AGU PUBLICATIONS

Water Resources Research

COMMENTARY

10.1002/2014WR016585

Special Section:
The 50th Anniversary of Water Resources Research

Key Points:
- Field observations may contain errors undetectable by standard quality control
- Iterative comparisons of models and observations are required to detect errors
- Scientific discourse must be honest about the nonlinear road to success

Correspondence to:
J. D. Lundquist,
jdlund@u.washington.edu

Citation:
Lundquist, J. D., N. E. Wayand, A. Massmann, M. P. Clark, F. Lott, and N. C. Cristea (2015), Diagnosis of insidious data disasters, *Water Resour. Res.*, *51*, 3815–3827, doi:10.1002/2014WR016585.

Received 21 OCT 2014
Accepted 19 MAR 2015
Accepted article online 24 MAR 2015
Published online 15 MAY 2015

Diagnosis of insidious data disasters

Jessica D. Lundquist[1], Nicholas E. Wayand[1], Adam Massmann[1,2], Martyn P. Clark[3], Fred Lott[1,4], and Nicoleta C. Cristea[1]

[1]Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington, USA, [2]Now at Atmospheric Science Department, State University of New York, Albany, New York, USA, [3]National Center for Atmospheric Research, Boulder, Colorado, USA, [4]Now at King County, Seattle, Washington, USA

Abstract Everyone taking field observations has a story of data collection gone wrong, and in most cases, the errors in the data are immediately obvious. A more challenging problem occurs when the errors are insidious, i.e., not readily detectable, and the error-laden data appear useful for model testing and development. We present two case studies, one related to the water balance in the snow-fed Tuolumne River, Sierra Nevada, California, combined with modeling using the Distributed Hydrology Soil Vegetation Model (DHSVM); and one related to the energy balance at Snoqualmie Pass, Washington, combined with modeling using the Structure for Unifying Multiple Modeling Alternatives (SUMMA). In the Tuolumne, modeled streamflow in 1 year was more than twice as large as observed; at Snoqualmie, modeled nighttime surface temperatures were biased by about $+10°C$. Both appeared to be modeling failures, until detective work uncovered observational errors. We conclude with a discussion of what these cases teach us about science in an age of specialized research, when one person collects data, a separate person conducts model simulations, and a computer is charged with data quality assurance.

1. Introduction

Measurements and modeling have gone hand-in-hand since before hydrology began as a formal science. In most early work [e.g., *Mulvany*, 1851; *United States Army Corps of Engineers (US ACE*), 1956], the same person took the measurements and developed the model, and iterated between them until all information collectively made sense. The principal challenges of hydrologic modeling—theoretical assumptions, data scarcity, insufficient computer capacity, and limitations of calibration procedures—were identified by *Freeze* [1978]. Over time, research has become more specialized, and now many people use a model developed by someone else, compare model simulations to data collected by another someone else, and if the two match, accept hypotheses and/or modeling advances, and if they do not match, find it difficult to identify a path forward. The exceptions, where data are collected by those who build and use the models, are largely associated with modeling small catchments [e.g., *Wigmosta and Burges*, 1997; *Western and Grayson*, 2000]. In many cases, the model is calibrated to achieve a match, potentially leading to compensatory errors [e.g., *Kirchner*, 2006; *Kampf and Burges*, 2010], or uncertainty bounds and structure are assumed for both the measurements and the model and its parameters [e.g., *Feyen et al.*, 2007; *Montanari and Koutsoyiannis*, 2012], and both are combined in a data assimilation framework [e.g., *Slater and Clark*, 2006; *Reichle*, 2008]. In some cases, the work is shelved, the apparent failure swept under the rug [*Andréassian et al.*, 2010]. In others, a problem is found and fixed, but only the final working solution is presented to the broader community [*Clark et al.*, 2011]. While mismatches between models and observations are a ubiquitous part of hydrologic science [e.g., *Gupta et al.*, 2008], frank discussion of how such mismatches should be evaluated and reconciled is still needed.

We present two case studies of apparent modeling failures, wherein all efforts at model calibration failed, where traditional data quality-control measures detected no problems, and where only extreme stubbornness and repetitive iteration between modeling and observations led to discovery of the root of the problem. The point of the paper is to exemplify that such odd cases occur. To quote from Arthur Conan Doyle's Sherlock Holmes: "when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth" [*Doyle*, 1890]. Section 2 presents the case studies, and section 3 discusses how we can use these examples to help frame future community efforts and priorities with regards to education, instrumentation, and modeling best practices.

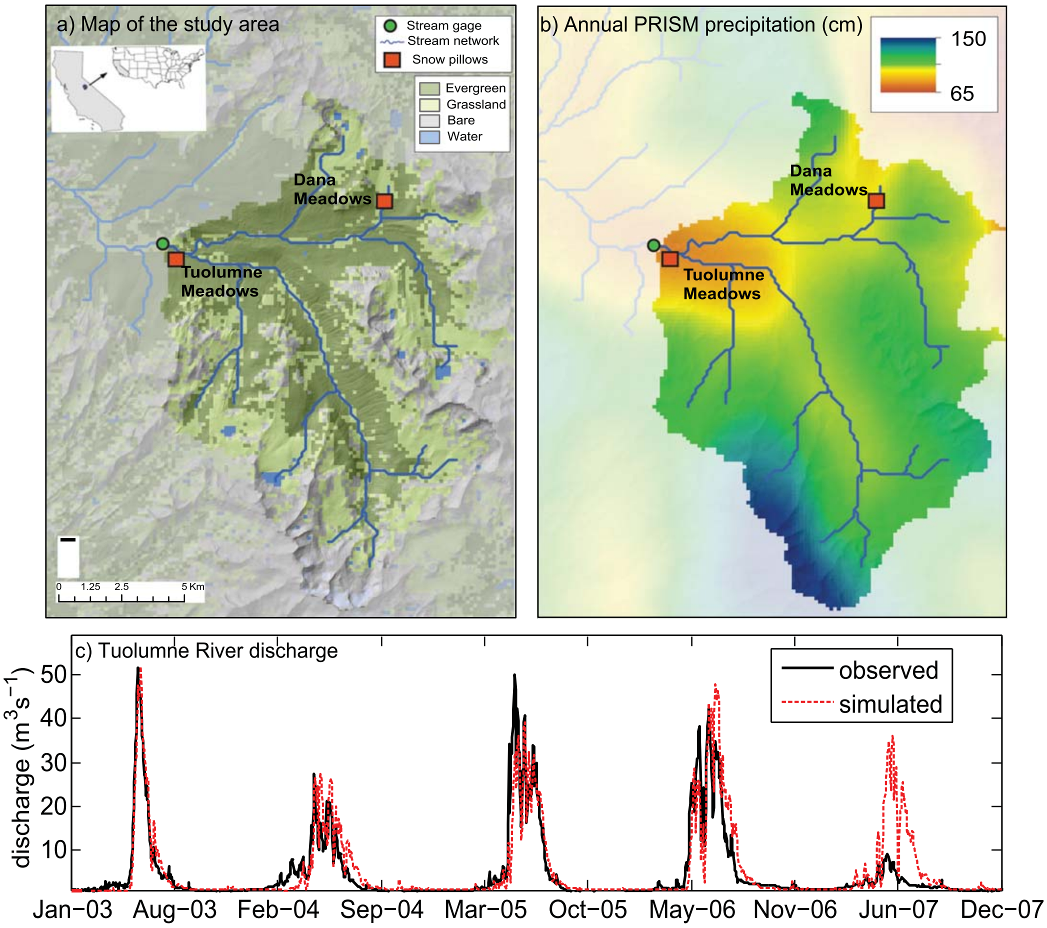

Figure 1. (a) Map of study area and instrument locations. (b) PRISM map, with basin boundaries of average annual precipitation distribution, interpolated to 150 m resolution. (c) Simulated and observed Tuolumne River discharge for the basin above Highway 120. Note: Slide Canyon, shown in Figure 2, is located slightly to the northwest of the map domain shown here.

