2008

Computational provenance in hydrologic science: a snow mapping example

Royal Society Publishing

J. Dozier, J. Frew, "Computational provenance in hydrologic science: a snow mapping example," Philosophical Transactions of the Royal Society A, v.367 (2009), pp. 1021-1033.

http://hdl.handle.net/10945/52428

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 1021–1033
doi:10.1098/rsta.2008.0187
Published online 16 December 2008

Computational provenance in hydrologic science: a snow mapping example

By Jeff Dozier* and James Frew

Donald Bren School of Environmental Science and Management, University of California, Santa Barbara, CA 93106-5131, USA

Computational provenance—a record of the antecedents and processing history of digital information—is key to properly documenting computer-based scientific research. To support investigations in hydrologic science, we produce the daily fractional snow-covered area from NASA's moderate-resolution imaging spectroradiometer (MODIS). From the MODIS reflectance data in seven wavelengths, we estimate the fraction of each 500 m pixel that snow covers. The daily products have data gaps and errors because of cloud cover and sensor viewing geometry, so we interpolate and smooth to produce our best estimate of the daily snow cover. To manage the data, we have developed the Earth System Science Server (ES^3), a software environment for data-intensive Earth science, with unique capabilities for automatically and transparently capturing and managing the provenance of arbitrary computations. Transparent acquisition avoids the scientists having to express their computations in specific languages or schemas in order for provenance to be acquired and maintained. ES^3 models provenance as relationships between processes and their input and output files. It is particularly suited to capturing the provenance of an evolving algorithm whose components span multiple languages and execution environments.

Keywords: snow; remote sensing; data management; provenance

1. Introduction

(*a*) *The snow mapping problem*

Of the seasonal changes that occur on the Earth's land surface, perhaps the most profound is the accumulation and melt of seasonal snow cover, affecting climate, weather and the water balance. Snow exerts a huge influence on the hydrologic cycle during the winter and spring for much of the Earth's land area. Near many mountain ranges, the seasonal snow cover is the dominant source of run-off, filling rivers and recharging aquifers that more than a billion people depend on for their water resources (Barnett *et al.* 2005).

König *et al.* (2001) and Dozier & Painter (2004) have reviewed developments in remote sensing of snow and ice. Among them is the use of snow-covered area from NASA's moderate-resolution imaging spectroradiometer (MODIS) in

* Author for correspondence (dozier@bren.ucsb.edu).

One contribution of 24 to a Discussion Meeting Issue 'The environmental eScience revolution'.

hydrologic analysis and modelling. Unlike surface measurements, satellite observations are able to show the distribution of snow over the topography. Nearly daily maps are necessary for hydrologic and climate models because of the dynamic nature of snow cover, which changes at a slower time scale than atmospheric phenomena but faster than other surface covers. The availability of daily global observations of snow cover was inconceivable prior to the satellite era. Nowadays, the global MODIS snow-cover product (Hall *et al.* 2002) is produced daily, and as an 8 day composite, at 500 m spatial resolution.

Snow-water equivalent is regularly estimated at coarse spatial resolution (12–50 km) from passive microwave data, including SSM/R, SSM/I and the EOS instrument AMSR-E, in a time series that goes back to 1978 (National Research Council 2007). However, at finer spatial resolutions necessary for the mountains, remotely sensing snow-water equivalent is an unsolved problem. Our measurements of snow-covered area can be combined with point measurements to spatially interpolate snow-water equivalent (Fassnacht *et al.* 2003; Molotch *et al.* 2004).

(*b*) *The computational provenance problem*

Computational provenance refers to knowledge of the origins and processing history of a computational artefact, such as a data product or an implementation of an algorithm (Bose & Frew 2005). Provenance is an essential part of metadata for Earth science data products, where both the source data and the processing algorithms change over time. These changes can result from errors (e.g. sensor malfunctions or incorrect algorithms) and from an evolving understanding of the underlying systems and processes (e.g. sensor recalibration or algorithm improvement). Occasionally, such changes are memorialized as product or algorithm 'versions', but more often they are only manifest in mysterious differences between data products that one would otherwise expect to be similar. Provenance allows us to better understand the impacts of changes in a processing chain, and to have higher confidence in the reliability of any specific data product.

The snow mapping problem provides an ideal application for provenance capture and management. In §2, we show the steps to remotely sense fractional snow cover from MODIS, and interpolate across time and space to fill in the missing values and account for errors introduced by off-nadir views. In §3, we describe our provenance management system, and, in §4, we show how it is applied to the generation of MODIS snow maps.

2. Snow-covered area

Two products comprise information needed from MODIS for hydrologic modelling in the snowmelt environment:

— From the daily satellite overpasses, we calculate the fractional snow cover for each MODIS pixel that is not obscured by clouds. Requiring spectral unmixing and multiple solutions of simultaneous linear equations, this 'MODSCAG' algorithm (MODIS snow-covered area and grain size) is computationally intensive.

