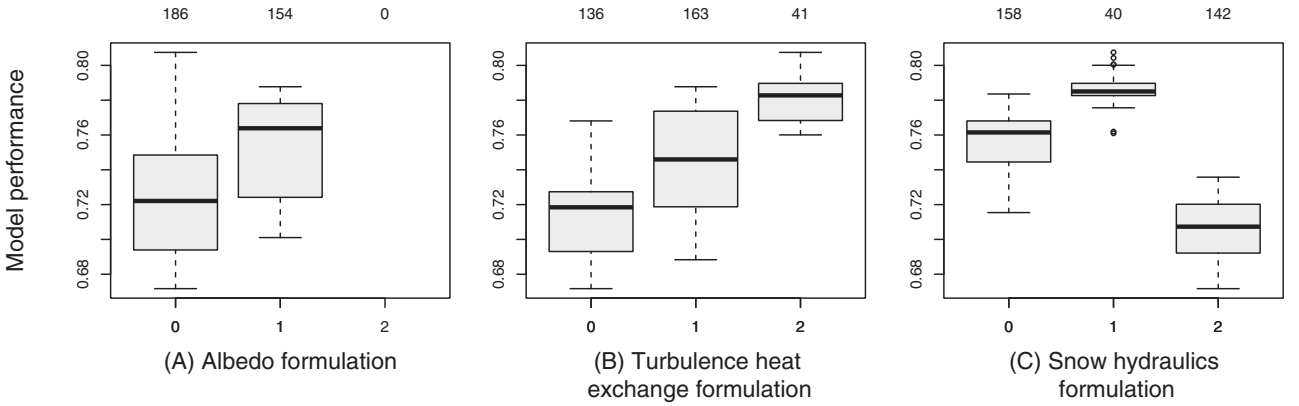

Figure 4. Box plots showing the model performance for runoff depending on the choice of parameterization for three different processes. The plots only show configurations which match the snow mass observations with high model efficiency (the 20% of JIM configurations with highest KGE for snow mass). The numbers above the box plots indicate the number of times each option was selected.

simulations with similarly high performance (see Table 1 for a summary of the processes and parameterizations in JIM). On the other hand, none of the model configurations that included the simplest albedo formulation (option 2) were in the top 20% of snow mass performing models. Second, the formula for calculating turbulent heat exchange using a constant exchange coefficient (option 2) produces better runoff simulations than the methods taking atmospheric stability into account (options 0 and 1). Third, using either one of the two bucket model options (options 0 and 1) for describing snow hydraulics gives the best results, compared to not including water retention in the snowpack (option 2). Finally, the choices of parameterizations in JIM for the remaining processes, for example, snow cover fraction and new snow density, do not seem to influence the quality of daily runoff predictions provided the model captures snow mass with sufficient accuracy (results not shown).

4.3. Relationship in Model Performance Between the Sites

The model performance for the different JIM configurations shows a positive correlation between the two sites for runoff, snow mass, and snow depth (black and gray dots in Figures 5a, 5b, and 5c). Figure 5d shows that for snow surface temperature, the models using a stability correction (options 0 and 1 represented by gray dots) perform better for Col de Porte than the models relying on a constant exchange coefficient (option 2 represented by black dots). For the performance of snow surface temperature at Col de Porte, our results differ from those presented by *Essery et al.* [2013] who judged the model performance using root mean squared error. However, our results become similar to the previous study if applying the same error measure. This example illustrates that the performance criteria for evaluating simulations can certainly influence model assessments. In contrast to Col de Porte, at Weissfluhjoch, the configurations using the simpler turbulence scheme (option 2) provide the highest model efficiency for snow surface temperature (Figure 5d). The SNOWPACK model, which employs the Monin-Obukhov formulation for surface heat and moisture fluxes, follows the behavior of the best performing JIM configurations which also use stability corrections for the turbulent fluxes. Thus, the methods for computing the turbulent heat and moisture exchange appears to lack transferability between the sites, indicating a potential for improving the energy-balance approach. However, reducing the uncertainty in the computations of this energy-balance component is difficult since we largely lack direct observations of the turbulent energy exchange over closed snow cover at different locations [*Stoessel et al.*, 2010].

4.4. Variation in Model Performance for Runoff Between Years

The model efficiency, computed for each year and model separately, varies over time for all model types (upper plot in Figures 6 and 7). In many years, the differences in performance between the models seem small; in some winters, all models agree well with the observations whereas in other winters, all models reproduce the measurements poorly. Indeed, the yearly KGE-statistic shows a positive correlation, ranging from 0.86 to 0.96 for Col de Porte and from 0.65 to 0.96 for Weissfluhjoch, between the three model types (the best JIM configuration, SNOWPACK, and TI-VDDF). The high correlation indicates that one or more

