

SCIENTIFIC DATA

OPEN

Data Descriptor: A cloud-free MODIS snow cover dataset for the contiguous United States from 2000 to 2017

Received: 16 March 2018
Accepted: 8 November 2018
Published: 15 January 2019

Hoang Tran¹, Phu Nguyen^{1,2}, Mohammed Ombadi¹, Kuo-lin Hsu¹, Soroosh Sorooshian¹ & Xia Qing¹

This article presents a cloud-free snow cover dataset with a daily temporal resolution and 0.05° spatial resolution from March 2000 to February 2017 over the contiguous United States (CONUS). The dataset was developed by completely removing clouds from the original NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Snow Cover Area product (MOD10C1) through a series of spatiotemporal filters followed by the Variational Interpolation (VI) algorithm; the filters and VI algorithm were evaluated using bootstrapping test. The dataset was validated over the period with the Landsat 7 ETM+ snow cover maps in the Seattle, Minneapolis, Rocky Mountains, and Sierra Nevada regions. The resulting cloud-free snow cover captured accurately dynamic changes of snow throughout the period in terms of Probability of Detection (POD) and False Alarm Ratio (FAR) with average values of 0.955 and 0.179 for POD and FAR, respectively. The dataset provides continuous inputs of snow cover area for hydrologic studies for almost two decades. The VI algorithm can be applied in other regions given that a proper validation can be performed.

Design Type(s)	modeling and simulation objective • process-based data transformation objective
Measurement Type(s)	snow
Technology Type(s)	digital curation
Factor Type(s)	temporal_interval
Sample Characteristic(s)	contiguous United States of America • climate system

¹Center for Hydrometeorology and Remote Sensing, Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California, USA. ²Nong Lam University, Ho Chi Minh City, Vietnam. Correspondence and requests for materials should be addressed to H.T. (email: htranvie@uci.edu)

Background & Summary

The Moderate Resolution Imaging Spectroradiometer (MODIS) snow cover area (SCA) product¹ serves as a reliable source of snow measurements for hydrologic studies as well as for data assimilation in climate models. Researchers used and evaluated the product in various regions including the Columbia and Missouri river basins², Austria³, Xinjiang, China⁴, and in Sierra Nevada^{5,6}. The datasets were also incorporated into a land surface model⁷ and a hydrology model⁸ for data assimilation.

Nevertheless, cloud obscuration limits the product's usage. Clouds block satellites from capturing the ground state of the earth surface (i.e. snow/land), especially during snow accumulation periods when clouds gather and block the snow cover extent information from the ground. Major concerns about MODIS snow maps polluted by clouds and snow/cloud discrimination have been repeatedly mentioned in its assessment studies^{9,10} as well as in the MODIS product user guide¹¹. Hence, researches focus on the removal of clouds from snow images have been conducted for years and have been classified into two main categories: model-driven methods and data-driven methods. To estimate snow extent beneath clouds, the model-driven methods rely on the relationships between snow and other factors such as: energy and mass exchanges^{12,13}, temperature¹⁴, and grain size¹⁵.

On the other hand, data-driven methods take into account the persistent characteristics of snow to remove clouds. For example, Dozier *et al.*¹⁶ considered snow data as a sparse space-time cube that could be filled by temporal cubic spline interpolation. Their results demonstrated that the interpolated and smoothed product has more consistent snow-covered area in the Tuolumne and Merced River basins throughout the water year 2005 than the raw, cloud cover filtered data. However, Dozier *et al.* approach is limited to 1-D interpolation primarily because of frequent zenith angles oscillation and slow computation. Later on, Gafurov and Bardossy¹⁷ introduced a series of six spatiotemporal filters to mitigate cloud cover from MODIS images in the Kokcha River basin in Afghanistan, detailed information of this study will be presented in subsequent sections. Parajka *et al.*¹⁸ proposed a regional snow-line method (SNOWL) utilizing elevation information for de-clouding; the method robustly recovered snow cover maps from cloud over Austria¹⁸. In another study, Hall *et al.*¹⁹ suggested a cloud-gap-filled (CGF) method to produce a snow cover map with cloud-persistence count (CPC) for each grid where lower CPC snow grids are more likely to have snow. While approaches from Parajka *et al.*¹⁸ and Hall *et al.*¹⁹ are simple and proven to be suitable for use in hydrological and global models, persistent cloudy conditions during snow accumulation periods may reduce their reliability significantly. More recently, Dong *et al.*²⁰ employed information from snow stations to estimate ground states of cloud-cover areas from MODIS in south-western Germany. Although this approach is effective in areas with dense snow networks, it requires manually determining thresholds for each station's predicting capability of nearby snow cover based on station location and elevation.

To completely remove clouds and delineate dynamic snow boundaries, Xia *et al.*²¹ implemented the Variational Interpolation (VI) method²² for interpolating the three-dimensional space-time cube of snow cover proposed by Dozier *et al.*¹⁶. Evaluation results in the Sierra Nevada mountain range demonstrated that the method was robust and accurate since during the accumulation period (25–27) March 2007, the “Cleared” images obtained from VI had an average omission error and commission error of about 22.5 and 2.1% respectively since the initial omission errors from the original Terra and Aqua images were 14.3 and 20.2% respectively. Meanwhile during the melting period (14–16) March 2009, original satellite observations were not obscured by clouds as severely as accumulated ones, hence the errors of omission and commission between ‘Cleared’ and MODIS images were similar (5.7 to 5% of omission errors and 0% for commission errors). However, the main drawback of the original VI method is the system instability which limits its implementation on a larger scale. In this study, an improved VI version is introduced by integrating MINimum RESidual (MINRES) iterations²³ to prevent the system from breaking up when applied to much broader scales.

The procedure of developing the cloud-free snow cover dataset consists of two parts. First, estimation of the ground states (i.e. snow/land) of cloud-hindered grids via five filters adopted from Gafurov *et al.*¹⁷. The resulting images are still obstructed by clouds but provide more details about snow boundaries which is beneficial to the second step. In the second stage, the improved VI method is implemented to reconstruct a three-dimensional time-varying snow cover boundary. Subsequently, this 3-d surface can be used to obtain cloud-free snow cover images.

Methods

The spatial domain of the dataset developed in this study is the contiguous United States (CONUS) which covers about 8,080,464.3 km², ranges between 24°30N and 49°25N in latitude and from 66°57W to 124°46W in longitude. During winter seasons, from November to the end of February, the snow cover extent for the whole CONUS varies from one million km² to four million km² which plays a crucial role in energy and hydrological cycles^{24,25}.

The main inputs in this study are products from MODIS/Terra Snow Cover Daily (MOD10C1) and MODIS/Aqua Snow Cover Daily (MYD10C1) version 6 (Data Citation 1) released in July 2016 by the National Snow & Ice Data Center (NSIDC). These daily Climate Modeling Grid (CMG) products are in a sequence of MODIS snow product suite¹¹, beginning with the 500 m resolution swath product (MOD10_L2).