— The daily maps of fractional snow cover form a time–space data cube, with missing values because of cloud cover or sensor noise, and some less reliable values because of the highly off-nadir viewing geometry. We interpolate and smooth this data cube to provide our best estimate of the fractional snow cover in each grid cell, every day. This product is more useful for inclusion in a distributed energy balance snowmelt model than the raw daily estimates.

(*a*) *Fractional snow cover*

Snow-covered area in the mountains usually varies at a spatial scale finer than that of the ground instantaneous field of view of the remote-sensing instrument. This spatial heterogeneity poses a 'mixed-pixel' problem, because the sensor may measure the radiance reflected from a mixture of snow, rock or soil, and vegetation. Painter *et al.* (2003) used spectral mixture analysis with the data from an airborne imaging spectrometer, AVIRIS, with 224 contiguous spectral bands, to estimate fractional snow-covered area for each pixel. While these results with AVIRIS demonstrated the ability to derive both snow cover and albedo at subpixel resolution, imaging spectrometer data are available too infrequently to use them regularly in hydrologic models. Multispectral sensors, such as the Landsat Thematic Mapper and MODIS, provide data over wider swaths and at more frequent intervals than imaging spectrometers. From MODIS, information is available at 500 m spatial resolution, at twice daily intervals using the data from two satellites, Terra and Aqua.

In the spectral mixture analysis, an endmember is the spectral reflectance of a pure surface cover. The MODIS algorithm uses a spectral library for snow generated with model calculations for monodispersions of spheres. We calculate the snow grains' single-scattering properties over each MODIS band with the Mie theory (Nussenzveig & Wiscombe 1980), and the reflectance of the snowpack with a discrete-ordinates radiative transfer model (DISORT; Stamnes *et al.* 1988) at the solar geometry of the image. The other endmembers are vegetation, rock or soil, lake ice and photometric shade to account for the variability of the illumination angle. For their reflectance values, we use the spectra measured in the field and laboratory at 1 nm spectral resolution, convolved to the MODIS bandpasses.

The spectral mixture analysis is based on a set of simultaneous linear equations that make up the components of a pixel's reflectance. The MODIS product suite includes MOD09, atmospherically corrected surface reflectance (Kotchenova & Vermote 2007), along with a cloud mask that, unfortunately, has errors of both omission and commission in snow–cloud discrimination. The spectral mixing equation is

$$R_{S,\lambda} = \sum_{i=1}^{N} F_i R_{\lambda,i} + \epsilon_\lambda,$$

where $R_{S,\lambda}$ is the pixel-averaged MODIS surface reflectance; F_i is the fraction of endmember i; $R_{\lambda,i}$ is the reflectance of endmember i at wavelength λ; N is the number of spectral endmembers; and ϵ_λ is the residual error at λ for the fit of the N endmembers. The least-squares fit to F_i can be solved by several standard methods. In full exploratory mode, there are 100 snow endmembers with effective grain radii from 10 to 1100 μm, 25 vegetation endmembers, 25 soil endmembers

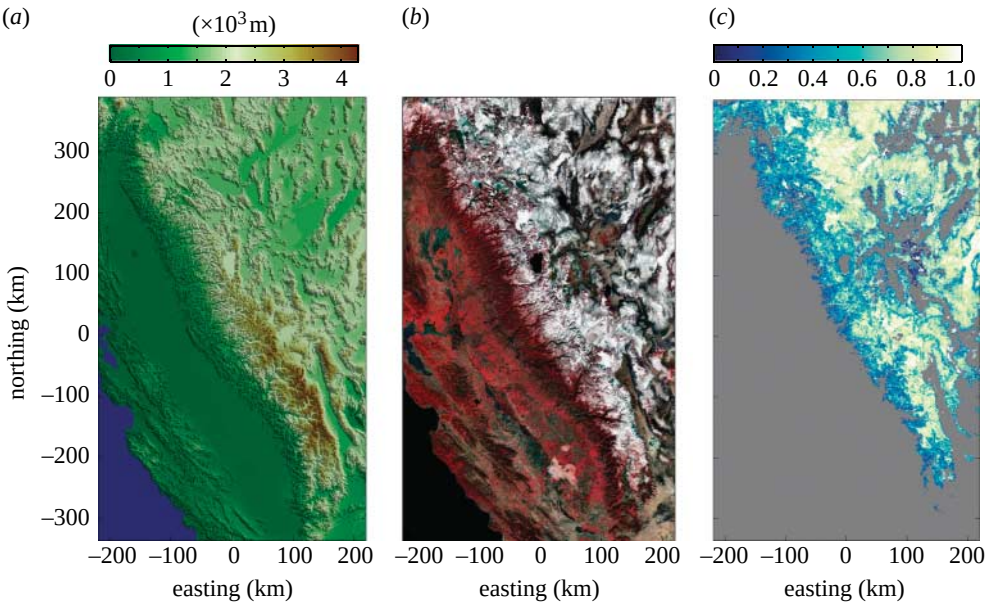

Figure 1. (*a*) Elevations of the Sierra Nevada, Coast Ranges and southern Cascades. (*b*) MODIS bands 2, 4 and 3 (0.841–0.876, 0.545–0.565 and 0.459–0.479 μm) on 19 January 2008. (*c*) Snow cover in the Sierra Nevada and western Nevada on 19 January 2008. The Albers equal-area projection, used by the California Data Exchange Center, has standard parallels at 40.5° N and 34.0° N, centre longitude 120° W.