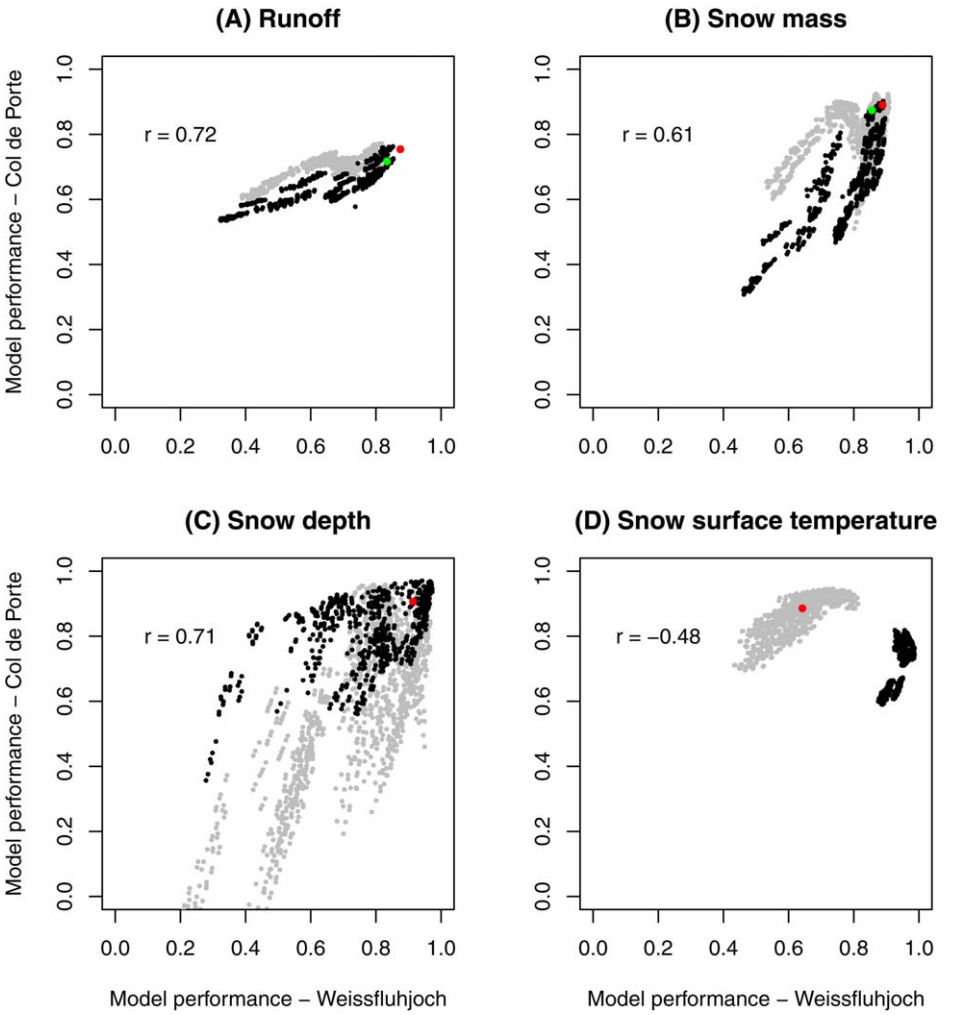

Figure 5. Correlation in model performance, measured by the KGE-statistic, between the two sites for four variables. The gray dots show the model configurations using the surface exchange schemes including atmospheric stability corrections (options 0 and 1), whereas the black dots show configurations using a constant exchange coefficient (option 2). The red dot indicates the SNOWPACK model results and the green dot indicates the TI-VDDF model results. Note that TI-VDDF does not simulate snow depth and snow surface temperature.

confounding factors influence the ability of all models to reproduce daily runoff. SNOWPACK (Figure 7a, red line) shows slightly better performance than the best JIM configuration (Figure 7a, blue line) and TI-VDDF (Figure 7a, green line). TI-VDDF shows the most variable results, with the highest performance in several years but also often the worst model efficiency in other years. The best JIM configuration shows the most consistent results, seldom outperforming the other models and at the same time rarely producing the worst results in individual years. For both sites, the yearly best performing JIM configuration (Figures 6a and 7a, black dashed line) outperforms the other models for most of the years. By applying an appropriate model averaging technique [i.e., *Rings et al*., 2012] more reliable snowmelt predictions including uncertainty estimates may be generated from an ensemble of well-performing JIM configurations.

At both sites, the model performance often seems to deteriorate due to a mismatch between the simulated mean runoff and observed mean runoff (Figures 6c and 7c). In particular for Col de Porte, the simulations show lower runoff than the observations for most winters after 1999–2000. Before this winter, precipitation was adjusted to match manually observed amounts of new snow, and later on corrected for undercatch using wind speed and air temperature measurements [*Morin et al*., 2012]. Thus, the method for correcting the precipitation measurements may have influenced the simulation quality. In general, the mismatch between simulated and observed runoff amounts, which is mostly independent of the model type, shows the importance of providing accurate precipitation input data to the models.