As reported in Riggs and Hall¹¹, the swath level snow mapping algorithm is based on the Normalized Difference Snow Index (NDSI)^{26,27}. NDSI is calculated for Terra/MODIS using band 4 and band 6, for Aqua/MODIS using band 4 and band 7.

$$NDSI = \frac{\text{Band4} - \text{Band6}}{\text{Band4} + \text{band6}} \quad (1)$$

The global criteria for snow is NDSI greater than 0.4 and near-infrared reflectance (band 2) greater than 0.11 and band 4 reflectance greater than 0.10. To increase snow detection sensitivity in forested landscapes, the MOD10_L2 product combines the Normalized Difference Snow Index (NDSI) range from 0.1 to 0.4 with the Normalized Difference Vegetation Index (NDVI)¹¹. After a pixel is classified as snow, “it is subjected to a series of screens to alleviate snow commission errors and flag uncertain snow detections”¹¹. More details about the screens can be found in the product user guide¹¹. Here, we only summarized the screens main thresholds: (1) Version 6 combines surface temperature and height screen, if snow pixels are in low elevation (<1300 m) and warm surfaces (>283 K), they are reversed to no-snow. This new surface temperature screen solves the problem of detecting snow in mountain ranges during spring and summer brought up by Rittger *et al.*^{5,10}; (2) If snow pixels have low reflectance (Very High Visible (VIS) of band 2 is ≤ 0.10 or band 4 is ≤ 0.11) or low illumination (solar zenith angles > 70°), they will be set as no-snow or night pixels; (3) Low NDSI or unusually high Short-Wave Infrared (SWIR) reflectance snow pixels will be converted back to no-snow pixels.

The next product in the sequence, MOD10A1, only selects one 'best' observation from all the MOD10_L2 swaths over a location using strict criteria including solar elevation, distance from nadir, and observation cover. Selecting an observation closest to nadir with maximum coverage of the cell⁹ could solve problems from the off nadir viewing from the MODIS reflectance product (MOD09) mentioned by Dozier *et al.*¹⁶.

The main input of the study, MOD10C1/MYD10C1, “maps 500 m MOD10A1 observations into 0.05° CMG cells. Outputs for a grid cell are determined by the percentage of counts of observations, snow or cloud, mapped in the cell”¹¹. The daily MODIS snow product suite has “an overall accuracy of about 93%, lower accuracy is found in forested areas”⁹. More recent studies from Liu *et al.*²⁶ and Rittger *et al.*⁵ also highlighted scenarios in which dense canopy limits the product ability to detect snow cover.

It should be noted that the launch dates of the two satellites, Terra and Aqua, are different with December 18, 1999 and May 4, 2002 for Terra and Aqua respectively. In this study, the term MODIS-SCA product will be used to represent both products, MOD10C1 and MYD10C1, from Terra and Aqua whenever the text refers to 2002 or later. When referencing prior to 2002, the term only represents products from Terra satellite.

Overall, the process of creating the cloud-free product from the MOD10C1/MYD10C1 product starts with reclassifying MODIS into one of three categories, namely snow, land (no-snow), and cloud based on the threshold of 50% fractional: if a grid has a percentage of snow greater than 50%, it is set as snow. If the sum of the snow and cloud fractions in one location is smaller than 50%, a grid is marked as land (no-snow). If neither snow nor no-snow, a grid is set as cloud. Next, the reclassified MODIS images passed through two subsequent steps based on a series of filters and the VI algorithm. These two steps are discussed in the following subsections.

Mitigated filters

The filters are used as a first step to retrieve cloud-free snow cover images. This method has been adopted from Gafurov and Bardossy¹⁷ and it consists of five filters to mitigate cloud obstruction: (1) combining Terra and Aqua snow cover images in a same day, (2) short-term temporal filter, (3) elevation filter, (4) neighborhood spatial filter, and (5) long-term temporal filter. Fig. 1 illustrates the flow of cloud polluted images through the filters and the functions applied by the filters. The importance of using the filters is that they provide the necessary information about snow boundaries in order for the VI algorithm to be applied.

The first filter implies an assumption that no snowmelt or snowfall occurred within two observations of MODIS in one day. Thus, as long as one satellite views a pixel as snow (or land), this ground status will be assigned to the pixel in the combined image. The formula is given as follows in equation (2):

$$S_{(x,y,t)} = \max(S_{(x,y,t)}^A, S_{(x,y,t)}^T) \quad (2)$$

Where x and y are spatial (i.e. longitude and latitude) coordinates of pixel S ; t is the day index of pixel S . S^A and S^T represent pixels from Aqua and Terra respectively. The second filter assigns the cloud grids as snow (or land) if the cloud covered pixel showed both snow (or land) in the preceding and succeeding days. This step is formulated as equation (3):

$$S_{(x,y,t)} = 1 \text{ if } (S_{(x,y,t-1)} = 1 \text{ and } S_{(x,y,t+1)} = 1) \quad (3)$$

In the third filter, a maximum and a minimum elevation lines defined as the highest and lowest elevation of snow grids in the image are determined. To ensure the snow lines are correctly determined, the condition of this filter is that at least 70% of the image is cloud free¹⁷. Otherwise this filter will be skipped. The filter assigns grids with a lower elevation than the minimum elevation line $H_{min}^S(t)$ as land. Likewise, grids with a higher elevation than the maximum elevation line $H_{max}^S(t)$ are assigned as snow. The

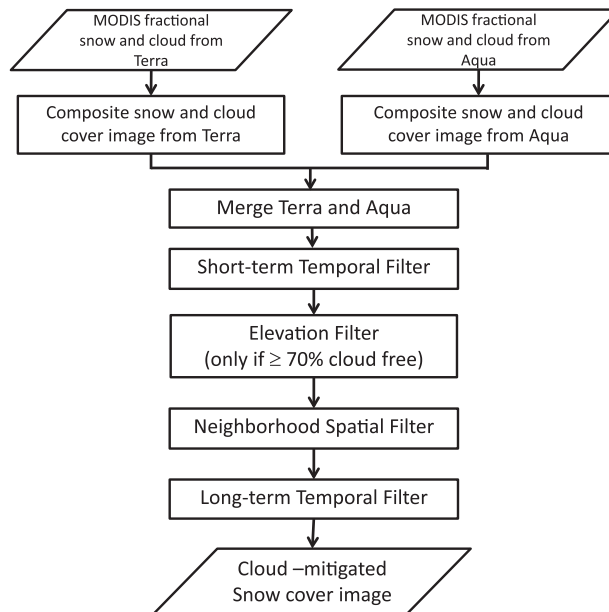


Figure 1. Flow chart of cloud mitigation filters. The process consists of five filters: (1) combining Terra and Aqua snow cover images in a same day, (2) short-term temporal filter, (3) elevation filter, (4) neighborhood spatial filter, and (5) long-term temporal filter.