and therefore approximately 6000 combinatorial possibilities. In operational mode, we reduce the number of snow endmembers by using intervals of reflectance in MODIS band 5 rather than intervals of grain radius, and we reduce the number of soil and vegetation endmembers because the time series allows us to constrain the range of possibilities. The snow endmember can vary from pixel to pixel, but within a pixel we assume that the snow endmember is the same; the snow endmembers are spectrally too similar to one another to resolve different snow endmembers in the same pixel. MODSCAG solves each possible combination of endmembers and chooses the combination with the lowest error for $M=7$ MODIS spectral bands, defined by

$$\text{r.m.s.e.} = \left(\frac{1}{M} \sum_{\lambda=1}^{M} \epsilon_{\lambda}^{2} \right)^{1/2}.$$

Figure 1 shows the topography of California's Sierra Nevada, a MODIS image from mid-January 2008, and the fractional snow cover derived from that image. We have compared MODSCAG's performance with fractional snow cover from Landsat at 30 m spatial resolution, which has in turn been verified against aerial photographs at sub-metre resolution (Rosenthal & Dozier 1996). The r.m.s. difference between Landsat and MODIS fractional snow cover is generally in the 0.07–0.12 range, with part of the error caused by misregistration between MODIS and Landsat.

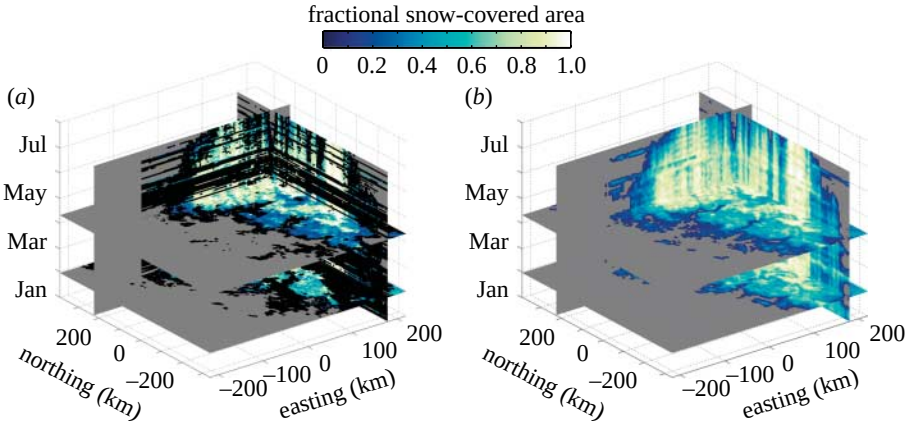

Figure 2. Representation of the process as interpolation and smoothing to fill in a sparse time–space data cube, with (*a*) the filtered cube, where cloudy or noisy pixels are black, and (*b*) the interpolated and smoothed cube. Data are for the combined Tuolumne and Merced River basins in the Sierra Nevada, January to July 2004.

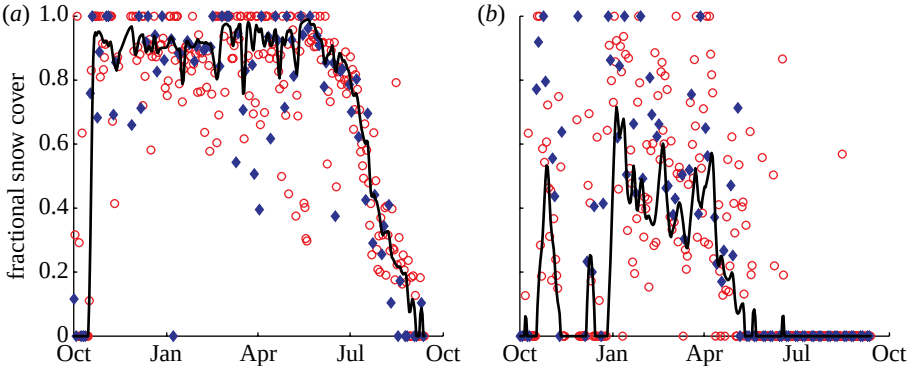

Figure 3. Raw data and interpolation and smoothing for two locations in the Tuolumne River drainage from October 2004 to July 2005. A point at (*a*) 3550 m near the eastern boundary of the basin and (*b*) 1800 m near the seasonal snowline. The points show the raw MODIS snow-cover measurements, but with cloudy and noisy data omitted. Diamonds and circles show the data at sensor zenith angles less than and greater than 25°, respectively. The lines show the smooth fit to the data, including interpolation by date and then modest spatial smoothing.