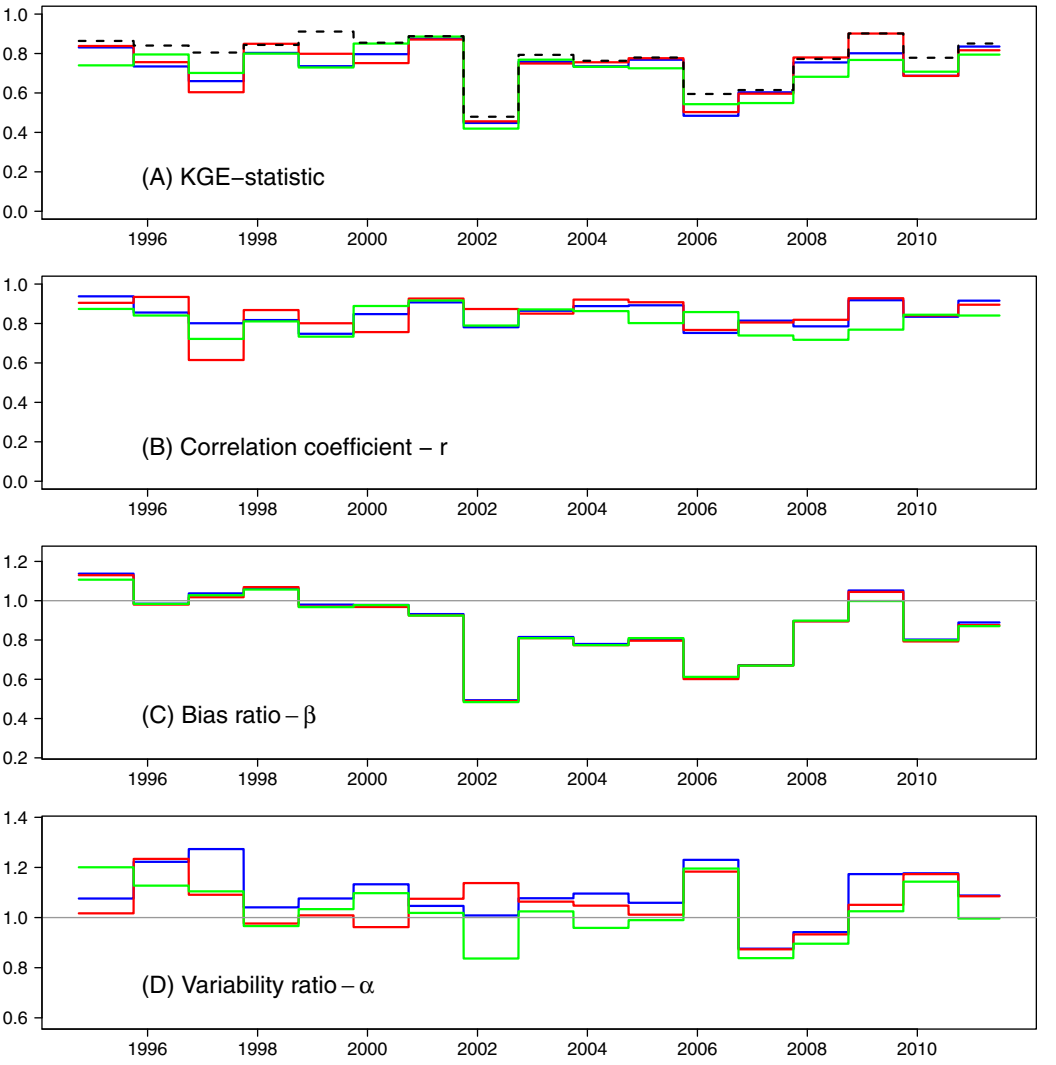

Figure 6. Model performance for daily runoff computed for each year individually at Col de Porte for the overall best performing JIM simulation (blue line), yearly best JIM results (black dashed line), SNOWPACK (red line), and TI-VDDF model (green line). The model performance is judged by the modified Kling-Gupta efficiency and its decomposition into the correlation coefficient (r), the ratio between the mean of the simulation and the mean of the observations (α), and the ratio between the coefficients of variation for simulations and observations (β).

4.5. Evaluation of Runoff Simulations in Categories

The contingency tables show that all three model types capture daily snowpack runoff with similar hit rates (Table 3). All models display many hits in the dry category due to the simple prediction of nonmelting conditions occurring during cold periods. In this category, depending on model and site, between 91.0% and 94.6% of the events were correctly simulated by the models. For the remaining categories, all models show many misses indicating the difficulty of predicting snowpack runoff within those categories robustly. For those classes, the fraction of hits within each category varied between 28.1% and 82.1% depending on model and site. The models show a slightly lower ability to capture the runoff in categories for Col de Porte than Weissfluhjoch. At Col de Porte, between 82.2% and 84.6% of all observed events were correctly captured by the simulations, whereas the corresponding number for Weissfluhjoch varied between 84.9% and 88.5%. All model types fail to forecast approximately half or more of the events observed in the high flow category. Depending on model and site, only between 28.1% and 51.4% of the observed high flow events were captured by the simulations with the lowest hit rate obtained by TI-VDDF at Col de Porte. For high flows at Weissfluhjoch, JIM and SNOWPACK show fewer false alarms than TI-VDDF. Both JIM and SNOWPACK show similar results for the high flow events.