formulas are given as follows in two equation (4) and (5):

$$S_{(x,y,t)} = 0 \text{ if } (H_{(x,y)} < H_{min}^S(t)) \quad (4)$$

$$S_{(x,y,t)} = 1 \text{ if } (H_{(x,y)} > H_{max}^S(t)) \quad (5)$$

where $H_{(x,y)}$ is the elevation of a pixel (x,y) location. The last two filters apply spatial and temporal processing respectively. Firstly, the fourth filter merges two neighborhood spatial filters of Gafurov and Bardossy into one. Specifically, if three out of four direct “side-bordering” pixels of the cloudy pixel indicate snow/land, the cloudy pixel will be set as snow/land. Furthermore, when considering all eight neighboring pixels, if any pixel both has lower elevation than the centre elevation and shows snow, the center pixel will also be assigned as snow. This step is formulated as equation (6):

$$S_{(x,y,t)} = 1 \text{ if } (S_{(x+k,y+k,t)} = 1 \text{ and } H_{(x+k,y+k)(k \in (-1,1))} < H_{(x,y)}) \quad (6)$$

In the last filter, a new long-term temporal filter was developed based on the fact that the annual snow-status time series can be separated into three types of periods: snow, land, and transition periods. For a snow period, the grids either show snow or cloud, therefore, cloudy pixels would be assigned as snow. On the other hand, for a land period, the grids either show land or cloud, therefore, cloudy pixels would be assigned as land. The selection of length of the period is subjective, but must be long enough to avoid phase change or long-lasting cloudy periods. For this study, we chose a 30-day window period. Interested readers should refer to (Gafurov and Bardossy¹⁷) for detailed information about the filters.

Variational Interpolation algorithm

The time-varying snow cover boundaries resulted from previous filters are modeled by the VI algorithm²² using a three-dimensional implicit function formulated as:

$$f(\vec{x}) \begin{cases} > 0 & \text{inside snow cover} \\ = 0 & \text{at snow boundaries} \\ < 0 & \text{outside snow cover} \end{cases} \quad (7)$$

where $\vec{x} = (x_1 \ x_2 \ t)^T \in R^3$, x_1 and x_2 are spatial coordinates on the projection plane, and t is the time. In three spatial dimensions, implicit functions deliver relatively simple techniques to generate complicated but useful surfaces²⁷. Once the snow cover implicit surface in space and time is determined, cloud-free images from selected days can then be obtained through cross sections of the surface²¹.

One thing to note is that interpolation from implicit surfaces depends heavily on the surface smoothness. In order to apply VI for snow cover, we have to make a hypothesis about the dynamic property of snow cover boundaries. Numerous theories in Physics such as the Principle of Least Action²⁸, Principle of Least Forcing²⁹, and the Variational Principle³⁰ proved that a natural process always operates

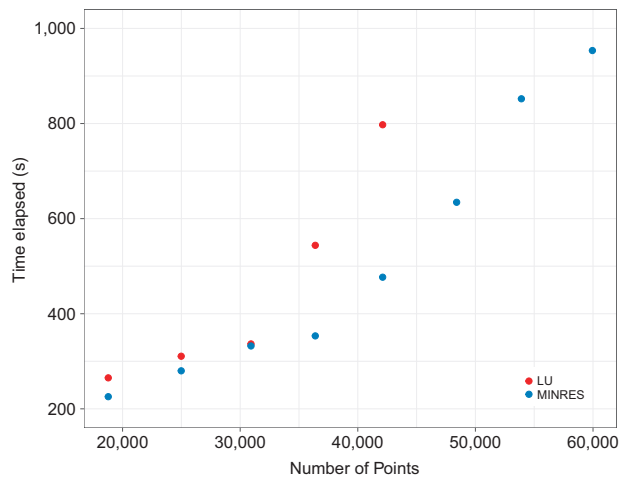


Figure 2. Results of computational efficiency comparison between the VI algorithm integrated with LU decomposition and the MINRES iteration. Computation time to solve a linear equation using LU decomposition (red) and MINRES iteration (blue) regarding the number of points collected into the equation.

in its most efficient way. As energy cost is one of the most crucial factors of efficiency, a natural surface should hold the minimum energy cost. Hence, it can be represented as a linear combination of the radial-basis function established at selected constraint points on the surface according to the following equation^{21,31}:

$$f(\vec{x}) = 0 \Rightarrow \sum_{i=1}^N w_i R(\vec{x} - \vec{x}_i) = 0 \quad (8)$$

where \mathbf{w} is a set of N weights and $R(\vec{x} - \vec{x}_i)$ is a selected radial-basis function established at N constraints points. We decided to use the thin plate function $R(\cdot) = r^2 \log r$ with $r = \|\vec{x} - \vec{x}_i\|$ to present the radial-basis function. With constraint points collected on snow boundaries in discrete times, the weights of those points can be computed by solving the linear system to create the implicit surface.

To provide VI with necessary constraint points, the Douglas-Peucker algorithm has been used³². This one parameter method is simple and widely used in vector graphics simplification and cartographic generalization. Given a relative distance dimension ϵ ($0 < \epsilon < 1$) and a starting curve of an ordered set of points, the algorithm recursively divides the curve to discard points closer than ϵ to line segments. The larger ϵ , the less points will be kept³². After experimenting a wide range of ϵ , a value of $\epsilon = 0.2$ has been chosen to preserve shapes of snow boundaries and reduce the number of points fed into the VI algorithm.

Improving the system stability for the VI algorithm

The performance of the original VI algorithm²¹ is unstable when a massive amount of constraint points is collected. For example, if constraint points were collected for the CONUS region in January 2009 (i.e., ~130,000 points), the original linear system (3) will become singular and the whole system will break down. A proposed solution for this problem is using MINRES algorithm by Paige and Saunders²³. MINRES is a Krylov subspace method for solving large symmetric systems. When applied to an inconsistent system (i.e., a singular symmetric linear problem), Paige *et al.*³³ and Choi³⁴ reported that MINRES maintains the system stability and provides a least-squares solution.

Paige *et al.*³³ analyzed the convergence behavior of the MINRES method in singular systems and concluded its residual monotonically decreases toward the origin satisfying several convergence properties. Interested readers should refer to Paige *et al.*³³.