(*b*) *Time–space interpolation*

Several confounding factors make the raw daily snow cover difficult to use in hydrologic models. These include cloud cover, sensor noise, artefacts caused by viewing geometry, angular effects on the signal caused by vegetation, other elements such as subpixel clouds and imperfections in the retrieval algorithms. However, the time series itself enables us to better estimate snow properties.

In effect, we can view the snow data as a sparse space–time cube that needs filtering, smoothing and interpolation (figure 2). *Filters* replace cloudy or noisy values in the cube with NaNs ('not a number') to set the right sparse values in the cube. After filtering, the task is to *interpolate* and *smooth* the legitimate values, which may have errors, to fill in the cube. Clouds are often extensive and

persistent in snow-covered areas, as is the case in winter in the Sierra Nevada. Sometimes, software errors in the MODIS data processing stream have contaminated data for a week or two. Occasional noisy images appear throughout the history of MODIS.

When clouds do not obscure the pixels, viewing angles can perturb the signal. Although the MOD09 product is resampled at 500 m resolution, in fact, the pixels are elongated in both the along-track and cross-track directions at off-nadir views. At the edge of the scan, the pixel is approximately 10 times as large as at nadir, elongated by factors of 2 and 5 in the along-track and cross-track directions, respectively. Therefore, the putative values of the MOD09 reflectance, and thereby the derived snow cover, incorporate values in neighbouring pixels. The neighbours may have more or less snow cover, so the error is not systematic. Topography and forest cover also affect the viewing geometry. Within a given viewing area, there can be considerable subpixel topographic variability. In forested areas, a sensor 'sees' greater fractions of the underlying ground at near-nadir views and lesser fractions as the view angle increases (Nolin 2004). Finally, MODSCAG probably has imperfections, especially at large view angles where the effect of subpixel topographic variability is amplified.

We start filling in the space–time cube by smoothing the values along the time axis. We choose this approach, rather than generalized three-dimensional interpolation or generic data assimilation, because MODIS sensor zenith angles oscillate between near-nadir and more than 65° in a regular 16 day pattern, whereas there is only smooth variability in viewing angles across an image on a particular day. From a set of fractional snow-cover values $\hat{S}(t)$ estimated at a discrete set of times t, a smoothing spline (de Boor 2007) for the best estimate of $S(t)$ minimizes

$$q\sum_{j=1}^{N} w_j[\hat{S}(t_j) - S(t_j)]^2 + (1-q)\int_{t_{\min}}^{t_{\max}} \xi(t)\left(\frac{\mathrm{d}^2 S}{\mathrm{d}t^2}\right)^2.$$

The limits of integration are the time period of the analysis. A period of 32 days covers two MODIS viewing angle cycles. The typical maximum is a whole snow season. The smoothing parameter q is in the range of 0–1. When $q=0$, $S(t)$ is a least-squares, straight-line fit to the data. When $q=1$, $S(t)$ is a natural cubic spline that goes through each datum point. We generally let the MATLAB function `csaps` choose $q=1/[1+(\Delta t)^3/6]$ adaptively based on the average spacing Δt between data points; therefore, q can vary spatially depending on the extent of cloud cover or missing data. The weights w also vary from 0 to 1, and $\xi(t)$ is a piecewise polynomial approximation to provide weights for continuous values of t. In our case, our confidence in the data decreases as the viewing angle increases, so we choose weights $w_j=\cos\theta_s(p_0/p_s)$, the cosine of the viewing angle from the pixel to the satellite multiplied by the ratio of a pixel's area at nadir to that on day j.

If the data are distributed well along the abscissa, a smoothing spline interpolates well; however, if there are large gaps in the input data, for example from cloud cover, a smoothing spline can yield unlikely values in those gaps. Therefore, after setting smoothed daily values for the fractional snow cover, we interpolate between the missing days using a piecewise interpolant (Fritsch & Carlson 1980). After interpolating and smoothing by date, we smooth the whole time–space cube with a Gaussian filter. Figure 3 shows time interpolation for two

locations in the 2004–2005 water year: one in the upper reaches of the Tuolumne River basin at an elevation of 3550 m and the other at 1800 m near the seasonal snowline. Because we have more confidence in the measures at near-nadir viewing angles, the graph identifies those days when the sensor viewing angle is within 25°, true for approximately one-third of the observations; at that angle the weight $w_j = 0.689$.

3. Computational provenance in ES^3

The Earth System Science Server (ES^3) is a software environment for data-intensive Earth science. ES^3 has unique capabilities for automatically and transparently capturing, managing and reconstructing the provenance of arbitrary, unmodified computational sequences (Frew *et al.* 2008). Automatic acquisition is critical to avoid the inaccuracies and incompleteness of human-specified provenance (i.e. annotation). Transparent (i.e. invisible to the user) acquisition avoids the computational scientists having to learn, and be constrained by, a specific language, schema or system in which their problem must be expressed in order for provenance to be captured and maintained. 'Unmodified' means that provenance capture by ES^3 requires no changes whatsoever to existing programs.