2. Case Studies (Detective Stories)

2.1. Water Balance: The Tuolumne Watershed and the Dana Meadows Snow Pillow

We set up the Distributed Hydrology Soil Vegetation Model (DHSVM) [*Wigmosta et al*., 1994] to simulate streamflow entering Tuolumne Meadows in Yosemite National Park, California, USA (see *Lowry et al*. [2010, 2011] for a description of project motivation and application). The basin area is 186 km^2 with elevations ranging from 2600 to 4000 m (Figure 1), with over 95% of annual precipitation falling in the winter in the form of snow. Because of known problems with precipitation-gauge undercatch in such snowy conditions [*Rasmussen et al*., 2012], we decided to use the daily snow accumulation on the Dana Meadows snow pillow (Figure 1a) to provide precipitation input to the model. This point-measured solid precipitation was then distributed to each 150 m grid cell in DHSVM according to the Parameter Regression on Independent Slopes Model (PRISM) 800 m climatology maps from 1971 to 2000 [*Daly et al*., 1994, 2008], interpolated to the 150 m scale (Figure 1b). Model parameters were set to literature-reported recommended values (see Appendix in *Cristea et al*. [2013] for list), and model-simulated discharge was compared to measured discharge, as determined from a stage-sensor and rating curve at the bridge where the Tuolumne River passes under Highway 120 (Figure 1c). The model simulated discharge well for the first 4 years of record, but in water year 2007, the total annual discharge was too high by more than a factor of 2 (Figure 1c).

Any adjusting of model parameters to match 2007 resulted in mismatch in the model simulation in all other years, so the only conclusion was that something odd happened in 2007. Many investigators would have excluded 2007 from their analysis, but we had received our first NSF grant to look at interactions between

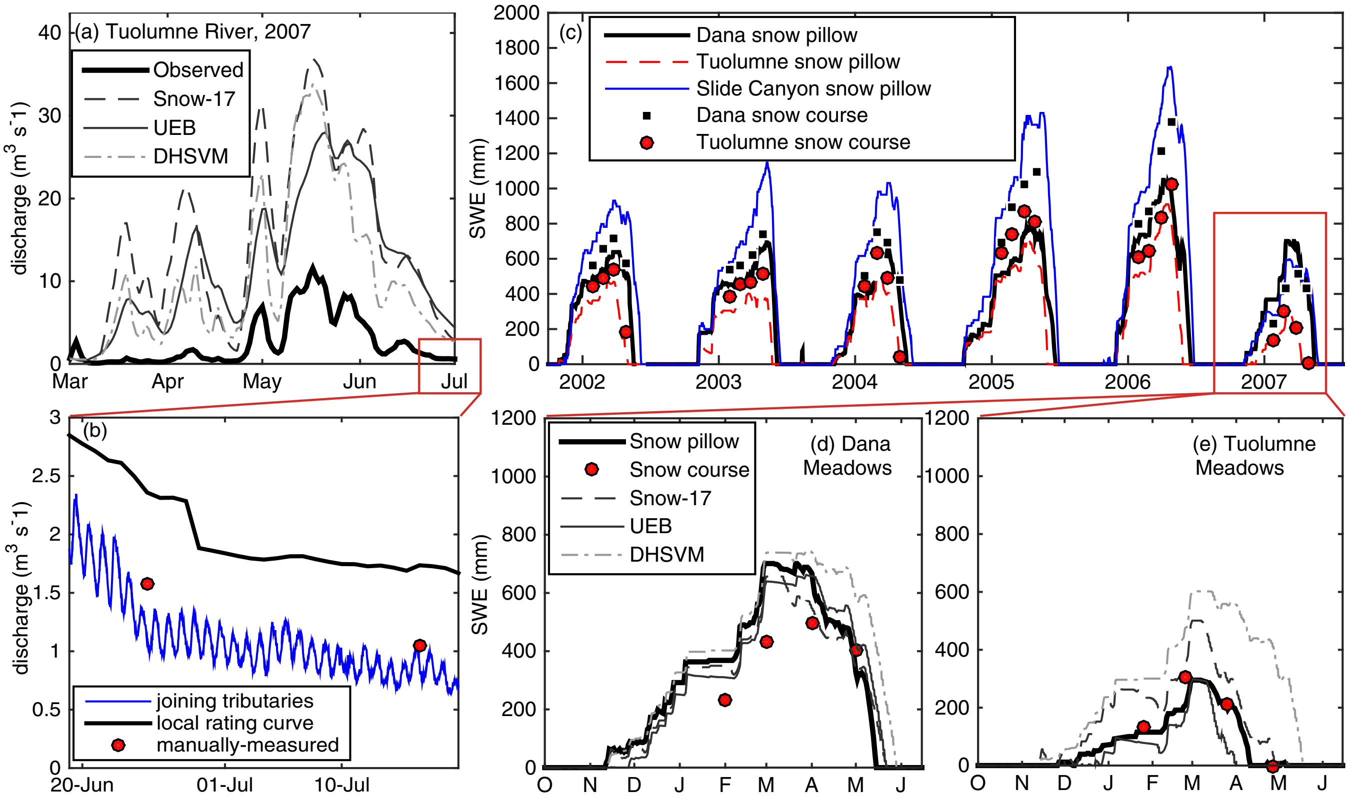

Figure 2. (a) Tuolumne River observed discharge and modeled discharge (from distributed versions of Snow-17, UEB, and DHSVM) for 2007, with period shown in Figure 2b indicated by box. (b) Comparison of observed Tuolumne River discharge estimated from the sum of discharge calculated from rating curves developed at the two tributaries that make up the main river, from the sum of manual wading measurements on these two tributaries, and from the local stage and rating curve at the Tuolumne River bridge at Highway 120. The observed discharge in Figure 2a is the same as the local rating curve method in Figure 2b. (c) Snow water equivalent (SWE) from three snow pillow stations in the area (locations shown in Figure 1a), along with colocated snow course measurements at Dana Meadows and Tuolumne Meadows from 2002 to 2007. Box indicates time period shown in more detail in Figures 2d and 2e. (d and e) Observed and simulated SWE for (d) Dana Meadows and (e) Tuolumne Meadows snow pillow locations for 2007 only.

snowmelt runoff, meadow groundwater, and meadow vegetation communities, and all of the meadow measurements began in 2007. This was the critical year to "get right."

Our first thought was that the observed discharge was incorrect that year. We had developed the rating curve ourselves and knew that a shifting of the stage recorder location or calibration, or a change to the channel cross-section geometry, could make our rating curve obsolete. We checked the manual discharge measurements for 2007 (Figures 2a and 2b), as well as measurements of stage and streamflow in the tributaries to the main Tuolumne River and water level measurements in a number of groundwater wells that were adjacent to the stream [*Loheide and Lundquist*, 2009]. All showed that the rating-curve-stage-based discharge was not bad (if anything, too high, Figure 2b)—the flows really were that low in 2007.

Our second thought was that the problem was with our chosen model, DHSVM. We were relatively new to modeling, and perhaps, for reasons unknown to us, DHSVM just failed in really dry years. We decided to implement two other snow models over the basin: the Utah Energy Balance, UEB, model [*Tarboton and Luce*, 1996] and Snow-17 [*Anderson*, 1976], with model implementations described in *Lott* [2008]. All three models simulated snow accumulation and ablation well in 2007 at the Dana Meadows snow pillow (Figure 2d), the site where we obtained the driving meteorological data, including the precipitation input to the model based on snow pillow increases. The three had notably different results at the Tuolumne Meadows snow pillow (Figure 2e), which was an independent observation not used in model forcing. The Tuolumne site is about 400 m lower in elevation than the Dana site, and model differences primarily arose due to different representations of midwinter melt events, which were pronounced in UEB and Snow-17 but did not occur in DHSVM or in the observations. All three models represented more snow accumulation than was observed at the site. Over the entire watershed, all three models overestimated 2007 discharge by about a factor of 2 (Figure 2a), despite reasonable fits with streamflow observations in most other years, so we could not blame DHSVM.

We checked the Dana snow pillow against manual snow course measurements and against other nearby stations and snow courses (Figure 2c). The Slide Canyon snow pillow and the Dana snow course, which usually recorded more snow than the Dana snow pillow, both recorded less snow than Dana in 2007, but no obvious problems were apparent.

Our problem was with the mass balance of our watershed, so given that the measured discharge and the snowfall input looked reasonable, we thought that evapotranspiration (ET) might have been drastically higher than expected in 2007. It was, after all, the driest year we had observed. The study of how ET varies in the Sierra between wet and dry years became its own paper [*Lundquist and Loheide*, 2011], and the answer was that ET variations were relatively small and could not possibly explain the factor of 2 difference we were worried about.