In our implementation of the MINRES algorithm, to ensure both the system stability and accuracy, besides setting the tolerance threshold of $1e-08$ (i.e. the smaller the threshold is, the more reliable the results become), we also specified the maximum number of iterations as 1000. Hence, when either of these conditions was reached, the algorithm terminated. In a small experiment, we compared the system stability of VI using both traditional LU decomposition and the MINRES method over the CONUS region in January 2009. The performance was measured by the time elapsed over the increment of the interpolation period. Both methods started by using five consecutive days as a calculation unit. However, as more points were collected, the ordinary LU decomposition resulted in a system broke down (Fig. 2). In contrast, the MINRES method demonstrated its superiority by maintaining the system stability and accelerating the computation time. It is worth noting that, in all singular cases, the MINRES terminated

Characteristic	Variables
Data type	Binary data (0 - for ground and 1 - for snow)
Data format	NetCDF
Spatial Coverage	23.90°N–50.30°N; 130.75°W–63.95°W
Spatial Resolution	0.05° × 0.05°
Temporal Coverage	March 10, 2000 to February 28, 2017
Temporal Resolution	1 day

Table 1. Descriptive characteristics of the cloud-free snow cover dataset.

before reaching the maximum number of iterations. This termination indicated that the system accuracy was also guaranteed.

Creating the cloud-free dataset

The cloud-free dataset was developed by applying the mitigating filters and the VI algorithm introduced in the previous subsections. Regarding the use of VI algorithm, wide ranges of interpolation periods were examined to obtain the optimum results in terms of accuracy; a period of 30 days has been selected. This interpolation period is used in a moving window approach such that the window is centered around the day under consideration in order to efficiently utilize information about snow boundaries. The interpolation period started from February 24, 2000; the whole dataset was created in more than one month, using the High Performance Clusters (HPC) of the University of California, Irvine.

Code availability

The code is developed using Python 2.7 programming language; specific toolboxes used for code development include image processing toolbox (<http://scikit-image.org/>) and scientific computing toolbox (<https://www.scipy.org/>).

Due to the immense computation and the required memory size, the code was developed to be implemented in High Performance Clusters (HPC) of the University of California, Irvine. Therefore, sharing the code is of limited practical usage. However, interested readers who plan to develop cloud-free snow product using the approach described in this article can contact the corresponding author. The corresponding author will help in the adaptation of code to other computing systems.

Moreover, we are planning to build a complete package that fully utilizes the code. Any update can be found on the corresponding author's github page (<https://github.com/hoangtv1899>).

Data Records

The cloud-free snow cover dataset is available to the public through an unrestricted data repository hosted by Figshare (Data Citation 2).

The characteristics of the dataset are provided in Table 1. It should be noted that there is a number of 12, 17, and 9 days missed in the years 2000, 2001, and 2002 respectively due to the unavailability in the original MODIS record.

Technical Validation

In this study, Landsat 7 ETM+ was used as a baseline to validate the snow cover dataset spatial continuity since Landsat has high spatial resolution (30 m) and full coverage of the dataset time range (2000 to 2017). To avoid a potential saturation of ETM+ visible bands¹⁵, only Landsat 7 Tier 1 images have been utilized. Data from October to March of each year with a cloud threshold of less than 15% was obtained from the U.S. Geological Survey website (<https://landsat.usgs.gov>). Landsat 7 Tier 1 product ensures the highest available data quality for time-series processing analysis (LEDAPS³⁵). Tier 1 Landsat data has RMSE no greater than 12 m and can be considered consistent and inter-calibrated across the full collection (<https://landsat.usgs.gov>).

In order to map snow cover from Landsat images, the SNOWMAP algorithm^{26,36} was used. It is based on the NDSI index which is calculated for Landsat using band 2 and band 5:

$$NDSI = \frac{\text{Band2} - \text{Band 5}}{\text{Band 2} + \text{band 5}} \quad (9)$$

When the NDSI is greater than or equal to 0.4 and band 4 reflectance value is greater than 11%, the pixel is classified as snow³⁷. The resolution of the Landsat snow cover is then up-scaled into 0.05° and reprojected into the geographic coordinate system with the spheroid of WGS84 to match with the MODIS-SCA product resolution and projection. The evaluation was conducted in four regions with different climate condition, elevation, and land cover, namely, the Seattle region, the Minneapolis region,

Region	Landsat no. and path/row	Date	POD	FAR	Initial cloud ratio in MODIS
Seattle	L7: 46/27	03/28/2004	0.886	0.280	0.035
		10/04/2009	0.896	0.130	0.012
		03/24/2014	0.886	0.236	0.133
		03/29/2016	0.882	0.280	0.027
Minneapolis	L7: 27/29	01/10/2001	1.000	0.130	0.063
		02/27/2001	0.999	0.019	0.285
		03/15/2001	0.988	0.005	0.762
		02/17/2003	0.864	0.063	0.288
		12/23/2005	0.996	0.210	0.223
		01/27/2007	0.994	0.270	0.777
		01/14/2008	1.000	0.195	0.103
		01/30/2008	0.982	0.220	0.856
		12/15/2008	0.999	0.245	0.858
		02/01/2009	1.000	0.186	0.001
		01/06/2011	0.993	0.268	0.110
		03/27/2011	0.959	0.208	0.110
		03/06/2015	0.916	0.246	0.140
		12/03/2015	1.000	0.269	0.021
Rocky	L7: 35/31	01/02/2001	0.994	0.111	0.021
		03/07/2001	0.948	0.168	0.179
		01/13/2005	1.000	0.178	0.113
		12/15/2005	0.995	0.175	0.440
		12/18/2006	1.000	0.228	0.250
		01/22/2008	0.998	0.086	0.046
		03/10/2008	0.981	0.192	0.164
		12/23/2008	0.999	0.215	0.358
		03/29/2009	0.950	0.246	0.389
		11/24/2009	0.993	0.215	0.083
		12/10/2009	0.991	0.134	0.052
		02/15/2011	0.964	0.149	0.115
		02/18/2012	0.998	0.095	0.104
		01/03/2013	1.000	0.134	0.063
		12/05/2013	0.995	0.092	0.072
		01/06/2014	1.000	0.222	0.265
		03/30/2015	0.903	0.149	0.011
		12/27/2015	1.000	0.120	0.024
		02/15/2017	0.860	0.240	0.169
Sierra Nevada	L7: 43/33	03/28/2000	0.924	0.127	0.183
		02/27/2001	0.904	0.144	0.033
		03/05/2003	0.909	0.204	0.033
		03/21/2003	0.923	0.202	0.652
		03/07/2004	0.935	0.135	0.023
		03/10/2005	0.940	0.162	0.023
		02/15/2008	0.909	0.250	0.039
		03/02/2008	0.902	0.151	0.100
		03/18/2008	0.933	0.181	0.221
		03/05/2009	0.985	0.155	0.176
		03/16/2013	0.928	0.230	0.040
		03/19/2014	0.878	0.224	0.256
		03/24/2016	0.889	0.164	0.464
		0.955	0.179	0.199	

Table 2. Landsat validation for the performance of cloud-free dataset.