Unlike almost all other provenance management systems (Bose & Frew 2005; Simmhan *et al.* 2005), ES^3 captures provenance information from running processes, as opposed to extracting it from static specifications such as scripts or workflows. ES^3 provenance management can thus be added to any existing scientific computations, without modifying or respecifying them.

(*a*) *ES provenance model*

ES^3 models provenance as a directed graph of processes and their input and output files. Here, we use 'process' in the operating system sense of a specific execution of a program. In other words, each execution of a program or access to a file yields a new set of provenance events.

Relationships between files and processes are captured by monitoring file system events (open, close, read, write, etc.). This monitoring can take place at many levels: operating system calls; application library calls; and arbitrary checkpoints within the source code. Any combination of monitoring levels may be active simultaneously, and all are transparent to the scientist–programmer using the system.

An ES provenance 'report' is a serialized graph of metadata representing the files and processes resulting from a specific invocation event (e.g. a 'job'). Nested processes (processes that spawn other processes) are correctly represented. In addition to retrieving the entire provenance of a job, ES^3 supports arbitrary forward (descendant) and/or reverse (ancestor) provenance retrieval, starting at any specified file or process.

(*b*) *ES^3 implementation*

ES^3 is implemented as a provenance-gathering client and a provenance-managing server (figure 4). The client, which runs in the same environment as the processes whose provenance is being tracked, is a set of *logger* processes that intercept raw messages from various *plugins* in the execution environment:

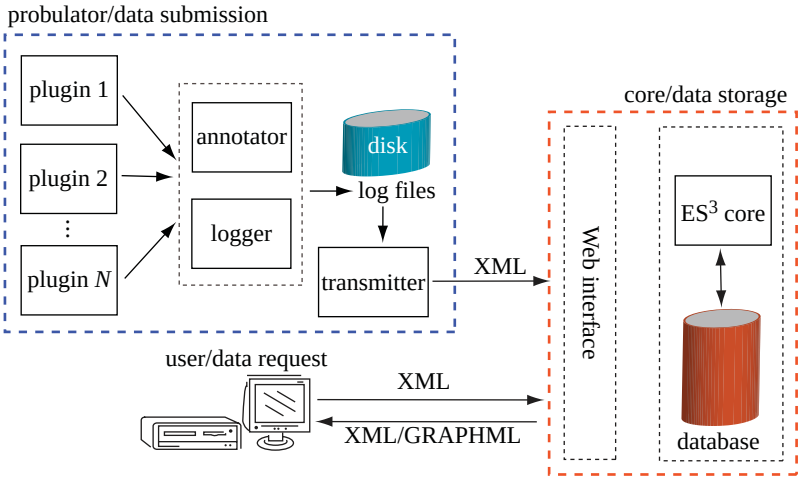

Figure 4. ES^3 architecture.

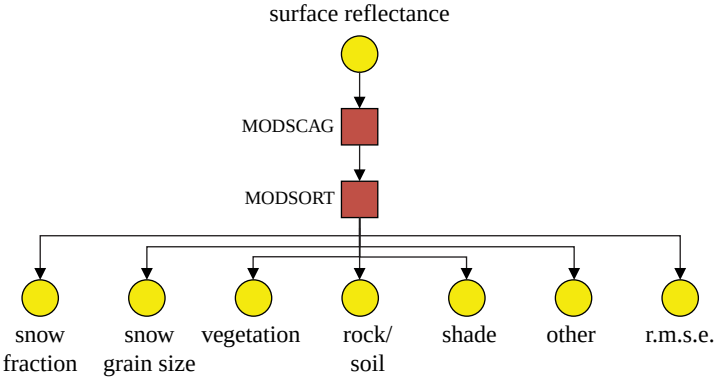

Figure 5. Top-level workflow for a fractional snow-covered area algorithm.

— Provenance-relevant operating system calls (e.g. `open( )`) are monitored via the `strace` system call tracing facility.
— Application library calls are monitored by transparently substituting instrumented versions of the libraries. ES^3 provides a set of instrumented libraries for the IDL (ITT Visual Information Solutions, Boulder, CO) and MATLAB (The MathWorks, Natick, MA) scientific programming environments. (Note that these instrumented libraries are mostly simple wrappers for the vendor-provided routines, and are thus relatively unaffected by software upgrades.)
— Arbitrary events (e.g. calls to specific functions) are monitored by automatically invoked source-to-source preprocessors that transparently insert monitoring statements in the interpreted source code. ES^3 provides preprocessors for the IDL and MATLAB source codes.

(Note that the execution time overhead incurred by these plugins is relatively minor; for example, it is dwarfed by the computational complexity of the MODSCAG algorithm. We would expect ES^3 provenance capture to prove burdensome for processes only where system call activity dominated their execution time.)