We next explored the possibility that shifting precipitation patterns in 2007 made the Dana Meadows snow pillow unrepresentative of basin-wide precipitation. In addition to 2007 being a very dry year, the storms had come primarily from the northwest, in contrast to the usual southwesterly flow during California snowfall [*Lott*, 2008]. California-Nevada River Forecasting Center volume forecast errors for the neighboring Merced River Basin were 40–90% higher than average in 2007 [*California-Nevada River Forecast Center*, 2007], raising the possibility that these different storm tracks may have made PRISM gridded precipitation, though relatively correct in most years, incorrect for 2007. Comparison of snow pillow records between Dana and other neighboring snow pillows (Figure 2c) highlighted that the relative amount of snow measured at Dana compared to its neighbors differed drastically in 2007. For example, Dana SWE was usually about half of that at Slide Canyon and about 20–25% more than that at Tuolumne (Figure 2c), but in 2007, Dana recorded more SWE than Slide Canyon and over twice as much SWE as Tuolumne. This observation led to a California-wide study of annual variations in spatial precipitation patterns [*Lundquist et al*., 2010], and while 2007 definitely had Sierra-wide precipitation patterns that did not match the patterns used in PRISM, this difference alone could not explain a factor of 2 error in the Tuolumne water balance.

Our attention shifted to exploring if something was wrong with the Dana snow pillow. We started contacting all of the field personnel we knew who worked on the snow pillow, and on 10 December 2008, Alan Flint (from the USGS) sent a photo he had taken on 16 October 2007—showing what appeared to be a small lodgepole pine growing out of the snow pillow. To appreciate the significance of this, one needs knowledge of how a snow pillow works. At most California snow sites, each snow pillow consists of four metal bladders filled with antifreeze (Figure 3a). These are hydraulically connected to each other with pipes that connect to a vertical reservoir of antifreeze with a float-activated recorder. As pressure on the metal bladders changes (generally from the weight of overlying snow), the float raises and lowers, and this turns a wheel attached to a data logger that relays the information back to the snow survey main office. To account for any potential linear drift in the system, the measured height at the end of each summer is set to represent 0 SWE because at that time no snow is present. We visited the snow pillow in June 2009 (as soon as we could upon hearing this information) and found and removed the tree (Figures 3c–3e). The tree was growing in between two of the metal plates, with substantial underground roots, and appeared to be applying pressure on the overall system (Figure 3b). The gradual increasing pressure of the tree had presumably been zeroed out during annual remote calibration, and no one had registered that this small pine could be a problem. In retrospect, the tree is identifiable in a photo (taken by F. Lott on 7 August 2007) in *Lott* [2008, Figure A.2.b], and despite our substantial desperation, no one noticed it. In summer 2009, we switched to using the Tuolumne Meadows precipitation gage as the basis for our basin-wide precipitation distributions, and the model performed with reasonable success from 2003 to 2009 (as demonstrated in *Cristea et al*. [2013]).

2.2. Energy Balance: Snoqualmie Pass and Longwave Irradiance

After our adventures in Tuolumne, we decided that we should study snow processes closer to home, at a site we could drive to regularly and check for trees or other problems before they disrupted years of our lives. In fall 2012, we upgraded the Washington Department of Transportation (WA DOT) Snoqualmie Pass avalanche snow measurement site (Figure 4) with new instrumentation to capture all aspects of the snow energy balance in a region that receives as much winter rain as it does snow. The goal was to provide

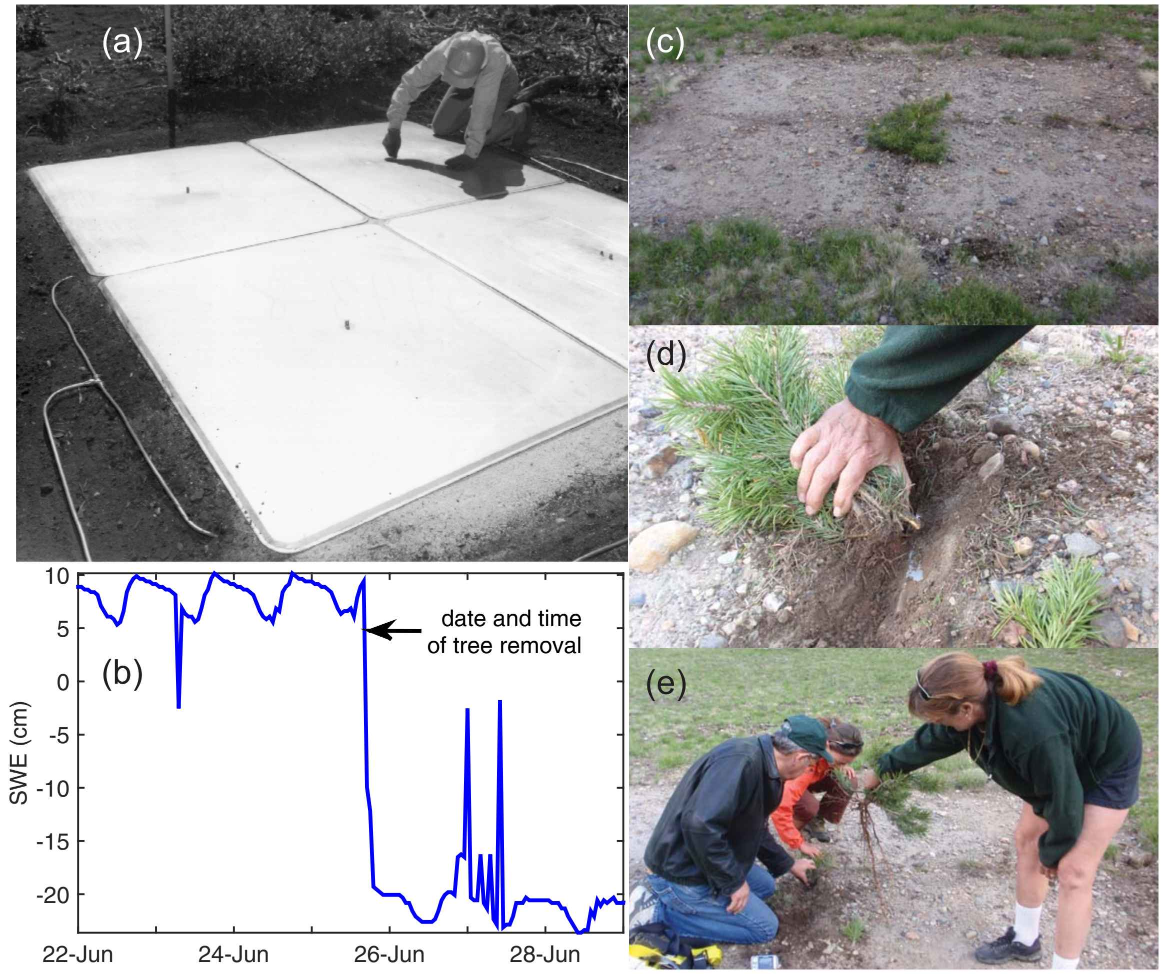

Figure 3. (a) Photo of a snow pillow installation (from California DWR). (b) Real-time record of SWE at Dana Meadows from the California Data Exchange Center (CDEC) on 22–28 June 2009. (c–e) Sequence of photos of the tree and its removal from the snow pillow; activities took place about 5 pm local time on 25 June 2009.

comprehensive surface observations for testing how well a new modular snow modeling framework being developed at NCAR (SUMMA) [*Clark et al.,* 2015a, 2015b] could work in a warm, mixed rain-snow environment.

After collecting and modeling our first water year (2013), the modeled snow melted too quickly compared to observations. The most glaring model-observational differences occurred on cold, clear, calm nights, when the observed snow surface temperature was about 10°C colder than the modeled surface temperature (Figure 5). This difference led to the model surface temperature reaching 0°C each day, resulting in melt, whereas the colder observed temperatures did not rise high enough to melt (Figure 5). Our first assumption was that the model needed fixing, but testing various model options and parameters for both turbulent fluxes [*Anderson*, 1976; *Louis*, 1979; *Mahrt*, 1987] and thermal conductivity [*Yen*, 1965; *Mellor*, 1977; *Jordan*, 1991; *Smirnova et al.*, 2000] led to little or no change and certainly did not solve our surface temperature problem. So we turned to theoretical arguments: (1) it was night, so there was no shortwave irradiance; (2) it was clear, so there was no precipitation; (3) no wind was measured, so the sensible and latent heat fluxes were likely very small. The change in surface temperature at this time had to be controlled by the net incoming and outgoing longwave (LW) irradiance, which was not sufficiently negative to cause the observed cooling.