Sierra Nevada

03 – 21 – 2003

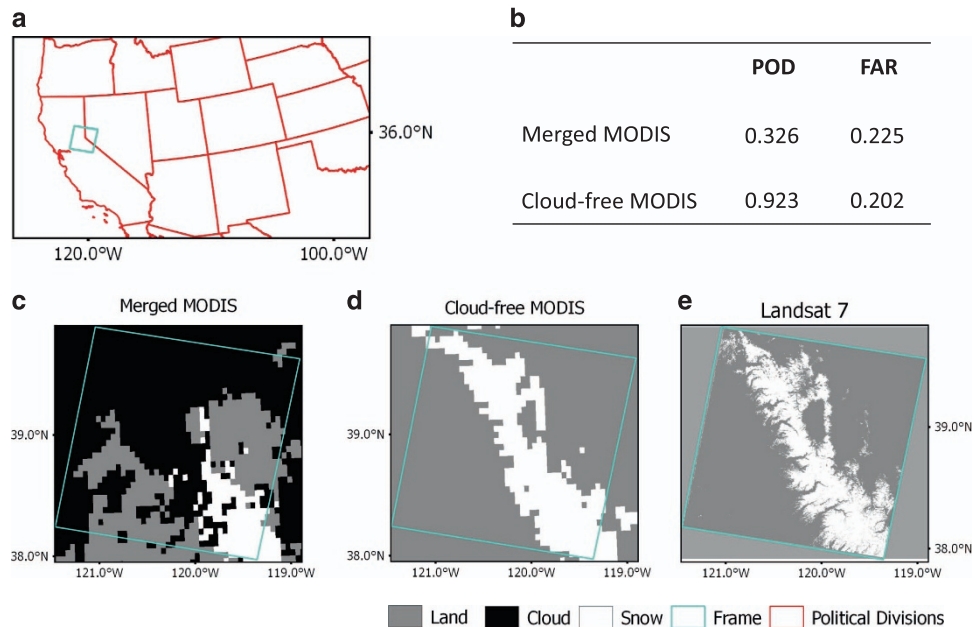


Figure 3. A validation example over the Sierra Nevada region on March 21, 2003. (a) Base map with state lines in red and a cyan rectangle to indicate the validation region. (b) Table showing Probability of Detection (POD) and False Alarm Ratio (FAR) results. (c) Merged MODIS image. (d) Cloud-free MODIS image. (e) Landsat7 image.

the Rocky Mountain, and the Sierra Nevada of California. The subsequent subsection illustrates the four regions and their characteristics.

Validation regions for snow-covered area

Landsat images from March 28, 2000 to February 15, 2017 were used for validation (See Table 2). The four selected regions for validation in this study include two high altitude regions, namely the Rocky and Sierra Nevada Mountains. The Rocky Mountains located between 40.80° and 42.73°N in latitude, 105.51° and 108.25°W in longitude. It has a high average altitude of 2500 m with a large area of grasslands. The region's snow regime is “predominantly continental with some pockets of intermountain characteristics”¹⁵. On the other hand, the Sierra Nevada Mountains (37.95° to 39.89°N, 118.84° to 121.45°W) is under great influence of maritime snow climates. Elevation ranges from 2m at the foothills to around 2800 m on the mountain (in the sample area) with equal portions of needle leaf forest, savannas, and grasslands.

Moreover, two additional regions were selected for validation. Firstly, a region around Minneapolis located between 43.62° and 45.60°N in latitude, 92.09° and 94.97°W in longitude was selected. It has an average elevation of 330 m and a land use pattern primarily consisting of cropland/natural vegetation mosaic and urban. Snow is the main form of precipitation from November through March with an annual state-wide average of 110 snow-cover days. Secondly, a region around Seattle located between 46.46° and 48.42°N in altitude, 123.32° and 123.04°W in longitude was selected. The region is known for its rainy climate with the Cascade Mountain range located on the east side, winters in this region are typically wet with significant snow accumulation in the mountain area. Land cover is mostly evergreen needle leaf forest and mixed forest.

From these four regions, we selected a final set of 50 tiles from over 150 Landsat tiles for best representing an overall snow climate. This selection is affected by a great variation in number of qualified Landsat tiles for each regions. In regions where snows occur frequently (i.e. Rocky Mountain, Sierra Nevada, and Minneapolis), there were small variations in Landsat tiles during each snow seasons, especially during snow melt periods when there were less clouds. On the contrary, in the Seattle region, there were many fewer Landsat tiles which were not contaminated by clouds but also contained significant snow boundaries.

Days	Filters	Percentage of Cloud Elimination	POD	FAR
Jan 24, 2006	Short-term temporal filter	0.071	0.985	0.016
	Elevation filter	0.021	0.983	0.017
	Spatial filter	0.105	0.981	0.017
	Long-term temporal filter	0.465	0.970	0.025
	VI algorithm	0.339	0.826	0.028
Jan 26, 2007	Short-term temporal filter	0.101	0.983	0.012
	Elevation filter	0.021	0.981	0.013
	Spatial filter	0.104	0.980	0.013
	Long-term temporal filter	0.397	0.981	0.013
	VI algorithm	0.376	0.953	0.017
Jan 31, 2009	Short-term temporal filter	0.135	0.984	0.013
	Elevation filter	0.025	0.981	0.014
	Spatial filter	0.080	0.980	0.014
	Long-term temporal filter	0.447	0.979	0.014
	VI algorithm	0.313	0.944	0.016
Mar 18, 2009	Short-term temporal filter	0.317	0.964	0.031
	Elevation filter	0.049	0.956	0.031
	Spatial filter	0.017	0.954	0.031
	Long-term temporal filter	0.530	0.949	0.031
	VI algorithm	0.087	0.867	0.035
Jan 18, 2013	Short-term temporal filter	0.234	0.988	0.010
	Elevation filter	0.037	0.988	0.010
	Spatial filter	0.050	0.986	0.011
	Long-term temporal filter	0.422	0.978	0.016
	VI algorithm	0.256	0.932	0.025
Mar 13, 2014	Short-term temporal filter	0.066	0.986	0.021
	Elevation filter	0.012	0.982	0.022
	Spatial filter	0.116	0.980	0.022
	Long-term temporal filter	0.453	0.965	0.049
	VI algorithm	0.353	0.917	0.051
Mar 6, 2015	Short-term temporal filter	0.146	0.986	0.012
	Elevation filter	0.035	0.981	0.013
	Spatial filter	0.090	0.979	0.014
	Long-term temporal filter	0.418	0.976	0.012
	VI algorithm	0.310	0.931	0.018
Mar 7, 2015	Short-term temporal filter	0.451	0.980	0.030
	Elevation filter	0.065	0.975	0.031
	Spatial filter	0.000	0.973	0.031
	Long-term temporal filter	0.377	0.971	0.028
	VI algorithm	0.107	0.933	0.029
Nov 23, 2015	Short-term temporal filter	0.107	0.985	0.013
	Elevation filter	0.020	0.980	0.014
	Spatial filter	0.094	0.979	0.014
	Long-term temporal filter	0.416	0.973	0.056
	VI algorithm	0.364	0.863	0.063

Table 3. Cross validation for each filter.