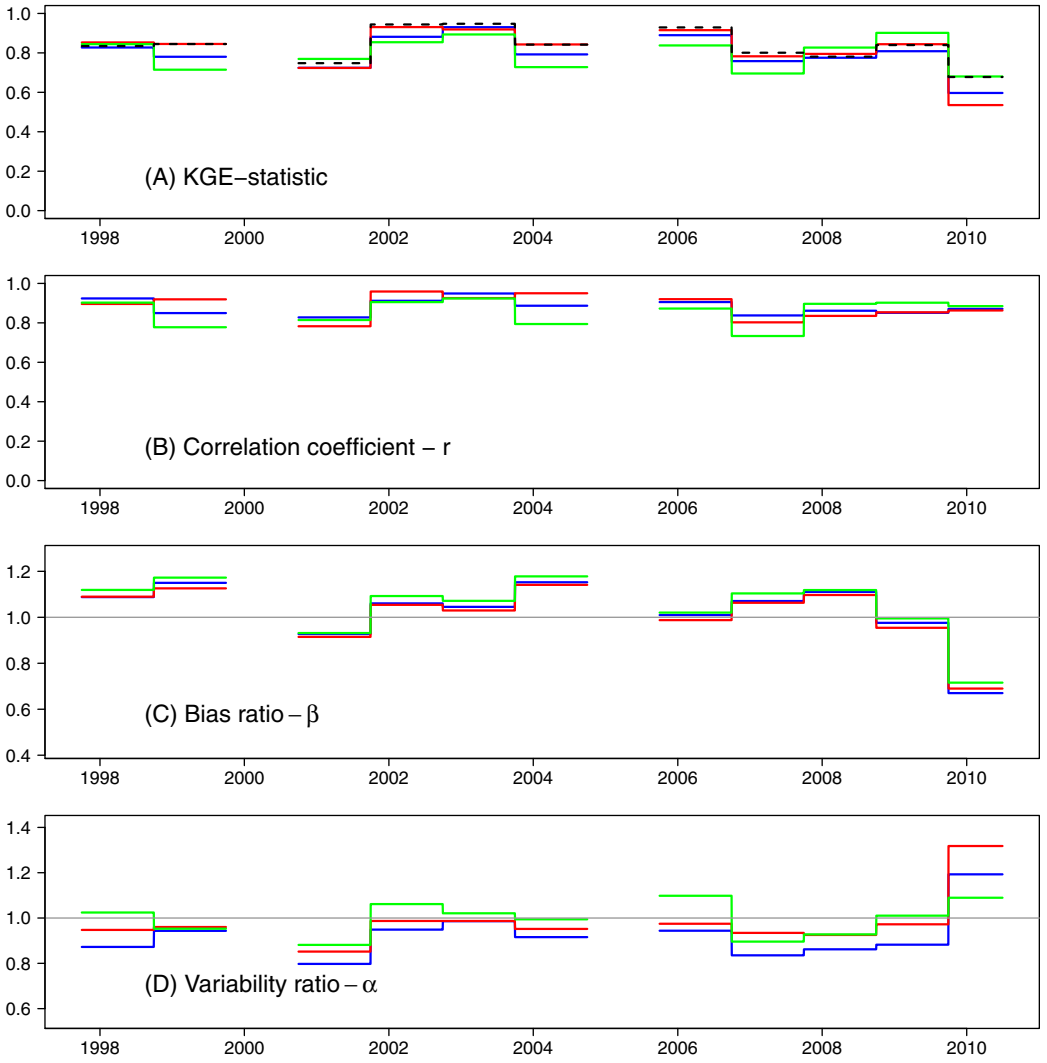

Figure 7. Model performance for daily runoff computed for each year individually at Weissfluhjoch for the overall best performing JIM simulation (blue line), yearly best JIM results (black dashed line), SNOWPACK (red line), and TI-VDDF model (green line). The model performance is judged by the modified Kling-Gupta efficiency and its decomposition into the correlation coefficient (r), the ratio between the mean of the simulation and the mean of the observations (α), and the ratio between the coefficients of variation for simulations and observations (β).

Visual inspection of the single observed events in the high flow category (data not shown) reveals that some of the misses occur during late-spring when the simulations show too early snow disappearance and that others occur during rain-on-snow events. Our study confirms earlier results which show the difficulty of accurately predicting runoff around the date for melt-out [e.g., *Jin et al*., 1999] and snowpack outflow during rain-on-snow events [e.g., *Barry et al*., 1990]. Finally, the lysimeter data might not be representative of the point-simulated runoff totals due to lateral and preferential water flow paths in the snowpack [*Kattelmann*, 2000; *Wever et al*., 2014].

The three different model types show skill (Gerrity-Score) in reproducing daily snowpack runoff (Table 4). Overall, the simulations show higher skill in capturing the categorical events at Weissfluhjoch than at Col de Porte. The main reason for this behavior is the simpler prediction of runoff from the thicker high-alpine snowpack not prone to the mid-winter freeze/thaw cycles encountered at the lower elevation site. The physically based models seem to outperform the temperature-index model, and SNOWPACK shows higher skill than JIM. For Weissfluhjoch, the difference between SNOWPACK and JIM is lower than for Col de Porte. The difference in skill between those two models may arise for at least two reasons. First, SNOWPACK was driven using measured albedo whereas JIM was driven with simulated albedo. Second, the SNOWPACK

Table 3. Categorical Contingency Tables for Dry (D), Low (L), Medium (M), and High (H) Flow Separated by the 75th, 85th, and 99th Percentiles of Observed Daily Snowpack Runoff[a]

		Col de Porte				Weissfluhjoch				
Simulations	D	2146	112	25	2	2378	102	10	0	JIM
	L	206	289	77	3	188	339	46	2	
	M	5	67	172	16	25	65	246	19	
	H	0	0	9	11	0	1	6	14	
										SNOWPACK
	D	2164	101	29	3	2451	100	21	2	
	L	185	315	82	1	126	324	24	1	
	M	8	52	164	15	14	83	253	14	
	H	0	0	8	13	0	0	10	18	
	D	2154	148	15	2	2393	153	5	0	TI-VDDF
	L	190	271	114	6	170	287	56	0	
	M	13	48	146	15	26	65	223	18	
	H	0	1	8	9	2	2	24	17	
		D	L	M	H	D	L	M	H	
					Observations					

[a]See also Figure 1.

version [*Wever et al*., 2014] used in this study includes a detailed description of water transport through snow which is not included in JIM.

Very few previous studies have evaluated long-term simulations of snowpack runoff against snow lysimeter observations [e.g., *Wever et al*., 2014]. In other studies, such comparisons have been made, but for much shorter periods and without the use of several complementary performance measures [e.g., *Albert and Krajeski*, 1998; *Barry et al*., 1990; *Foerster et al*., 2014]. For practical applications, the use of different skill measures as presented in this study can provide information about several relevant aspects of the model behavior. For example, the ability of models to predict flooding events has not been thoroughly assessed in earlier snow model intercomparison projects. Thus, we encourage future studies to follow and extend the analysis presented here using multiple performance metrics and skill measures.

4.6. Evaluation of Daily and Subdaily Runoff Predictions

For flood forecasting of large watersheds, running a hydrological model at a daily time step should suffice. For such a time step, both JIM and SNOWPACK reproduce the observed lysimeter runoff with similar performance (Figure 8). In general, these models perform slightly better at Weissfluhjoch than Col de Porte likely because the snowpack at the lower elevated site undergoes more melt events during mid-winter than at the high alpine site. As pointed out above, some of the deviations between simulations and observations may arise due to, for example, lateral flows in the snowpack influencing the lysimeter measurements. Another cause for those deviations can be the spatial variability of the snow cover. The site Col de Porte is equipped with two lysimeters. Those two lysimeters show mutual agreement (RMSE = 5.6 mm/24 h, r = 0.87, KGE = 0.83, GS = 0.77) which is only slightly better than the agreement between the simulations and observations presented in Figure 8. Thus, even with error free lysimeter observations, we should not expect a perfect match between the observation and simulation results due to spatial variability of the snow cover.

For reliable flood forecasting of small watersheds, hydrological models require shorter time steps than for larger catchments. For subdaily time steps, SNOWPACK shows higher performance than JIM (Figure 9). SNOWPACK includes a detailed description of liquid water transport through the snowpack whereas JIM represents this process with a so-called bucket approach. For shorter periods than those presented here, JIM reproduces the observations with even lower performance (data not shown). *Wever et al*. [2014] showed that physically based methods using Richards equations outperformed snowpack runoff simulations using simple bucket approaches for hourly observations. Thus, more effort should be invested in improving and testing the description of liquid water transport through the snowpack in JIM to increase the model performance for short time steps.