We had purchased a Kipp and Zonen CNR4 sensor from Campbell Scientific, which provides 4-stream radiation data, based on both upward and downward-pointing longwave and shortwave radiometers, all conveniently mounted together on one boom (Figure 4a). We regretted this mounting configuration as soon as

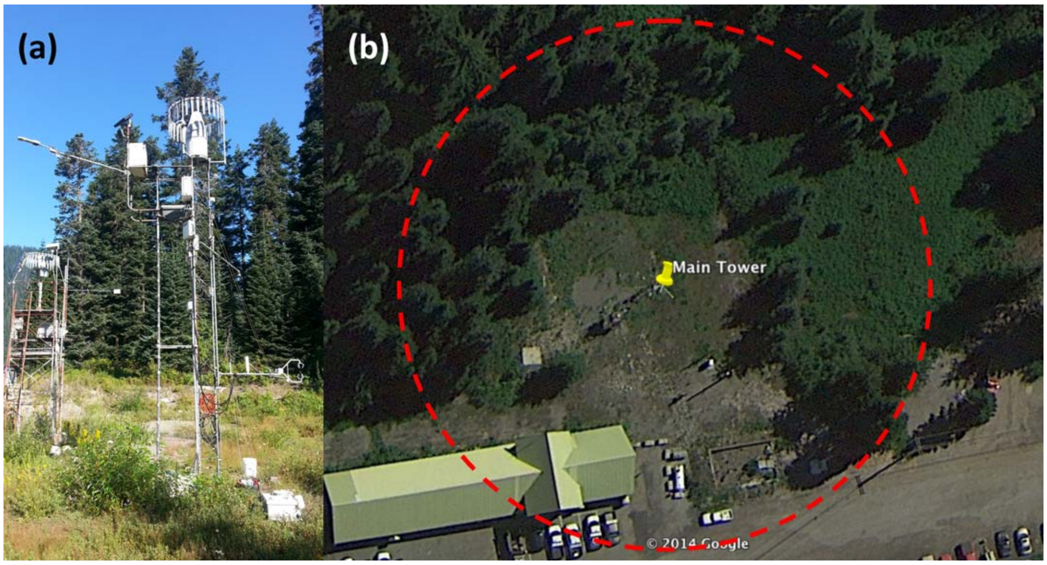

Figure 4. (a) Photo of energy balance station. (b) Google Earth (copyright 2014) image of the site, with thumbtack representing the tower location, and red dashed circle showing the calculated viewshed of the downward-looking longwave radiometer (which would get smaller as snow depth increases).

we tried to put it on our tower. Our study plot is relatively small, with surrounding trees and buildings (Figure 4b), so the very wide-angle viewshed (150° for upwelling, 180° for downwelling) of the pyrgeometers was problematic. If we put the instrument high on the tower, the upward-pointing sensors would have an unobstructed view of the sky, but the downward-pointing sensors would be influenced by some trees, building parts, and parking lot space (Figure 4b, see thumbtack and circle). With this in mind, we decided to mount the sensor high and use the CNR4 LW only for incoming irradiance and to use an Apogee SI-111 Infrared Radiometer (mounted on the same boom but with a smaller view angle, standard field of view, of 22°) to determine snow surface temperature and outgoing LW (assuming an emissivity of 0.99). The downward-pointing LW on the CNR4 did not agree with the Apogee IR measurements. In particular, during known melting periods, the Apogee sensor reported 0°C, while the CNR4 reported a snow surface temperature exceeding 0°C. We attributed this first to the warmer building and parking lot (Figure 4b) but then calculated the percentage of the entire viewing area that these contributed, and it was too small to account for the total difference. We began to doubt the accuracy of our brand-new sensor. Could it have been calibrated incorrectly? If our measurements reported higher incoming LW than really occurred, our model-measurement discrepancies and our intermeasurement discrepancies could be explained.

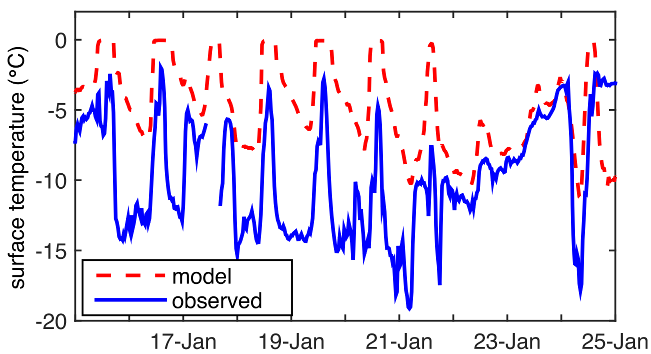

Figure 5. Observed (using an Apogee point IR sensor) and modeled snow surface temperature. Various turbulent flux and thermal conductivity parameterizations (detailed in text) resulted in variations so small that they overlaid the model line on the scale of this graph.

We began sending out queries. After considerable discussion with a senior radiometry expert at NOAA about how radiometers could and should be calibrated, we were told, "By the way, agreement or not with a model does not seem to me to be a good way to choose the correctness of a measurement." Despite this warning, we could not come up with a better explanation and continued to doubt our measurement. Fortunately, another laboratory at the university (see acknowledgements) had been using longwave radiometers much longer and more intensively than

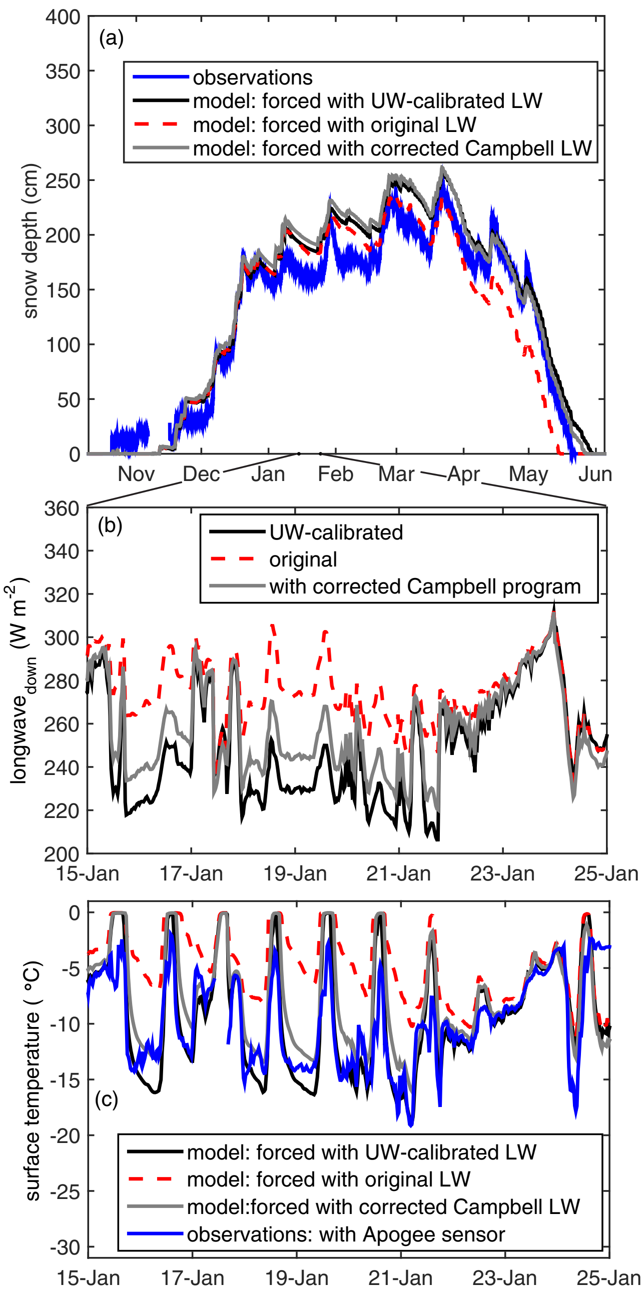

Figure 6. (a) Observed snow depth and modeled snow depth using three different sources of incoming longwave irradiance. (b) Observed incoming LW as determined by original and corrected calibration coefficients (based both on UW calibration and the Campbell programming correction). (c) Modeled surface temperature compared to that observed by the Apogee sensor for the three incoming LW series shown in Figure 6b.

we had, and they kindly let us use their blackbody chamber for calibration. Based on our local calibration, the coefficients were not correct. We sent this information to Kipp and Zonen, inquiring if they thought that the original calibration might be in error. To this inquiry, they informed us that a CNR4 pyrgeometer is "calibrated against a hot plate and not outdoors" and that "a hot plate does not simulate the sky properly." They reminded us that, in relation to this, the manual clearly explains that daily uncertainty in the measurement is 10%. Although 10% could explain quite large errors (~30 W m^{-2}), our differences were larger than this (Figure 6b).