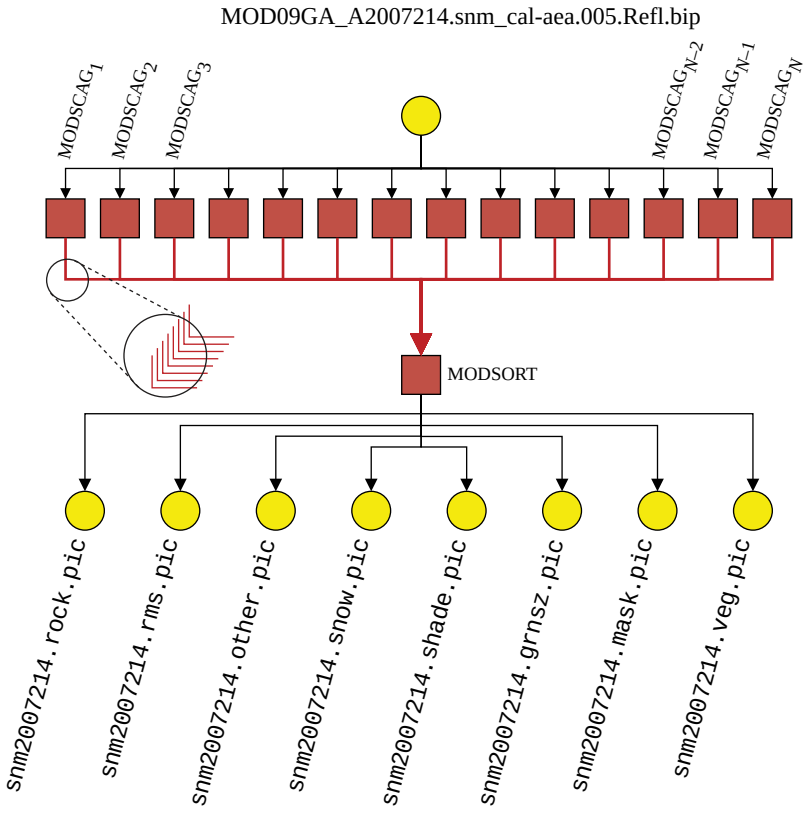

Figure 6. Provenance of the snow-covered area algorithm (MODSCAG nesting expanded).

The logger writes these provenance messages to local log files. A separate *annotator* client optionally examines the files and directories being accessed by the instrumented processes, retrieves certain non-provenance metadata (e.g. README files and source code comments) and adds them to the log files.

A common *transmitter* client asynchronously scans the log files, assembles the provenance events into a time-ordered stream, assigns unique identifiers (UUIDs) to each file and process being tracked and submits a raw provenance report to the *ES*^{3} *server*. Since these reports contain only metadata, they are quite compact.

Queries sent to the ES^3 server return provenance reports in an XML-based serialized graph format, directed either forward (descendants) or backward (ancestors) from a specified data or process object. In addition to a graphical display, as shown in figure 5, these reports are amenable to post-processing, such as determining the differences between two provenance graphs (see Frew *et al.* (2008) for an example).

It should be emphasized that provenance in ES^3 is a *post hoc* description of a processing sequence, not an executable workflow. While it is possible to reconstruct and re-execute a computational sequence from an ES^3 provenance report, there is no guarantee that the files and programs referenced in the report are accessible, or even still exist. The ES^3 server includes a general storage

facility that can be used to preserve snapshots of files or programs, and allow robust references to them in the provenance reports, but this is not required for provenance to be valid.

4. Capturing MODSCAG provenance

We use ES^3 to track the provenance of a MODSCAG run using satellite imagery of portions of the Sierra Nevada (figure 1). The snow-covered area product involves processing steps implemented in IDL, C and UNIX shell scripts, and the space–time interpolation uses MATLAB. The algorithms are under active development. Figure 5 shows an idealized top-level workflow for MODSCAG. A satellite image of surface reflectance is processed by MODSCAG into multiple estimates of the surface composition of each pixel. MODSORT selects the best of these estimates for each pixel and creates a suite of output grids whose cell values are the percentage of snow (figure 1) and other components present at the corresponding surface location, as well as estimates of snow grain size, classification error, and whether the input pixel was too deeply shaded by the surrounding terrain to be usable.

ES^3 is particularly useful for elucidating 'hidden' provenance—dependencies between files and processes that are not explicitly stated in the workflows or scripts that invoke the processes—and for managing highly nested provenance graphs. Requesting forward provenance for an actual MODIS image (figure 6) reveals that the MODSCAG workflow step actually comprises 30 separate invocations of the MODSCAG program, each using different starting assumptions about surface composition, which MODSORT merges into a single set of output files.

The ES^3 request that yielded figure 6 included a restriction to avoid expanding nested workflows. Relaxing this restriction for an entire MODSCAG workflow would yield a provenance graph too complex for a printed page. Instead, figure 7 shows the combined forward and reverse provenance for a single one of the 30 MODSCAG program invocations. Imagine variations in figure 7 replacing all 30 processes in figure 6 to get an idea of the complexity of a complete MODSCAG 'run'.

Note that since figure 7 is a portion of a much larger provenance graph, it provides sufficient information for some provenance assertions but not others. For example, it correctly shows that the file `snm2007214.snow.pic` is derived from the image `MOD09GA.A2007214.snm_cal-aea.005.Refl.bip`, but does not show any of `snm2007214.snow.pic's` possible antecedents from any of the other 29 MODSCAG invocations.