Table 4. Gerrity-Score Computed for the Two Sites Individually and Combined for the Three Different Model Types

	Col de Porte	Weissfluhjoch	Average of Both Sites
JIM	0.57	0.66	0.62
SNOWPACK	0.59	0.71	0.65
TI-VDDF	0.51	0.67	0.59

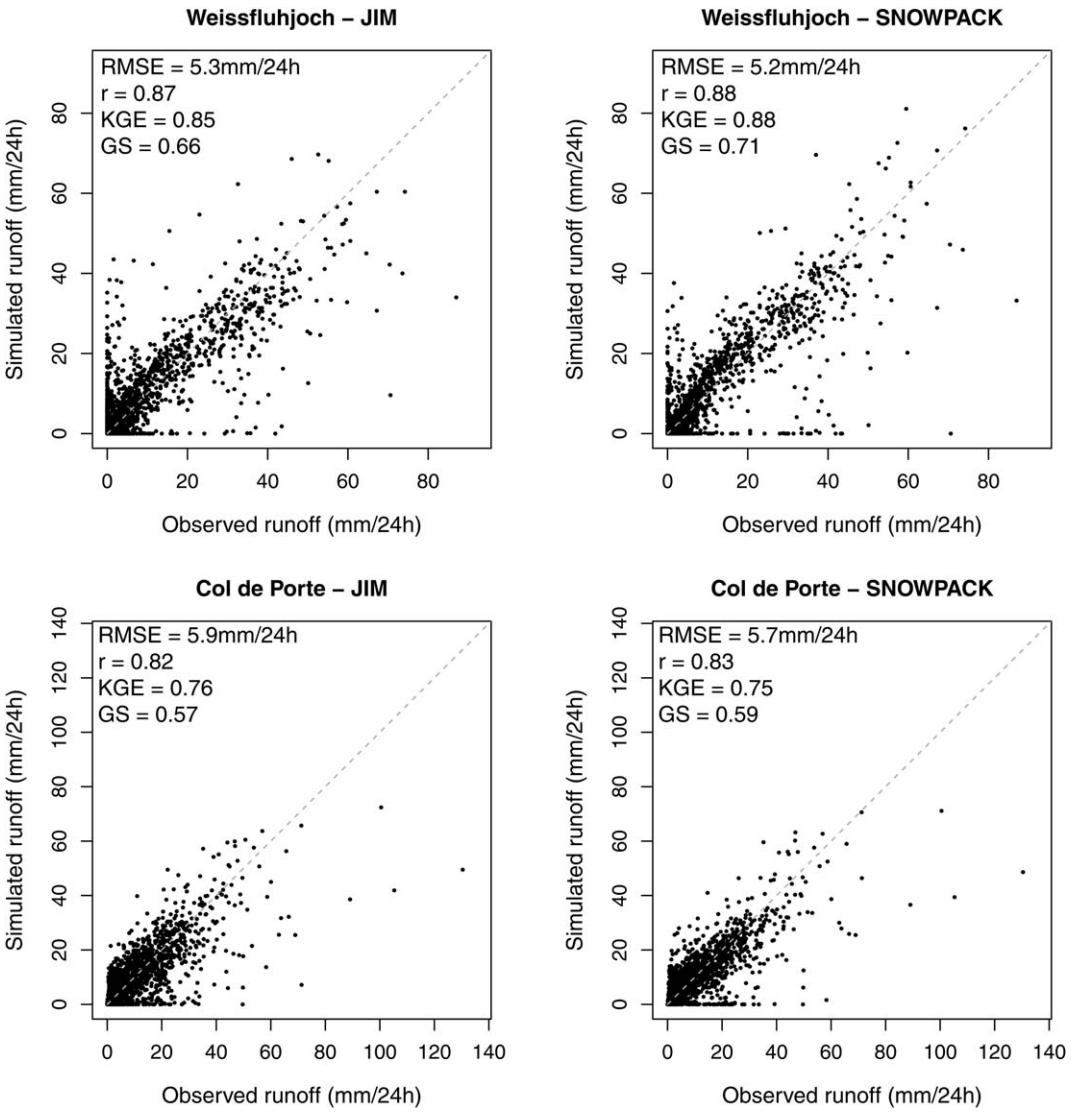

Figure 8. Scatter plots showing simulated against observed 24 h sums of lysimeter runoff.

5. Summary and Conclusions

In this study, we compare and evaluate three different types of snow models with a focus on hydrological applications:

1. The temperature-index method represented by two different implementations, one using a constant degree-day factor (TI-CDDF) following, e.g., *Warscher et al*. [2013] and one including a seasonally varying degree-day (TI-VDDF) following, e.g., *Slater and Clark* [2006].
2. A multimodel framework (JIM) providing an ensemble of snowpack energy-balance models with varying complexity depending on the selection of process representations [*Essery et al*., 2013].
3. A complex snow-physics model (SNOWPACK) with a detailed description of the layered snowpack [*Bartelt and Lehning*, 2002; *Wever et al*., 2014].

Our selection of models covers a large range of existing snow models. We evaluated the models at two sites in the European Alps at mid and high-elevations using high quality long-term recordings of input and validation data.

In our study, for daily predictions of snowpack runoff and snow mass, model complexity is not a determinant for high model performance. The best JIM configurations and SNOWPACK, as well as the temperature-