Within 1 week of this e-mail exchange (approximately 1 year after we purchased our sensors), we received an e-mail announcement from Campbell Scientific that read, "We recently discovered the inaccuracy of the programming examples in the CNR4 manual for the CR1000, CR3000, and CR5000, as well as the inaccuracy of the programs created by Short Cut build… [that] apply incorrect multipliers to the sensors being measured giving erroneous values." Indeed, our Campbell data logger program followed these incorrect examples, and when we implemented the recommended corrections, the LW appeared much closer to that we had calibrated ourselves (Figure 6). Both the self-calibrated LW and the program-corrected LW allowed the model to produce plausible results (Figure 6c), and we finally felt confident enough to publish our data set [*Lundquist et al.*, 2014] and continue with our model evaluation.

3. Discussion

As we look forward to the next 50 years of hydrologic science, we anticipate increased data collection, storage, and availability [e.g., *Burges*, 2011; *Killeen and Blatecky*, 2011; *Hey*, 2012]. We also expect higher-resolution modeling and increased data assimilation [e.g., *Wood et al.*, 2011]. The examples presented here at first seem outlandish, but in conference presentations, the majority of the audience nods with knowing understanding, with wry smiles and comments of "I've been there too." Unconventional data errors are likely more common than has been documented. They elude detection by a standard computerized QC analysis [e.g., *Meek and Hatfield*, 1994], but have the potential to hinder scientific advancement if not taken seriously [e.g., *Kampf and Burges*, 2010, *Beven et al*., 2011; *Beven and Westerberg*, 2011; *Kauffeldt et al*., 2013].

We have fast computers and high-speed connectivity, but we are in danger of losing the essential skills that came as second nature to scientists in the past [e.g., *Dunne and Black*, 1970a, 1970b]. Because they had to draw graphs by hand and write or type most every field number, they had time (perhaps forced time) to think about where each number came from and to think about its validity. The challenge now remains: can we train a computer to do this? Can we employ enough people to do this? If we do not have to plot numbers by hand, how can we efficiently combine computer-based methods with human intuition to catch hard-to-anticipate errors?

The hydrologic modeling community has long discussed ways to extract enough relevant information from data to identify and correct model errors and misconceptions [e.g., *Gupta et al*., 2008]. The "Court of Miracles of Hydrology" meeting and special issue [*Andréassian et al*., 2010] was one recent effort to highlight lessons to be learned from failures. Many authors discussed cases where a mismatch between a catchment and a model turned out to identify a catchment process that was not represented in the model (e.g., karst flow [*Goswami and O'Connor*, 2010]; or geological heterogeneity [*Refsgaard and Hansen*, 2010]), and thus a "hydrological monster" provided a way to improve hydrologic science. Many similar insights have also been detailed in earlier work (e.g., lack of observations, scaling issues) [*Horton*, 1931]; issues with the unit hydrograph and groundwater separation [*Hoyt et al*., 1936]; many examples of mismatches [*Meinzer*, 1949; *Langbein and Wells*, 1955]; development of the Stanford Watershed model [*Crawford and Linsley*, 1966]; assumptions in choice of an aquifer model [*Theis*, 1967]; scaling issues, stationarity, need for adequate and accurate data [*Philip*, 1975]; scaling issues and spatial variability [*Seyfried and Wilcox*, 1995]; and the importance of systematic iteration between models and data [*James and Burges*, 1982, Figure 11.2]).

However, while these issues date back to the beginning of hydrologic science, no clear, straight-forward, universal solutions have emerged. The ad hoc detective work described in our case studies was neither systematic nor time-efficient and would be impractical for someone modeling over larger scales. Another group of researchers would likely have followed a different path of investigation and may have arrived at different conclusions. The sections below highlight general guidelines, drawn from the greater literature as well as these case-study examples, which could help detect similar errors in new situations.

3.1. Time-Invariant Parameters

Many authors have recommended looking at key hydrologic signatures, such as the runoff ratio (discharge divided by precipitation), to diagnose model divergence from observations [e.g., *Yilmaz et al*., 2008]. *Gupta et al*. [1998] recommend testing whether a model can represent observations using a temporally invariant parameter set. These criteria would readily have identified the problem with our Tuolumne water balance, and these criteria have been successfully used in other studies. For example, forcing data sets were problematic in both the first and second NOAA Distributed Model Intercomparison Projects (DMIP and DMIP-2). In the first, *Yilmaz et al*. [2008] found that their model consistently overestimated observed total streamflow, with particular overestimation in years with very low flows. They found that the potential evapotranspiration (PET) data provided in the DMIP data set were constant every year, whereas PET values from an alternative data set were consistently larger, varied between years, and resulted in simulations more closely matching observations. The gridded precipitation data set for the Sierra Nevada in DMIP-2 led to even larger problems, with loss of a key precipitation station in 2003 resulting in model discharge errors differing in sign before and after that year [*Mizukami and Smith*, 2012]. Anecdotally, B. McGurk et al. (personal communication 2014) recently told us that he could not calibrate the PRMS model for the upper Tuolumne watershed using the same set of parameters before 2007 and after 2007 (although he could make it work with two different sets of parameters). His first assumption was that the methodology for the streamflow rating curve had changed, but we quickly concluded that we had inadvertently given him our forcing time series derived from the Dana snow pillow with the tree some time ago and then neglected to inform him of the essential necessary corrections. Because most data errors develop over time or in response to specific events, the temporal consistency of a model is critical to identifying problems.

3.2. Multicriteria Modeling—Multiple-Independent Observations

There have been numerous recommendations for model evaluation to examine matches to multiple criteria, rather than the single-criterion of "goodness-of-fit" to streamflow [e.g., *Gupta et al*., 2008]. Such criteria could incorporate signature patterns in the data, such as recession characteristics or covariance structures, or could include the use of qualitative observations, e.g., "soft data" [*Seibert and McDonnell*, 2002], to update conceptual understanding and revise the model. Similarly, observations need to be evaluated with multiple criteria and with comparison to other, independently obtained observations [e.g., *Gandin*, 1988; *Steinacker et al*., 2011]. For example, if we had placed two CNR4 radiometers on our Snoqualmie tower, they would have both been subjected to the same data logger programming error and would have produced nearly identical incorrect values, while the comparison of the CNR4 to the Apogee infrared sensor clearly identified a problem. Multiple precipitation gauges of the same type are likely subject to the same gauge undercatch [e.g., *McMillan et al*., 2012], but skilled implementation of different types of gauges [e.g., *Sieck et al*., 2007; *Kampf and Burges*, 2010] may overcome this limitation. Similarly, manual snow course measurements near a snow pillow, as at Dana, indicated that something about the snow pillow might have changed.

3.3. The Human Dimension: Attention, Experience, Understanding, and the Ability to Question Patterns With New Eyes

Holländer et al. [2014] investigated how a group of modelers using different models responded to varying levels of information about a particularly hard-to-model catchment. They found that the human modeler himself/herself, including "soft data" entailing a modeler visit to the catchment to better conceptualize it, was more important in accurately modeling the catchment than were data sets designed to allow proper matching of multiple, physically based model parameters. The implication is that investing in the human modelers' education and time devoted to a project is very valuable and cannot be replaced by better computers, better optimization algorithms, or better observational data sets. Value judgments, particularly with regards to what is an acceptable observation or an acceptable model representation in a given system, must be made by humans, who will always be shaped by individual past experience and by present attention to the problem at hand.

3.4. Reversing the Conception That "Failure to Match a Model Is no Reason to Doubt the Validity of an Observation"

Most prior studies focus on diagnostic model evaluation and on failure to represent key watershed processes. Our case studies differ in that our first assumption, that the model must be wrong or missing key physical processes, was not true. Rather, key data streams (snowfall and incoming longwave irradiance) were hopelessly in error. We found modeling to play an important role in helping to identify data errors. This is in the full spirit of the use of the earliest computer models [e.g., *Crawford and Linsley*, 1966] to identify likely measurement errors or to show the consequences of an increase or decrease of precipitation on hydrologic response. Rather than quickly assuming that disagreement with a model is no way to identify a data error (as we were told during our Snoqualmie Pass adventure), care should be taken before discarding either an uncertain model or uncertain data [e.g., *Montanari and Di Baldassarre*, 2013]. Beyond simply considering stochastic uncertainties in the data, diagnostic model evaluation should be accompanied with diagnostic data evaluation, and the true errors inherent in data may require quite complicated functions to quantify. The pendulum of popularity seems to swing between overly trusting models to overly trusting data, when the truth will always be somewhere in between.