Although not illustrated here, we should mention that arbitrary metadata can be associated with any component (node or edge) of a provenance graph (e.g. a data object node often contains the URI for the associated file). The provenance graph can thus be a convenient 'scaffold' for all of the metadata available for a particular computation.

5. Conclusion

As climate and land-use change and populations grow, the empirical methods of managing water, which are based on historical relationships between point measurements and run-off, are likely to become less accurate, a conundrum that

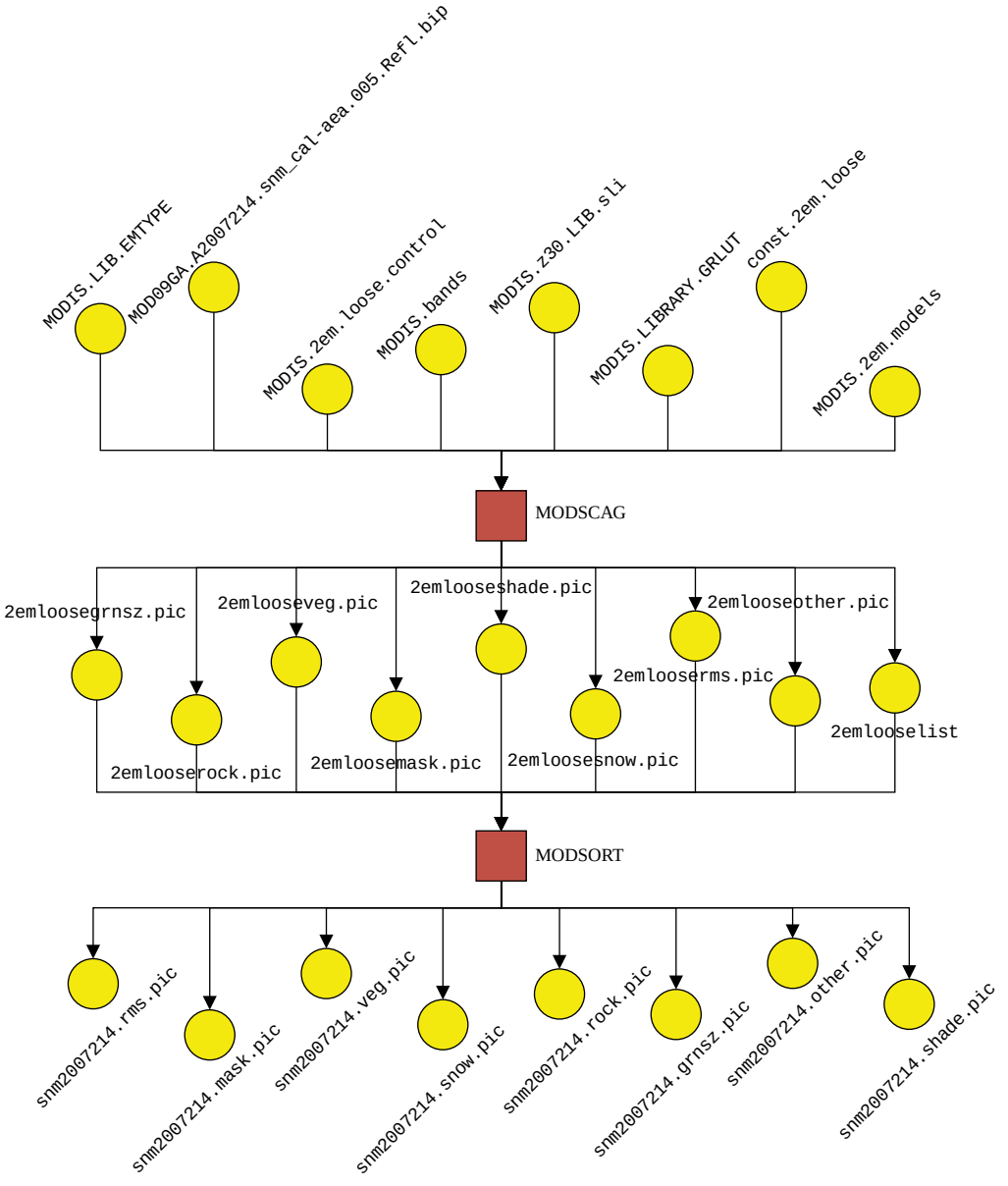

Figure 7. Provenance of the algorithm (single MODSCAG invocation detail).

Milly *et al.* (2008) characterize as 'stationarity is dead'. Hence, the usefulness of distributed snowmelt models based on a judicious integration of remotely sensed and surface measurements will increase (Bales *et al.* 2006).

However, the translation of reflectance measurements from MODIS into a product that is useful for hydrologic analyses involves complicated, somewhat arcane knowledge. Because snow changes dynamically, daily data allow one to overcome the limitations imposed by clouds and off-nadir viewing to reconstruct a daily time series. The daily product is useful for a variety of hydrologic models

and analyses, including interpolation of spatially distributed snow-water equivalent, without the need for the user to interpolate and filter the patchy daily maps.