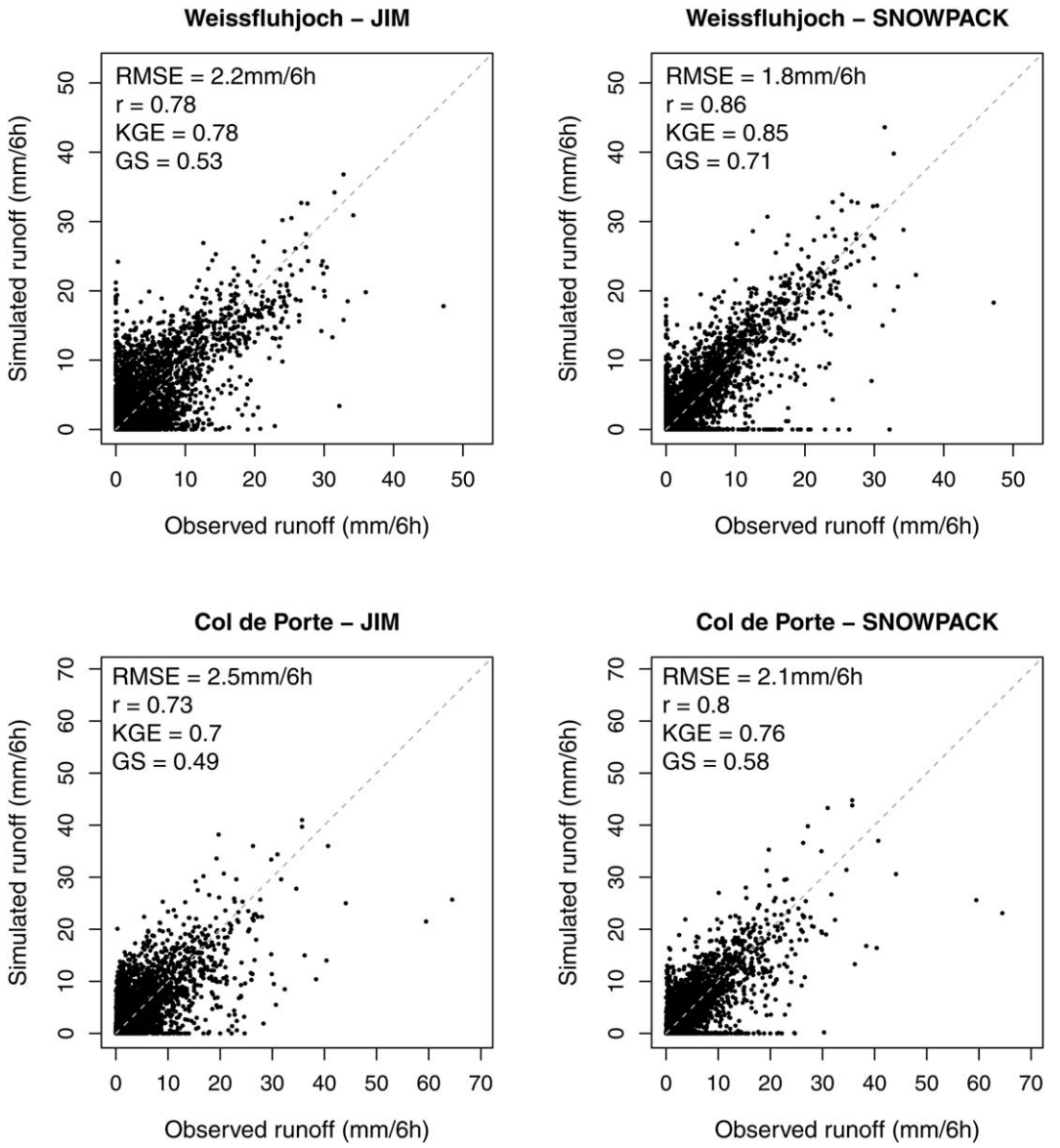

Figure 9. Scatter plots showing simulated against observed 6 h sums of lysimeter runoff.

index method including a variable melt factor can provide similar model efficiency (Figure 2) and similar ability to reproduce high runoff events (Tables 3 and 4). For daily runoff and snow mass, the ensemble of JIM models provided a model performance with a Kling-Gupta efficiency for individual models approximately ranging from 0.41 to 0.84 (Figure 2). For testing whether new model developments provide additional skill, the JIM ensemble provides an excellent benchmark.

JIM results show that models which reproduce snow mass accurately also tend to capture snowpack runoff with high performance (Figure 3a). On the contrary, clear relationships were missing when comparing model performance for snowpack runoff with snow depth and snow surface temperature (Figures 3b and 3c). Thus, our study suggests (a) that modeling daily snowpack runoff reliably does not require very accurate snow density and snow surface temperature simulations, and (b) that many frequently observed variables (e.g., snow depth) provide only limited information for evaluating snow models intended for hydrologic applications. These conclusions though are based on data from two alpine sites. Site-specificity and potential surface-temperature compensations within the turbulent exchange parameterizations deserve consideration in interpreting these results. Different conditions, such as sites with thin, cold, low density snow covers, or other areas where turbulent exchanges are

more dominant might be more sensitive to accurate portrayals of snow depth, density, and surface temperature.

All model performance measures for snow mass, snowpack runoff, and snow depth exhibit a positive correlation between the two sites (Figures 5a–5c). Performance measures for snow surface temperature, however, did not exhibit the same between-site correlation (Figure 5d). The methods for computing the turbulent heat fluxes, which is the least validated component in the snowpack energy-balance, greatly influences the snow surface temperature (Figure 5d) and snowpack runoff simulations (Figure 4b). Thus, more studies should focus on reducing the uncertainty in the simulation of this energy-balance component, which would require direct measurements of turbulent heat exchange over snow [e.g., *Helgason and Pomeroy*, 2012; *Mahat et al*., 2013; *Reba et al*., 2014].

The ability to reproduce daily runoff varied strongly between the years but simultaneously for all different model types (Figures 6 and 7). Inaccurate precipitation input data or nonrepresentative snowpack runoff data appears to have degraded the validation of all models in some years. While model development seems to receive more attention than acquiring accurate input and validation data, our results suggest that both are necessary at the same time to further advance existing modeling approaches.

Our study demonstrates that for snowpack runoff predictions at daily timescale, an appropriately set up energy-balance model with a simplified snowpack structure can provide nearly identical performance as a much more complex snow-physics model (Figure 8). Even at 6 h time scale, the model efficiencies can be similar (Figure 9). Energy-balance models with a simplified snowpack structure have a much shorter runtime and include far fewer parameters than the snow-physics model (Table 2). Thus, for many applications, this simpler model type may constitute the optimal trade-off between model performance, computational constraints, and model complexity limits required by data assimilation frameworks [e.g., *Magnusson et al*., 2014; *Slater and Clark*, 2006]. However, we could show that an appropriate combination of process representations is mandatory to achieve a high model performance with this type of model. The JIM multimodel framework of energy-balance models used here provides an excellent tool for identifying appropriate model structures, i.e., the correct combination of parameterizations. Such inferences are less straightforward to make in typical model intercomparison studies [i.e., *Etchevers et al*., 2004] due to intermodel differences in numerical schemes and parameter values masking the influence of differences in the parameterizations themselves. In this study, we demonstrated the model selection process with regards to snow hydrological applications. We suggest using such a framework to optimize specific model applications under given boundary conditions such as data availability, properties of interest, and computational constraints.