3.5. The Story Behind the Scenery—Open-Access Meta-Data and Discourse on How Science Is Actually Done

While many pages of scientific discourse have been published on best-practices for incorporating potentially error-ridden data in modeling [e.g., *Kuczera*, 1996; *Vrugt et al*., 2008; *Beven et al*., 2011; *Montanari and Di Baldassarre*, 2013], much less has been written about how instruments are actually installed, maintained, and quality-controlled, likely because technicians are paid to fix problems rather than write about them. While key papers exist on typical errors in discharge and precipitation [e.g., *Sieck et al*., 2007; *Rasmussen et al*., 2012; *McMillan et al*., 2012], similar information on measurements of the energy balance are scarce [e.g., *Augustine et al*., 2000].

The hydrologic community is currently advancing on these fronts; for example, the new data policies for NSF grants and for all AGU journals (including WRR) require that all data and model code related to an investigation be made publicly available. The standards for such publicly available data are still being refined, but to ensure high data quality, systematic recording and publication of metadata will be essential, particularly with regards to instrument types, maintenance, and replacements. Such metadata could be linked to documents describing the mechanics and underlying physical principles of the instruments deployed. In our examples, the snow pillow responded to lateral pressure (from the tree) in addition to the vertical pressure (weight) of snow, and the standard factory calibration procedures differed between radiometer types and manufacturers—the Apogee sensor was calibrated against a range of temperatures (>5 points) in the factory [*Apogee*, 2014], while the Kipp and Zonen CNR4 used a 1 point calibration with an assumed 0 intercept.

As people publish data, they need to also publish the best possible information about uncertainties and potential errors associated with such data, and how specific data collected compares to accepted measurement standards (e.g., those published by the World Meteorological Organization (WMO)'s Commission for Instruments and Methods of Observation (CIMA)) [*World Meteorological Organization (WMO)*, 2008] when they exist. In cases when standards do not exist (such as new instrumentation or technology), the leadership of organizations such as the U.S. Geological Survey (USGS) in developing and testing new devices provides an excellent standard. For example, USGS's development of protocols and field training for use of the now ubiquitous Acoustic Doppler Current Profilers (ADCPs) for streamflow measurement [e.g., *Oberg and Mueller*, 2007] provides one template of how the community can proceed.

More generally, scientific discourse needs to be more honest about the often nonlinear rocky road to research success. In the words of *Medawar* [1979], "the scientific paper may be a fraud because it misrepresents the process of thought that accompanied or gave rise to the work that is described in the paper." Honesty is critical so that beginning scientists know that adventures such as those described here are normal rather than a sign of personal failure. As the supplemental material for papers is expanded to contain data and metadata, it could also contain anecdotal stories of how the authors actually came to write the final paper. For example, the BBC's *Planet Earth* videos each contain brief follow-on vignettes of how the videographers actually captured the animal action in the final film, revealing a career that is often more challenging, less glamorous, and requiring much greater patience than one might originally assume. Current scientific writing style demands that the published paper be written as if the final hypotheses tested and methodology followed were known from the start, when most investigators discount multiple methodologies and disprove multiple hypotheses along the way, or perhaps find some of the actual published methods so difficult and/or time consuming that they plan to never follow that methodology again. These are the stories that are frequently exchanged with laughter over appropriate beverages after scientific meetings, but a key point here is that beginning investigators, and particularly under-represented minorities, are very seldom invited to those tables. Public exposure of even a subset of such stories could help beginning scientists avoid repeating many of the mistakes well trodden by more senior scientists.

3.6. More, More, More

The current trend is to measure more environmental variables, both in situ and via satellite and aerial remote sensing, at increasingly high spatial and temporal resolutions and for a larger number of catchments. This trend requires "balancing depth with breadth" [*Gupta et al*., 2014], which presents key challenges on how to synthesize the idiosyncrasies in individual catchments (both processes and data) and how we can focus on advancing representation of dominant hydrologic processes across larger areas. The examples here beg the question, how will we ensure quality control of this huge amount of data when we already struggle with a single point measurement station? The first five points help identify how we might begin to tackle this question.

A closely related question is whether more is necessarily better. Quantity is a much easier metric to assess than quality. In academic performance evaluations, it takes considerable effort to determine the meaning and contribution of a body of work to the field and is much more straight-forward to count numbers of publications and numbers of citations. Similarly, counting numbers or density of weather or stream stations are easier than assessing the quality of any one of those stations. Fortunately, as station density increases, the correlation between stations also increases, and groups of stations can be used collectively for quality control. Such techniques, many of which are already well developed [e.g., *Steinacker et al*., 2011; *Pielke et al*., 2007, and references therein], can be incorporated into "big data" handling strategies to flag measurements that do not match our expectations of local or regional coherence. Networks of sensors can also be designed to have individual components sensitive to different types of errors (as discussed in point two), and inescapable correlated errors, such as from a remote sensing platform subject to the same atmospheric biases, can be quantified. Thus, the construction of data network error models can help inform computers on how to maximize the information content through the entire network across time and space. These issues are actively being discussed by the data assimilation community (e.g., the ESA Workshop on Correlated Observation Errors in Data Assimilation, April 2014 at the University of Reading), but more work needs to be done to determine which techniques work best and to incorporate them into hydrologic research and practice.

As the pendulum of popularity has swung toward "more is better," it has also swung toward "new is better." For many principal investigators, it is easier to fund new instrumentation than to maintain an existing instrument site. The government agencies that traditionally maintain long-term observations have unstable funding; for example, the number of USGS streamgages dropped over the time period from 1968 to 1981 and only partially recovered after increased funding in 2001 [*USGS*, 2009]. Funding for the National Resource Conservation Service (NRCS) SNOTEL and snow survey program has also been uncertain [e.g., *Udall press release*, 2013]. Research grants are motivated by new ideas and new, innovative measurements—maintenance of an existing field site must be tied to continuing new innovative use of such data every 3 years. Such trends result in the loss of long-term, high-quality measurements. Hydrologic understanding, modeling, and prediction depend on the quality of our measurements. Therefore, as a community, we need to advocate for instrument maintenance and quality control. While technology can certainly guide and aid such efforts, it cannot replace site visits, instrument recalibration, and critical thinking in the face of new observations and discrepancies between multiple sources of information (models, different data streams, and conceptual ideas of how a watershed actually works).

4. Looking Forward to the Next 50 Years

To summarize, we must proceed with humility and patience as we continue to tackle issues that have plagued hydrology since the science began, taking advantage of ever-increasing data, modeling, and communication capabilities without letting the sheer magnitude of information to overwhelm our ability to see clearly. We must carefully and continuously evaluate both observations and models, through personal understanding, quantitative testing, soft data, multiple diagnostic signatures, and theoretical expectations [*Seibert and McDonnell*, 2002; *Gupta et al*., 2008; *Clark et al*., 2011]. We must maintain funding for the more mundane tasks of instrument maintenance and basic education while also pioneering new measurement techniques and new ways to evaluate and synthesize ever-increasing information streams. Finally, as we have tried to exemplify in this paper, we must maintain a sense of humor, honesty, and open discourse throughout our efforts, since science is as much about the process and the questions as about the final epiphany.