The complexity of the transformations that must be applied to render satellite observations useful to the scientist makes it imperative to automate, as far as possible, the acquisition and management of the associated metadata. We have demonstrated the ability to capture provenance metadata for a significant portion of the MODIS product workflow, without any effort on the part of the scientists developing and refining these products. Furthermore, the provenance we automatically collect can be rendered graphically in a form that is easily intelligible to those who must evaluate the product's fitness for use.

Our work is supported by NASA Cooperative Agreements NNG0C52A and NNG04GE66G and Naval Postgraduate School grant N00244-07-1-0013.

References

Bales, R. C., Molotch, N. P., Painter, T. H., Dettinger, M. D., Rice, R. & Dozier, J. 2006 Mountain hydrology of the western United States. *Water Resour. Res.* **42**, W08432. (doi:10.1029/2005WR004387)

Barnett, T. P., Adam, J. C. & Lettenmaier, D. P. 2005 Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature* **438**, 303–309. (doi:10.1038/nature04141)

Bose, R. & Frew, J. 2005 Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.* **37**, 1–28. (doi:10.1145/1057977.1057978)

de Boor, C. 2007 *Spline Toolbox 3: user's guide*. Natick, MA: The MathWorks.

Dozier, J. & Painter, T. H. 2004 Multispectral and hyperspectral remote sensing of alpine snow properties. *Annu. Rev. Earth Planet. Sci.* **32**, 465–494. (doi:10.1146/annurev.earth.32.101802.120404)

Fassnacht, S. R., Dressler, K. A. & Bales, R. C. 2003 Snow water equivalent interpolation for the Colorado River Basin from snow telemetry (SNOTEL) data. *Water Resour. Res.* **39**, 1208. (doi:10.1029/2002WR001512)

Frew, J., Metzger, D. & Slaughter, P. 2008 Automatic capture and reconstruction of computational provenance. *Concurrency Comput. Pract. Exp.* **20**, 485–496. (doi:10.1002/cpe.1247)

Fritsch, F. N. & Carlson, R. E. 1980 Monotone piecewise cubic interpolation. *SIAM J. Numer. Anal.* **17**, 238–246. (doi:10.1137/0717021)

Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGiromamo, N. & Bayr, K. J. 2002 MODIS snow-cover products. *Remote Sens. Environ.* **83**, 181–194. (doi:10.1016/S0034-4257(02)00095-0)

König, M., Winther, J.-G. & Isaksson, E. 2001 Measuring snow and glacier ice properties from satellite. *Rev. Geophys.* **39**, 1–28. (doi:10.1029/1999RG000076)

Kotchenova, S. Y. & Vermote, E. F. 2007 Validation of a vector version of the 6S radiative transfer code for atmospheric correction of satellite data, part II: homogeneous lambertian and anisotropic surfaces. *Appl. Opt.* **46**, 4455–4464. (doi:10.1364/AO.46.004455)

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P. & Stouffer, R. J. 2008 Stationarity is dead: whither water management? *Science* **319**, 573–574. (doi:10.1126/science.1151915)

Molotch, N. P., Fassnacht, S. R., Bales, R. C. & Helfrich, S. R. 2004 Estimating the distribution of snow water equivalent and snow extent beneath cloud cover in the Salt-Verde River basin, Arizona. *Hydrol. Proc.* **18**, 1595–1611. (doi:10.1002/hyp.1408)

National Research Council 2007 *Earth observations from space: the first 50 years of scientific achievement*. Washington, DC: National Academies Press.

Nolin, A. W. 2004 Towards retrieval of forest cover density over snow from the multi-angle imaging spectroradiometer (MISR). *Hydrol. Proc.* **18**, 3623–3636. (doi:10.1002/hyp.5803)

Nussenzveig, H. M. & Wiscombe, W. J. 1980 Efficiency factors in Mie scattering. *Phys. Rev. Lett.* **45**, 1490–1494. (doi:10.1002/hyp.1408)

Painter, T. H., Dozier, J., Roberts, D. A., Davis, R. E. & Green, R. O. 2003 Retrieval of subpixel snow-covered area and grain size from imaging spectrometer data. *Remote Sens. Environ.* **85**, 64–77. (doi:10.1016/S0034-4257(02)00187-6)

Rosenthal, W. & Dozier, J. 1996 Automated mapping of montane snow cover at subpixel resolution from the Landsat Thematic Mapper. *Water Resour. Res.* **32**, 115–130. (doi:10.1029/95WR02718)

Simmhan, Y. L., Plale, B. & Gannon, D. 2005 A survey of data provenance in e-science. *ACM SIGMOD Rec.* **34**, 31–36. (doi:10.1145/1084805.1084812)

Stamnes, K., Tsay, S.-C., Wiscombe, W. J. & Jayaweera, K. 1988 Numerically stable algorithm for discrete-ordinate-method radiative transfer in multiple scattering and emitting layered media. *Appl. Opt.* **27**, 2502–2509.