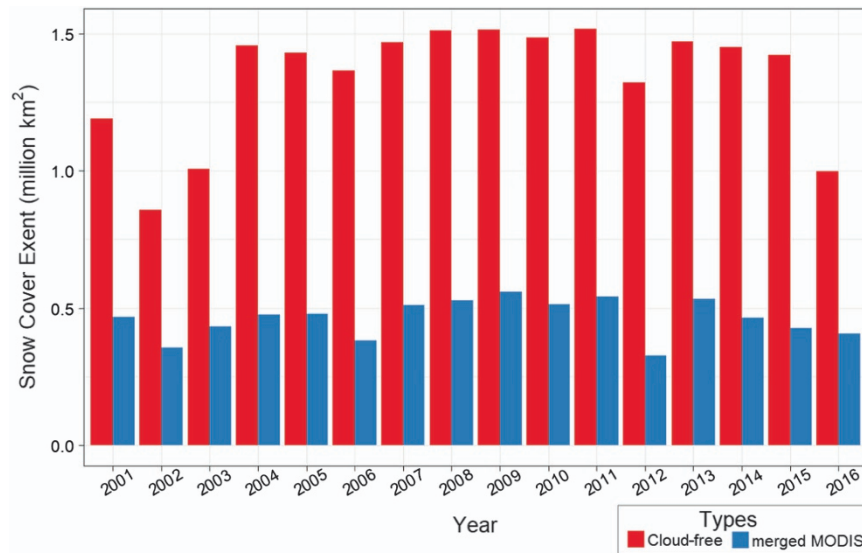


Figure 4. Comparison of annual average snow cover extent area over the CONUS computed from merged MODIS and the cloud-free snow cover product. The figure compares the annual average snow cover extent area over CONUS (in million km²) as computed from (a) the cloud-free snow cover product developed in this study (red bars) and (b) merged MODIS product (blue bars). The time period of the comparison is from 2001 to 2016.

Validation results

We compared the performance of the cloud-free with Landsat using two categorical validation indices, Probability of Detection (POD) and False Alarm Ratio (FAR).

$$POD = \frac{Hit}{Hit + Miss} \quad (10)$$

$$FAR = \frac{False}{Hit + False} \quad (11)$$

Over the validation scenarios, POD ranged from 0.860 to 1.000 with an average of 0.955. Modest results of POD with a mean of 0.888 came from the Seattle region since this area has a complex topography and dense forests which hindered the original MODIS-SCA product snow detection. Meanwhile, regions in high elevations or frequent snow areas showed high POD. Since the VI algorithm retrieved ground states of cloud hindered pixels, it is also important to validate the FAR of the cloud-free dataset. Across 50 validation scenes, the dataset yielded a reasonable average FAR of 0.179 with the highest FAR of 0.28 for 2 days, March 28, 2004 and March 29, 2016, in the Seattle region. The dataset modest performance in the Seattle region is justified by the rapid-varied topography and dense needle leaf forests in this region. These characteristics impose difficulties in mapping snow for this region from satellites^{5,9,26,38}.

In Table 2, we also computed the percentage of cloud in the combined MODIS images from Terra and Aqua for each region to demonstrate the effect of the VI algorithm. In general, since we only selected Landsat images with no or low cloud percentage, the corresponding merged MODIS images are likely to contain less cloud (Table 2). However, days when cloud cover polluted the merged MODIS images heavily (e.g., January 30 and December 15, 2008 in Minneapolis or March 21, 2003 in Sierra Nevada), the VI algorithm effectively recovered snow boundaries to correlate well with the Landsat snow-cover maps. Figure 3 shows an example of snow boundary recovered from cloud cover using VI to match with the Landsat image.

Bootstrap testing

In order to evaluate the accuracy of the mitigated filters and the VI algorithm, a cross-validation method is used. The validation process consists of three steps. Firstly, a record of synthetic cloud-covered images was developed. This was performed by selecting combined MODIS snow images over CONUS that satisfies two conditions, namely 20% or less cloud cover area and 4% or more snow cover area. Subsequently, these images were overlaid by cloud cover extracted from MODIS images with a cloud pollution rate higher than 80% to create the synthetic record. Secondly, each filter and the VI algorithm was applied sequentially to remove clouds from images in the synthetic record. Thirdly, the resulting image after each step was evaluated using the previously selected low cloud cover MODIS images. Table 3 shows the evaluation results for each filter and the VI algorithm for each day averaged across different cloud cover scenarios.