Acknowledgments
The Dana Meadows and other snow pillow and snow course historic data are available at http://cdec.water.ca.gov/cgi-progs/selectQuery (enter station ID "DAN" for Dana, "TUM" for Tuolumne, and "SLI" for Slide Canyon). All codes for running DHSVM are available on GitHub: https://github.com/UW-Hydro/DHSVM. The correct Snoqualmie Pass data are available in the University of Washington Research Works Archive at: http://hdl.handle.net/1773/25611. Links to the latest version of the SUMMA model can be found at http://www.ral.ucar.edu/projects/summa/. This research was supported by the National Science Foundation (NSF) under EAR-1215771, EAR-1215809, and CBET-0729830. Any opinions, findings, conclusions, or recommendation expressed here are those of the authors and do not necessarily reflect the views of NSF. We thank Frank Gehrke and the California Department of Water Resources (CA DWR) Cooperative Snow Surveys team for their help with understanding and maintaining snow pillows and their associated instrumentation, and John Stimberis and the Washington Department of Transportation Avalanche team for help with maintaining the Snoqualmie Pass measurement site. We thank Alan and Lorrie Flint for telling us about the tree in the snow pillow, and both the Flints and Courtney Moore for helping us to pull it out. We thank the CA DWR team for not scolding us for performing such unscheduled maintenance, and our collaborators at the "other UW" (University of Wisconsin, Madison), Steve Loheide and Chris Lowry (now at State University of New York, Buffalo), for their incredible patience as we sorted out the modeling that was simply a necessary boundary condition to their meadow studies. We thank Andy Jessup, Dan Clark, Chris Chickadel, and the Air-Sea Interaction and Remote Sensing group of the Applied Physics Laboratory (APL) at the University of Washington for their generous sharing of their longwave sensor calibration facilities. Finally, we thank two anonymous reviewers in addition to Jan Seibert, Thorsten Wagener, Steve Burges, and the Mountain Hydrology Research Group for their thoughtful comments and edits on the manuscript.

References

Anderson, E. A. (1976), A point energy and mass balance model of a snow cover, *NOAA Tech. Rep. NWS 19*, U.S. Department of Commerce, 150 pp., Silver Spring, Md.

Andréassian, V., C. Perrin, E. Parent, and A. Bárdossy (2010), Editorial—The court of miracles of hydrology: Can failure stories contribute to hydrological science?, *Hydrol. Sci. J.*, *55*(6), 849–856.

Apogee (2014), Apogee Instruments Owner's Manual: Infrared Radiometer. [Available at http://www.apogeeinstruments.com/content/SI-100manual.pdf.]

Augustine, J. A., J. J. DeLuisi, and C. N. Long (2000), SURFRAD—A national surface radiation budget network for atmospheric research, *Bull. Am. Meteorol. Soc.*, *81*(10), 2341–2357, doi:10.1175/1520-0477(2000)081<2341:SANSRB>2.3.CO;2.

Beven, K., and I. Westerberg (2011), On red herrings and real herrings: Disinformation and information in hydrological inference, *Hydrol. Processes*, *25*(10), 1676–1680.

Beven, K., P. J. Smith, and A. Wood (2011), On the colour and spin of epistemic error (and what we might do about it), *Hydrol. Earth Syst. Sci. Discuss.*, *8*(3), 5355–5386.

Burges, S. J. (2011), Invited perspective: Why I am an optimist, *Water Resour. Res.*, *47*, W00H11, doi:10.1029/2010WR009984.

California-Nevada River Forecast Center (2007), April-July Volume Water Supply Forecasts, 2007 Verification, technical report. California-Nevada River Forecast Center, Sacramento, Calif. [Available at http://www.cnrfc.noaa.gov/products/water_supply/2007/2007_verification.pdf.]

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.

Clark, M. P., et al. (2015a), A unified approach to process-based hydrologic modeling: 1. Modeling concept, *Water Resour. Res.*, *51*, doi:10.1002/2015WR017198.

Clark, M. P., et al. (2015b), A unified approach for process-based hydrologic modeling: 2. Model implementation and example applications, *Water Resour. Res.*, *51*, doi:10.1002/2015WR017200.

Crawford, N. H., and R. K. Linsley (1966), Digital simulation in hydrology: Stanford watershed model IV, *Tech. Rep. 39*, 210 pp., Dep. of Civ. Eng., Stanford Univ., Stanford, Calif.

Cristea, N. C., J. D. Lundquist, S. P. Loheide, C. S. Lowry, C. E. Moore, and M. S. Raleigh (2013), Modeling how vegetation cover affects climate change impacts on streamflow timing and magnitude in the snowmelt-dominated upper Tuolumne Basin, Sierra Nevada, *Hydrol. Processes*, *28*, 3896–3918, doi:10.1002/hyp.9909.

Daly, C., R. P. Neilson, and D. L. Phillips (1994), A statistical-topographic model for mapping climatological precipitation over mountainous terrain, *J. Appl. Meteorol.*, *33*, 140–158, doi:10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. P. Pasteris (2008), Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States, *Int. J. Climatol.*, *28*, 2031–2064, doi:10.1002/joc.1688.

Doyle, A. C. (1890), *The Sign of the Four*, 100 pp., Spencer Blackett, London, U. K.

Dunne, T., and R. D. Black (1970a), An experimental investigation of runoff production in permeable soils, *Water Resour. Res.*, *6*, 478–490, doi:10.1029/WR006i002p00478.

Dunne, T., and R. D. Black (1970b), Partial area contributions to storm runoff in a small New England watershed, *Water Resour. Res.*, *6*, 1296–1311, doi:10.1029/WR006i005p01296.

Feyen, L., J. A. Vrugt, B. O. Nuallain, J. van der Knijff, and A. De Roo (2007), Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model, *J. Hydrol.*, *332*, 276–289, doi:10.1016/j.jhydrol.2006.07.004.

Freeze, R. A. (1978), Mathematical models of hillslope hydrology, in *Hillslope Hydrology*, edited by M. J. Kirkby, pp. 177–226, John Wiley, Hoboken, N. J.

Gandin, L. S. (1988), Complex quality control of meteorological observations, *Mon. Weather Rev.*, *116*(5), 1137–1156, doi:10.1175/1520-0493(1988)116<1137:CQCOMO>2.0.CO;2.

Goswami, M., and K. M. O'Connor (2010), A "monster" that made the SMAR conceptual model "right for the wrong reasons," *Hydrol. Sci. J.*, *55*(6), 913–927.

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

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*, 3802–3813, doi:10.1002/hyp.6989.

Gupta, H. V., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian (2014), Large-sample hydrology: A need to balance depth with breadth, *Hydrol. Earth Syst. Sci.*, *18*, 463–477, doi:10.5194/hess-18–463-2014.

Holländer, H. M., et al. (2014), Impact of modellers' decisions on hydrological a priori predictions, *Hydrol. Earth Syst. Sci.*, *18*, 2065–2085, doi:10.5194/hess-18–2065-2014.

Hey, T. (2012), The Fourth Paradigm - Data-Intensive Scientific Discovery, E-Science and Information Management - Communications in Computer and Information Science, vol. 317, 1 pp., doi:10.1007/978-3-642-33299-9_1.

Horton, R. E. (1931), The field, scope, and status of the science of hydrology, *Eos Trans. AGU*, *12*, 189–202.

Hoyt, W. G., et al. (1936), Studies of relations of rainfall and run-off in the United States, *U.S. Geol. Surv. Water Supply Pap. 772*, 301 pp., U.S. Gov. Print. Off., Washington, D. C.

James, L. D., and S. J. Burges (1982), Selection, calibration, and testing of hydrologic models, in *Hydrologic Modeling of Small Watersheds, Hydrol. Model. Monogr.*, edited by C. T. Haan, H. P. Johnson, and D. L. Brakensiek, chap. 11, pp. 437–472, Am. Soc. of Agric. Eng., St. Joseph, Mich.

Jordan, R. (1991), A one-dimensional temperature model for a snow cover, *Tech. Doc. SNTHERM.89*, 49 pp., U.S. Army Corps of Eng. Cold Reg. Res. and Eng. Lab., Hanover, N. H.

Kampf, S. K., and S. J. Burges (2010), Quantifying the water balance in a planar hillslope plot: Effects of measurement errors on flow prediction, *J. Hydrol.*, *380*, 191–202, doi:10.1016/j.jhydrol.2009.10.036.

Kauffeldt, A., S. Halldin, A. Rodhe, C. Y. Xu, and I. K. Westerberg (2013), Disinformative data in large-scale hydrological modeling, *Hydrol. Earth Syst. Sci.*, *17*(7), 2845–2857.

Killeen, T., and A. Blatecky (2011), Dear colleague letter: The "Earth Cube"—Towards a National Data Infrastructure for Earth System Science, *NSF 11–065*. NSF, Arlington, Va. [Available at http://www.nsf.gov/pubs/2011/nsf11065/nsf11065.jsp?org=NSF.]

Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.

Kuczera, G. (1996), Correlated rating curve error in flood frequency inference, *Water Resour. Res.*, *32*, 2119–2127, doi:10.1029/96WR00804.

Langbein, W. B., and J. V. B. Wells (1955), The water in the rivers and creeks, in *Water: The Yearbook of Agriculture*, edited by A. Stefferud, pp. 52–62, U.S. Dep. of Agric., Washington, D. C.