Acknowledgments
Data used in this study can be made available upon request. We thank Martyn Clark and two additional reviewers for the helpful comments on this paper. This study was partly funded by the Federal Office of the Environment FOEN.

References

Albert, M., and G. Krajeski (1998), A fast, physically based point snowmelt model for use in distributed applications, *Hydrol. Processes*, *12*(10–11), 1809–1824.

Barry, R., M. Prevost, J. Stein, and A. P. Plamondon (1990), Application of a snow cover energy and mass balance model in a balsam fir forest, *Water Resour. Res.*, *26*(5), 1079–1092.

Bartelt, P., and M. Lehning (2002), A physical SNOWPACK model for the Swiss avalanche warning Part I: Numerical model, *Cold Reg. Sci. Technol.*, *35*(3), 123–145.

Bloschl, G., D. Gutknecht, and R. Kirnbauer (1991), Distributed snowmelt simulations in an alpine catchment. 2. Parameter study and model predictions, *Water Resour. Res.*, *27*(12), 3181–3188.

Brown, R. D., B. Brasnett, and D. Robinson (2003), Gridded North American monthly snow depth and snow water equivalent for GCM evaluation, *Atmos. Ocean*, *41*(1), 1–14.

Carrera, M. L., S. Belair, V. Fortin, B. Bilodeau, D. Charpentier, and I. Dore (2010), Evaluation of snowpack simulations over the Canadian rockies with an experimental hydrometeorological modeling system, *J. Hydrometeorol.*, *11*(5), 1123–1140.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

Cox, G. M., J. M. Gibbons, A. T. A. Wood, J. Craigon, S. J. Ramsden, and N. M. J. Crout (2006), Towards the systematic simplification of mechanistic models, *Ecol. Modell.*, *198*(1–2), 240–246.

De Michele, C., F. Avanzi, A. Ghezzi, and C. Jommi (2013), Investigating the dynamics of bulk snow density in dry and wet conditions using a one-dimensional model, *Cryosphere*, *7*(2), 433–444.

Dutra, E., G. Balsamo, P. Viterbo, P. M. A. Miranda, A. Beljaars, C. Schaer, and K. Elder (2010), An improved snow scheme for the ECMWF land surface model: Description and offline validation, *J. Hydrometeorol.*, *11*(4), 899–916.

Essery, R., S. Morin, Y. Lejeune, and C. B. Menard (2013), A comparison of 1701 snow models using observations from an alpine site, *Adv. Water Resour.*, *55*, 131–148.

Etchevers, P., et al. (2004), Validation of the energy budget of an alpine snowpack simulated by several snow models (SnowMIP project), *Ann. Glaciol.*, *38*, 150–158.

Foerster, K., G. Meon, T. Marke, and U. Strasser (2014), Effect of meteorological forcing and snow model complexity on hydrological simulations in the Sieber catchment (Harz Mountains, Germany), *Hydrol. Earth Syst. Sci.*, *18*(11), 4703–4720.