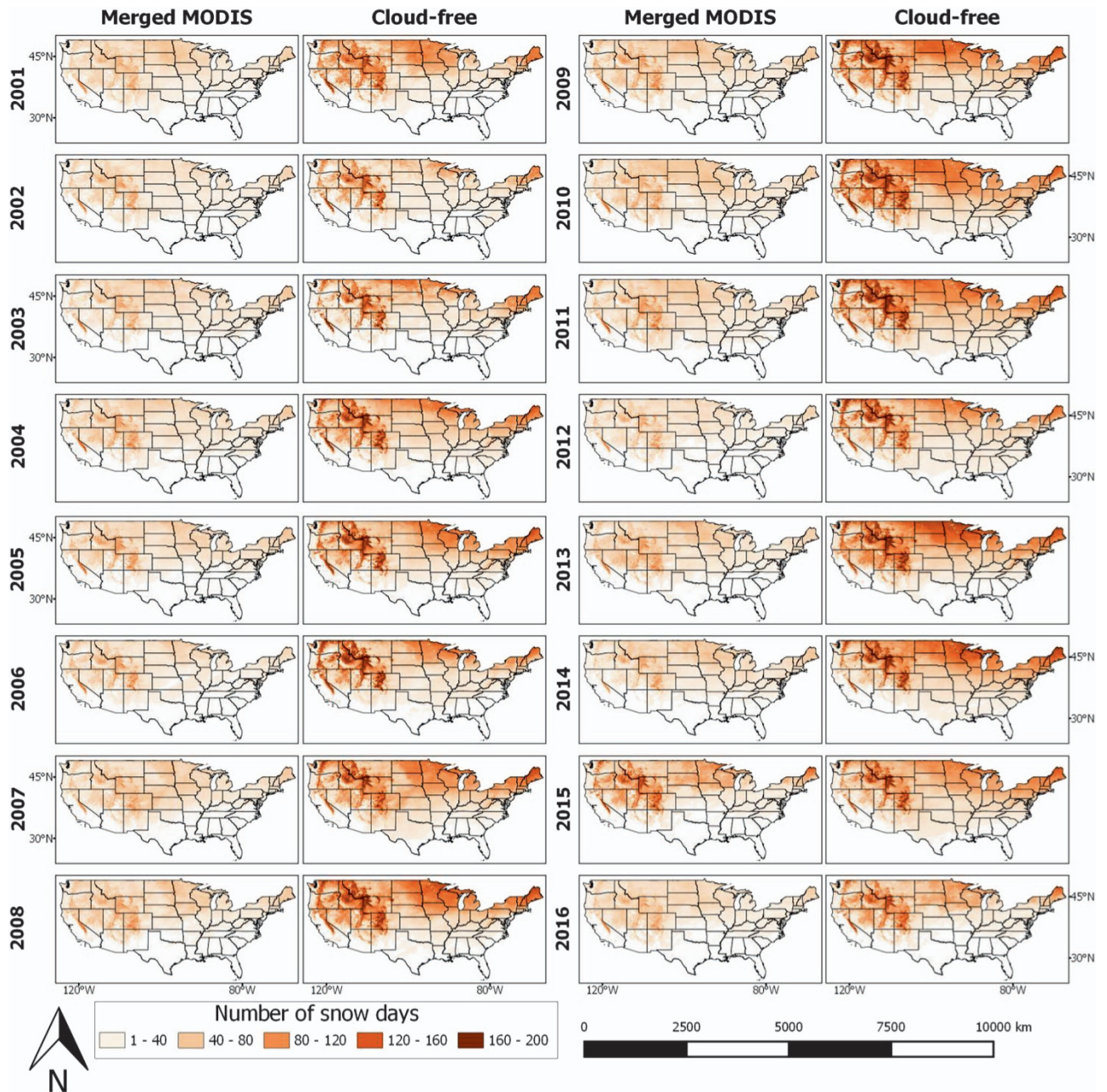


Figure 5. Comparison of the annual number of snow days over CONUS computed from merged MODIS and the cloud-free snow cover product. The figure compares the annual number of snow days over CONUS during the period 2001–2016 as computed from (a) Merged MODIS product and (b) the cloud-free snow cover product developed in this study.

The metrics used for evaluation include POD, FAR and cloud removal ratio. The results shown in Table 3 demonstrates that each filter contributes to the cloud removal while maintaining high accuracy (i. e. POD and FAR). After complete removal of clouds, the average accuracy metrics of the images are 0.907 and 0.031 for POD and FAR respectively. Furthermore, it can be seen that the VI algorithm generally is the main factor in clouds removal with an average percentage of 0.278. Despite of this high cloud removal ratio, the VI algorithm maintains high accuracy with an average change of -0.075 and $+0.014$ in POD and FAR respectively.

Usage Notes

The dataset produced in this study is useful in hydrological studies due to its adequate resolutions and validated accuracy as discussed in the previous section. In this section, we provide two simple applications of the dataset to serve as an example of the potential usages.

CONUS Snow Cover Extent

The annual average snow cover extent over CONUS, measured in million square kilometers, from both the merged MODIS and cloud-free snow cover datasets was compared. The comparison was performed from 2001 to 2016. Results show that statistics of snow cover using the merged MODIS could be substantially different from the cloud-free maps. As shown in Fig. 4, the merged MODIS images (blue) show significantly less snow cover extent than the cloud-free dataset (red). The cloud-free dataset maintains an average snow cover of 1.342 million km² in the period (2001–2016) compared to 0.462 million km² from merged MODIS. This shows that using this dataset has major implications in quantifying the amount of snow for different hydrologic processes.

Annual number of snow days

In order to examine the differences in the two datasets beyond the yearly and monthly average of snow extent, we also studied the number of snow days over CONUS. From this perspective, we could compare the snow spatial distribution of the two products. As shown in Fig. 5, during the period (2001–2016), the merged MODIS estimates 100 days as the annual number of snow days in the mountainous area of the Western US which has an altitude range of (910–1830 m). However, the cloud-free images estimate the number of snow days in the mountain states and the Sierra Nevada area as 160 to 200 days per year.

In two drought years, 2002 and 2003, there was considerably less snow days in the western mountains and the Midwest states, such as Colorado, Wyoming, and South Dakota, as illustrated by the cloud-free maps. Moreover, the images also show shorter snow seasons from the Northeast to the Northwest of the CONUS.

The cloud-free dataset provides considerably different estimates regarding both the amount of snow as well as the number of snowy days. This has considerable implications in the results of hydrologic and climate modeling studies.