Loheide, S. P., II, and J. D. Lundquist (2009), Snowmelt-induced diel fluxes through the hyporheic zone, *Water Resour. Res.*, *45*, W07404, doi:10.1029/2008WR007329.

Lott, F. (2008), Adaptability of snowmelt models to changing climate in complex terrain, master thesis, 82 pp., Univ. of Wash, Seattle, Wash. [Available at http://depts.washington.edu/mtnhydr/publications/LottMSCEthesis2008final.pdf.]

Louis, J.-F. (1979), A parametric model of vertical eddy fluxes in the atmosphere, *Boundary Layer Meteorol.*, *17*(2), 187–202.

Lowry, C., J. Deems, S. Loheide, and J. D. Lundquist (2010), Linking snowmelt derived recharge and groundwater flow in a high elevation meadow system, Sierra Nevada Mountains, California, *Hydrol. Processes*, *24*, 2821–2833.

Lowry, C., S. P. Loheide, C. Moore, and J. D. Lundquist (2011), Groundwater controls on vegetation composition and patterning in mountain meadows, *Water Resour. Res.*, *47*, W00J11, doi:10.1029/2010WR010086.

Lundquist, J. D., and S. P. Loheide (2011), How evaporative water losses vary between wet and dry water years as a function of elevation in the Sierra Nevada, California and critical factors for modeling, *Water Resour. Res.*, *47*, W00H09, doi:10.1029/2010WR010050.

Lundquist, J. D., J. Minder, P. J. Neiman, and E. Sukovich (2010), Relationships between barrier jet heights, orographic precipitation gradients, and streamflow in the Northern Sierra Nevada, *J. Hydrometeorol.*, *11*, 1141–1156, doi:10.1175/2010JHM1264.1.

Lundquist, J. D., A. K. Massmann, J. Stimberis, and N. E. Wayand (2014), Surface meteorological and snow observations at Snoqualmie Pass, Univ. of Wash. Seattle, Wash. [Available at http://hdl.handle.net/1773/25611.]

Mahrt, L. (1987), Grid-averaged surface fluxes, *Mon. Weather Rev.*, *115*, 1550–1560.

McMillan, H., T. Krueger, and J. Freer (2012), Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, *Hydrol. Processes*, *26*, 4078–4111, doi:10.1002/hyp.9384.

Medawar, P. B. (1979), Advice to a Young Scientist, Basic Books, 109 pp., N. Y. [Available at http://www.albany.edu/~scifraud/data/sci_fraud_2927.html.]

Meek, D. W., and J. L. Hatfield (1994), Data quality checking for single station meteorological databases, *Agric. For. Meteorol.*, *69*, 85–109.
Meinzer, O. E. (Ed.) (1949), Hydrology *(McGraw-Hill 1942)*, 712 pp., Dover, N. Y.
Mellor, M. (1977), Engineering properties of snow, *J. Glaciol.*, *19*(81), 15–66.
Mizukami, N., and M. B. Smith (2012), Analysis of inconsistencies in multi-year gridded quantitative precipitation estimate over complex terrain and its impact on hydrologic modeling, *J. Hydrol.*, *428–429*, 129–141.
Montanari, A., and G. Di Baldassarre (2013), Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty, *Adv. Water Resour.*, *51*, 498–504.
Montanari, A., and D. Koutsoyiannis (2012), A blueprint for process-based modeling of uncertain hydrological systems, *Water Resour. Res.*, *48*, W09555, doi:10.1029/2011WR011412.
Mulvany, T. J. (1851), On the use of self-registering rain and flood gauges, *Trans. Inst. Civ. Eng. Ireland*, *4*, 1–8.
Oberg, K., and D. S. Mueller (2007), Validation of streamflow measurements made with acoustic Doppler current profilers, *J. Hydraul. Eng.*, *133*(12), 1421–1432, doi:10.1061/(ASCE)0733–9429(2007)133:12(1421).
Philip, J. R. (1975), Some remarks on science and catchment prediction, in *Prediction in Catchment Hydrology*, edited by T. G. Chapman and F. X. Dunin, pp. 23–30, Aust. Acad. of Sci., Canberra.
Pielke, R., Sr., et al. (2007), Documentation of uncertainties and biases associated with surface temperature measurement sites for climate change assessment, *Bull. Am. Meteorol. Soc.*, *88*, 913–928, doi:10.1175/BAMS-88-6-913.
Rasmussen, R., et al. (2012), How well are we measuring snow: The NOAA/FAA/NCAR winter precipitation test bed, *Bull. Am. Meteorol. Soc.*, *93*, 811–829, doi:10.1175/BAMS-D-11–00052.1.
Refsgaard, J. C., and J. R. Hansen (2010), A good-looking catchment can turn into a modeller's nightmare, *Hydrol. Sci. J.*, *55*(6), 899–912.
Reichle, R. H. (2008), Data assimilation methods in the Earth sciences, *Adv. Water Resour.*, *31*(11), 1411–1418.
Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resour. Res.*, *38*(11), 1241, doi:10.1029/2001WR000978.
Seyfried, M. S., and B. P. Wilcox (1995), Scale and the nature of spatial variability: Field examples having implications for hydrologic modeling, *Water Resour. Res.*, *31*, 173–184, doi:10.1029/94WR02025.
Sieck, L. C., S. J. Burges, and M. Steiner (2007), Challenges in obtaining reliable measurements of point rainfall, *Water Resour. Res.*, *43*, W01420, doi:10.1029/2005WR004519.
Slater, A. G., and M. P. Clark (2006), Snow data assimilation via an ensemble Kalman filter, *J. Hydrometeorol.*, *7*, 478–493, doi:10.1175/JHM505.1.
Smirnova, T. G., J. M. Brown, S. G. Benjamin, and D. Kim (2000), Parameterization of cold-season processes in the MAPS land-surface scheme, *J. Geophys. Res.*, *105*, 4077–4086, doi:10.1029/1999JD901047.
Steinacker, R., D. Mayer, and A. Steiner (2011), Data quality control based on self-consistency, *Mon. Weather Rev.*, *139*, 3974–3991, doi:10.1175/MWR-D-10–05024.1.
Tarboton, D., and C. Luce (1996), Utah energy balance snow accumulation and melt model (UEB), in *Computer Model Technical Description Users Guide*, 64 pp., Utah Water Res. Lab. and USDA For. Serv. Intermed. Res. Stn., Logan, Utah.
Theis, C. V. (1967), Aquifer and models, in *Symposium on Groundwater Hydrology*, edited by M. A. Marino, pp. 138–148, Am. Water Resour. Assoc., Middleburg, Va.
Udall press release (2013), Washington D. C. [Available at http://votesmart.org/publicstatement/835844/udall-nrcs-announce-plan-to-maintain-funding-for-snow-data-water-supply-monitoring-program#.VR76pkLWt74.]
United States Army Corps of Engineers (USACE) (1956), *Snow Hydrology: Summary Report of the Snow Investigations*, Portland, Oreg.
USGS (2009). United States Geological Survey Streamflow Information Program, Reston, Va. [Available at http://water.usgs.gov/nsip/history1.html.]
Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.
Western, A., and R. Grayson (2000), Soil moisture and runoff processes at Tarrawarra, in *Spatial Patterns in Catchment Hydrology, Observations and Modelling*, edited by R. Grayson and G. Bloschl, pp. 209–246, Cambridge Univ. Press, Cambridge, U. K.
Wigmosta, M. S., and S. J. Burges (1997), An adaptive modeling and monitoring approach to describe the hydrologic behavior of small catchments, *J. Hydrol.*, *202*, 48–77.
Wigmosta, M. S., L. W. Vail, and D. P. Lettenmaier (1994), A distributed hydrology-vegetation model for complex terrain, *Water Resour. Res.*, *30*, 1665–1679, doi:10.1029/94WR00436.
Wood, E. F., et al. (2011), Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water, *Water Resour. Res.*, *47*, W05301, doi:10.1029/2010WR010090.
World Meteorological Organization (WMO) (2008), Guide to meteorological instruments and methods of observation, Geneva, Switzerland. [Available at http://www.wmo.int/pages/prog/www/IMOP/CIMO-Guide.html.]
Yen, Y.-C. (1965), Effective thermal conductivity and water vapor diffusivity of naturally compacted snow, *J. Geophys. Res.*, *70*, 1821–1825, doi:10.1029/JZ070i008p01821.
Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resour. Res.*, *44*, W09417, doi:10.1029/2007WR006716.