Franz, K. J., T. S. Hogue, and S. Sorooshian (2008), Operational snow modeling: Addressing the challenges of an energy balance model for National Weather Service forecasts, *J. Hydrol.*, *360*(1–4), 48–66.
Gerrity, J. P. (1992), A note on Gandin and Murphy equitable skill score, *Mon. Weather Rev.*, *120*(11), 2709–2712.
Granger, R. J., R. Essery, and J. W. Porneroy (2006), Boundary-layer growth over snow and soil patches: Field observations, *Hydrol. Processes*, *20*(4), 943–951.
Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*(1–2), 80–91.
Helgason, W., and J. Pomeroy (2012), Problems closing the energy balance over a homogeneous snow cover during midwinter, *J. Hydrometeorol.*, *13*(2), 557–572.
Hock, R. (2003), Temperature index melt modelling in mountain areas, *J. Hydrol.*, *282*(1–4), 104–115.
Huss, M., D. Farinotti, A. Bauder, and M. Funk (2008), Modelling runoff from highly glacierized alpine drainage basins in a changing climate, *Hydrol. Processes*, *22*(19), 3888–3902.
Jin, J., X. Gao, Z. L. Yang, R. C. Bales, S. Sorooshian, and R. E. Dickinson (1999), Comparative analyses of physically based snowmelt models for climate simulations, *J. Clim.*, *12*(8), 2643–2657.
Kattelmann, R. (2000), Snowmelt lysimeters in the evaluation of snowmelt models, *Ann. Glaciol.*, *31*, 406–410.
Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, *Water Resour. Res.*, *47*, W11511, doi:10.1029/2011WR010748.
Kling, H., M. Fuchs, and M. Paulin (2012), Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, *J. Hydrol.*, *424*, 264–277.
Kokkonen, T., H. Koivusalo, A. J. Jakeman, and J. Norton (2006), Construction of a degree-day snow model in the light of the 10 iterative steps in model development, in Proceedings of the iEMSs Third Biennial Meeting: "Summit on Environmental Modelling and Software," *[CD-ROM]*, edited by A. Voinov, A. J. Jakeman, and A. E. Rizzoli, 12 pp., Int. Environ. Modell. and Software Soc., Burlington, Vermont.
Kumar, M., D. Marks, J. Dozier, M. Reba, and A. Winstral (2013), Evaluation of distributed hydrologic impacts of temperature-index and energy-based snow models, *Adv. Water Resour.*, *56*, 77–89.
Kuusisto, E. (1980), On the values and variability of degree-day melting factor in Finland, *Nord. Hydrol.*, *11*(5), 235–242.
Lehning, M., P. Bartelt, B. Brown, C. Fierz, and P. Satyawali (2002), A physical SNOWPACK model for the Swiss avalanche warning Part II: Snow microstructure, *Cold Reg. Sci. Technol.*, *35*(3), 147–167.
Lehning, M., I. Voelksch, D. Gustafsson, T. A. Nguyen, M. Staehli, and M. Zappa (2006), ALPINE3D: A detailed model of mountain surface processes and its application to snow hydrology, *Hydrol. Processes*, *20*(10), 2111–2128.
Liu, Y., C. D. Peters-Lidard, S. Kumar, J. L. Foster, M. Shaw, Y. Tian, and G. M. Fall (2013), Assimilating satellite-based snow depth and snow cover products for improving snow predictions in Alaska, *Adv. Water Resour.*, *54*, 208–227.
Magnusson, J., D. Gustafsson, F. Husler, and T. Jonas (2014), Assimilation of point SWE data into a distributed snow cover model comparing two contrasting methods, *Water Resour. Res.*, *50*, 7816–7835, doi:10.1002/2014WR015302.
Mahat, V., D. G. Tarboton, and N. P. Molotch (2013), Testing above- and below-canopy representations of turbulent fluxes in an energy balance snowmelt model, *Water Resour. Res.*, *49*, 1107–1122, doi:10.1002/wrcr.20073.
Morin, S., Y. Lejeune, B. Lesaffre, J. M. Panel, D. Poncet, P. David, and M. Sudul (2012), An 18-yr long (1993–2011) snow and meteorological dataset from a mid-altitude mountain site (Col de Porte, France, 1325 m alt.) for driving and evaluating snowpack models, *Earth Syst. Sci. Data*, *4*(1), 13–21.
Rango, A., and J. Martinec (1995), Revisiting the degree-day method for snowmelt computations, *Water Resour. Bull.*, *31*(4), 657–669.
Reba, M. L., D. Marks, T. E. Link, J. Pomeroy, and A. Winstral (2014), Sensitivity of model parameterizations for simulated latent heat flux at the snow surface for complex mountain sites, *Hydrol. Processes*, *28*(3), 868–881.
Rings, J., J. A. Vrugt, G. Schoups, J. A. Huisman, and H. Vereecken (2012), Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments, *Water Resour. Res.*, *48*, W05520, doi:10.1029/2011WR011607.
Rutter, N., et al. (2009), Evaluation of forest snow processes models (SnowMIP2), *J. Geophys. Res.*, *114*, D06111, doi:10.1029/2008JD011063.
Schmucki, E., C. Marty, C. Fierz, and M. Lehning (2014), Evaluation of modelled snow depth and snow water equivalent at three contrasting sites in Switzerland using SNOWPACK simulations driven by different meteorological data input, *Cold Reg. Sci. Technol.*, *99*, 27–37.
Seibert, J. (2001), On the need for benchmarks in hydrological modelling, *Hydrol. Processes*, *15*(6), 1063–1064.
Shrestha, M., L. Wang, T. Koike, Y. Xue, and Y. Hirabayashi (2010), Improving the snow physics of WEB-DHM and its point evaluation at the SnowMIP sites, *Hydrol. Earth Syst. Sci.*, *14*(12), 2577–2594.
Slater, A. G., et al. (2001), The representation of snow in land surface schemes: Results from PILPS 2(d), *J. Hydrometeorol.*, *2*(1), 7–25.
Slater, A. G., and M. P. Clark (2006), Snow data assimilation via an ensemble Kalman filter, *J. Hydrometeorol.*, *7*(3), 478–493.
Stoessel, F., M. Guala, C. Fierz, C. Manes, and M. Lehning (2010), Micrometeorological and morphological observations of surface hoar dynamics on a mountain snow cover, *Water Resour. Res.*, *46*, W04511, doi:10.1029/2009WR008198.
Tobin, C., B. Schaefli, L. Nicotina, S. Simoni, G. Barrenetxea, R. Smith, M. Parlange, and A. Rinaldo (2013), Improving the degree-day method for sub-daily melt simulations with physically-based diurnal variations, *Adv. Water Resour.*, *55*, 149–164.
Tuteja, N. K., and C. Cunnane (1999), A quasi physical snowmelt runoff modelling system for small catchments, *Hydrol. Processes*, *13*(12–13), 1961–1975.
Vionnet, V., E. Brun, S. Morin, A. Boone, S. Faroux, P. Le Moigne, E. Martin, and J. M. Willemet (2012), The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2, *Geosci. Model Dev.*, *5*(3), 773–791.
Walter, M. T., E. S. Brooks, D. K. McCool, L. G. King, M. Molnau, and J. Boll (2005), Process-based snowmelt modeling: Does it require more input data than temperature-index modeling?, *J. Hydrol.*, *300*(1–4), 65–75.
Warscher, M., U. Strasser, G. Kraller, T. Marke, H. Franz, and H. Kunstmann (2013), Performance of complex snow cover descriptions in a distributed hydrological model system: A case study for the high Alpine terrain of the Berchtesgaden Alps, *Water Resour. Res.*, *49*, 2619–2637, doi:10.1002/wrcr.20219.
Wever, N., C. Fierz, C. Mitterer, H. Hirashima, and M. Lehning (2014), Solving Richards equation for snow improves snowpack meltwater runoff estimations in detailed multi-layer snowpack model, *Cryosphere*, *8*(1), 257–274.
Yossef, N. C., L. P. H. van Beek, J. C. J. Kwadijk, and M. F. P. Bierkens (2012), Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes, *Hydrol. Earth Syst. Sci.*, *16*(11), 4233–4246.
Zanotti, F., S. Endrizzi, G. Bertoldi, and R. Rigon (2004), The GEOTOP snow module, *Hydrol. Processes*, *18*(18), 3667–3679.
Zappa, M., F. Pos, U. Strasser, P. Warmerdam, and J. Gurtz (2003), Seasonal water balance of an Alpine catchment as evaluated by different methods for spatially distributed snowmelt modelling, *Nord. Hydrol.*, *34*(3), 179–202.
Zeinivand, H., and F. De Smedt (2009), Hydrological modeling of snow accumulation and melting on river basin scale, *Water Resour. Manage.*, *23*(11), 2271–2287.