References

- Hall, D. K., Riggs, G. A., Salomonson, V. V., DiGirolamo, N. E. & Bayr, K. J. MODIS snow-cover products. *Remote Sensing of Environment* **83**, 181–194 (2002).
- Maurer, E. P., Rhoads, J. D., Dubayah, R. O. & Lettenmaier, D. P. Evaluation of the snow-covered area data product from MODIS. *Hydrological Processes* **17**, 59–71 (2003).
- Parajka, J. & Blöschl, G. Validation of MODIS snow cover images over Austria. *Hydrology and Earth System Sciences* **10**, 679–689 (2006).
- Wang, X., Xie, H., Liang, T. & Huang, X. Comparison and validation of MODIS standard and new combination of Terra and Aqua snow cover products in northern Xinjiang, China. *Hydrological Processes* **23**(3), 419–429 (2009).
- Rittger, K., Painter, H. T. & Dozier, J. Assessment of methods for mapping snow cover from MODIS. *Advances in Water Resources* **51**, 367–380 (2013).
- Micheletty, P. D., Kinoshita, A. M. & Hogue, T. S. Application of MODIS snow cover products: wildfire impacts on snow and melt in the Sierra Nevada. *Hydrology and Earth System Sciences* **18**(11), 4601–4615 (2014).
- Rodell, M. & Houser, P. R. Updating a Land Surface Model with MODIS-Derived Snow Cover. *Journal of Hydrometeorology* **5**, 1064–1075 (2004).
- Andreadis, K. M. & Lettenmaier, D. P. Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources* **29**(6), 872–886 (2006).
- Hall, D. K. & Riggs, G. A. Accuracy assessment of the MODIS snow products. *Hydrological Processes* **21**, 1534–1547 (2007).
- Riggs, G. A., Hall, D. K. & Roman, M. O. Overview of NASA's MODIS and Visible Infrared Imaging Radiometer Suite (VIIRS) snow-cover Earth System Data Records. *Earth System Science Data* **9**, 765–777 (2017).
- Riggs, G. A. & Hall, D. K. MODIS Snow Products Collection 6 User Guide. *National Snow & Ice Data Center*, <https://nsidc.org/sites/nsidc.org/files/files/MODIS-snow-user-guide-C6.pdf> (2015).
- Cline, D. W. & Carroll, T. R. Inference of snow cover beneath obscuring clouds using optical remote sensing and a distributed snow energy and mass balance model. *Journal of Geophysical Research: Atmospheres* **104**(D16), 19631–19644 (1999).
- Barrett, A. P. *National operational hydrologic remote sensing center snow data assimilation system (SNODAS) products (digital media) at NSIDC*. NSIDC Special Report 11 National Snow & Ice Data Center, (2003).
- Molotch, N., Fassnacht, S., Bales, R. & Helfrich, S. R. Estimating the distribution of snow water equivalent and snow extent beneath cloud-cover in the Salt-Verde River basin, Arizona. *Hydrological Processes* **18**(9), 1595–1611 (2004).
- Painter, T. H. *et al.* Retrieval of subpixel snow covered area, grain size, and albedo from MODIS. *Remote Sensing of Environment* **113**, 868–879 (2009).
- Dozier, J., Painter, T. H., Rittger, K. & Frew, J. E. Time–space continuity of daily maps of fractional snow cover. *Advances in Water Resources* **31**(11), 1515–1526 (2008).
- Gafurov, A. & Bardossy, A. Cloud removal methodology from MODIS snow cover product. *Hydrology and Earth System Sciences* **13**, 1361–1373 (2009).
- Parajka, J., Pepe, M., Rampini, A., Rossi, S. & Blöschl, G. A regional snow-line method for estimating snow cover from MODIS during cloud cover. *Journal of Hydrology* **381**(3–4), 203–212 (2010).
- Hall, D., Riggs, G., Foster, J. & Kumar, S. Development and evaluation of a cloud-gap-filled. *Remote Sensing of Environment* **114**(3), 496–503 (2010).
- Dong, C. & Menzel, L. Producing cloud-free MODIS snow cover products with conditional probability interpolation and meteorological data. *Remote Sensing of Environment* **186**(1), 439–451 (2016).
- Xia, Q., Gao, X., Chu, W. & Sorooshian, S. Estimation of daily cloud-free, snow-covered areas from MODIS based on variational interpolation. *Water Resources Research* **48**(9) (2012).
- Turk, G. & O'Brien, J. F. Shape Transformation Using Variational Implicit Functions. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* 335–342 (SIGGRAPH, 1999).
- Paige, C. C. & Saunders, M. A. Solution of Sparse Indefinite Systems of Linear Equations. *SIAM Journal on Numerical Analysis* **12**(4), 617–629 (1975).
- Crane, R. G. & Anderson, M. R. Satellite discrimination of snow/cloud surfaces. *International Journal of Remote Sensing* **5**(1), 213–223 (1984).

25. Dozier, J. Spectral signature of alpine snow cover from the landsat thematic mapper. *Remote Sensing of Environment* **28**, 9–22 (1989).
26. Liu, Y., Key, J. R. & Wang, X. The influence of changes in cloud cover on recent surface temperature trends in the Arctic. *Journal of Climate* **21**(4), 705–715 (2008).
27. Gomes, A., Voiculescu, I., Jorge, J., Wyvill, B. & Galbraith, C. *Implicit Curves and Surfaces: Mathematics, Data Structures and Algorithms*. (Springer-Verlag: London, 2009).
28. De Maupertuis, P. *Accord de différentes lois de la nature qui avaient jusqu'ici paru incompatibles*. (Memoires de l'Academie des Sciences de Paris: Paris, 1744).
29. Gauss, C. F. Ueber ein allgemeines Grundgesetz der Mechanik. *Journal für die reine und angewandte Mathematik* **4**, 232–235 (1829).
30. Lanczos, C. *The Variational Principles of Mechanics*. (Dover Publications: New York, 1970).
31. Duchon, J. in *Constructive Theory of Functions of Several Variables. Lecture Notes in Mathematics* **571**, 85–100 edited by Schempp, W. & Zeller, K. (Springer, Berlin, 1977).
32. Douglas, D. & Peucker, T. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization* **10**(2), 112–122 (1973).
33. Paige, C. C., Parlett, B. N. & Van der Vorst, H. A. Approximate solutions and eigenvalue bounds from Krylov subspaces. *Numerical Linear Algebra with Applications* **2**(2), 115–133 (1995).
34. Choi, S.-C. *Iterative Methods for Singular Linear Equations and Least-Squares Problems*. PhD thesis. (Stanford University, 2006).
35. Masek, J. G. *et al.* A Landsat Surface Reflectance Dataset for North America, 1990–2000. *IEEE Geoscience and Remote Sensing Letters* **3**(1), 68–72 (2006).
36. Hall, D. K., Salomonson, V. V. & Riggs, G. A. Development of Methods for Mapping Global Snow Cover Using Moderate Resolution Imaging Spectroradiometer Data. *Remote Sensing of Environment* **54**(2), 127–140 (1995).
37. Huang, X., Liang, T., Zhang, X. & Guo, Z. Validation of MODIS snow cover products using Landsat and ground measurements during the 2001–2005 snow seasons over northern Xinjiang, China. *International Journal of Remote Sensing* **32**(1), 133–152 (2011).
38. Liu, Y., Key, J. R., Frey, R. A., Ackerman, S. A. & Menzel, P. W. Nighttime polar cloud detection with MODIS. *Remote sensing of environment* **92**(2), 181–194 (2004).

Data Citations

1. Hall, D. K. & Riggs, A. NASA National Snow and Ice Data Center DAAC <https://doi.org/10.5067/modis/myd10c1.006> (2016).
2. Tran, H. *et al.* [figshare](https://doi.org/10.6084/m9.figshare.5902381.v4) <https://doi.org/10.6084/m9.figshare.5902381.v4> (2018).

Acknowledgements

This research was partially supported by the ICIWARM of the US Army corps of Engineering, UNESCO's G-WADI program, Cooperative Institute for Climate and Satellites (CICS) program (NOAA prime award #NA14NES4320003, sub award #2014-2913-03) for OHD-NWS student fellowship, Army Research Office (award # W911NF-11-1-0422), National Science Foundation (NSF award #1331915), Department of Energy (DoE prime award # DE-IA0000018) and California Energy Commission (CEC Award #300-15-005).

Author Contributions

Hoang Tran collated input data, developed the algorithm and the dataset, managed the validation process and drafted the manuscript. Phu Nguyen instigated the study and supervised the project throughout its processes. Mohammed Ombadi audited data quality and a main contributor to the writing of the manuscript. Kuo-lin Hsu and Soroosh Sorooshian advised the project, reviewed the manuscript and provided valuable feedback and suggestions. Qing Xia developed the original algorithm and provided technical suggestions.

Additional Information

Competing interests: The authors declare no competing interests.

How to cite this article: Tran, H. *et al.* A cloud-free MODIS snow cover dataset for the contiguous United States from 2000 to 2017. *Sci. Data*. **6**:180300 doi: 10.1038/sdata.2018.300 (2019).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files made available in this article.

© The Author(s) 2019