Received: 3 December 2019 | Revised: 19 June 2020 | Accepted: 2 February 2021

DOI: 10.1002/cpe.6245

RESEARCH ARTICLE

WILEY

Assessing data change in scientific datasets

Juliane Müller[1] | Boris Faybishenko[2] | Deborah Agarwal[1] | Stephen Bailey[3] | Chongya Jiang[4] | Youngryel Ryu[4] | Craig Tull[1] | Lavanya Ramakrishnan[1]

[1]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California

[2]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California

[3]Physics Division, Lawrence Berkeley National Laboratory, Berkeley, California

[4]Seoul National University, Seoul, Republic of Korea

Correspondence
Juliane Müller, Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720.
Email: julianemueller@lbl.gov

Funding information
Department of Energy, Office of Science and Office of Advanced Scientific Computing Research, Grant/Award Number: DE-AC02-05CH11231

Summary

Scientific datasets are growing rapidly and becoming critical to next-generation scientific discoveries. The validity of scientific results relies on the quality of data used and data are often subject to change, for example, due to observation additions, quality assessments, or processing software updates. The effects of data change are not well understood and difficult to predict. Datasets are often repeatedly updated and recomputing derived data products quickly becomes time consuming and resource intensive and may in some cases not even be necessary, thus delaying scientific advance. Despite its importance, there is a lack of systematic approaches for best comparing data versions to quantify the changes, and ad-hoc or manual processes are commonly used. In this article, we propose a novel hierarchical approach for analyzing data changes, including real-time (online) and offline analyses. We employ a variety of fast-to-compute numerical analyses, graphical data change representations, and more resource-intensive recomputations of a subset of the data product. We illustrate the application of our approach using three scientific diverse use cases, namely, satellite, cosmological, and x-ray data. The results show that a variety of data change metrics should be employed to enable a comprehensive representation and qualitative evaluation of data changes.

KEYWORDS

data management, data versions, hierarchical data change analysis, QA/QC, scientific data change analysis

1 | INTRODUCTION

Next-generation scientific discoveries are increasingly relying on processing of data from experiments, observations and simulations. The validity of scientific results relies on the quality of data. However, many scientific communities experience "data change."[1] Data are often published in the form of versions. A new version of a dataset may mean, for example, that new entries have been added to the dataset (time series or survey data), a new processing software has been used for quality assessment and control of the data, or the settings of a measurement device have been found to be incorrect and the data taken from the device and previously published had to be corrected.[2]

The need to take into account data changes can have huge implications on compute resources and researcher's time for reprocessing, storage requirements for managing multiple versions of the data, and the science results obtained from using the data. For example, in the environmental sciences, satellite data from the moderate resolution imaging spectroradiometer (MODIS) is used to calculate evapotranspiration (ET).[3] ET is an

important parameter of the water balance equation, and its evaluation is critical for improving the accuracy of climate models. For a *skillful* user who is already familiar with using the ET calculation software, it may take 1.5–2 months (including time to download, check, and prepare all the required data). For example, the solar radiation computation, needed for computing ET, is the most compute resource intensive part, that can take up to 1 month to complete if massively parallel processing is available. This processing time is expected to increase in the future, as more data are being collected. Moreover, as the provided datasets may change (e.g., different resolution), the numerical codes have to be updated, leading to delays for reprocessing and perhaps multiple iterations to verify whether the updated software works as intended.

When a new version of the data is released, the user has to decide whether it is necessary to repeat the whole process of downloading and analyzing the data, and on what grounds this decision should be based. Data users often do not have the suitable tools at hand to quantify the data change, and detailed information on how the data were changed by the data producer is often not provided. The commonly accepted approach is to redo all computations with the new dataset. Reprocessing is computationally expensive and it requires a significant amount of storage space for the new sets of data and derived products, and, perhaps, it may not even be necessary if the changes in generated data are insignificant. In the absence of frameworks and best practices, the process of data change analyses, which is growing to be important for scientific data, is largely ad-hoc.

The goal of our research is to draw attention to the importance of data change and its impact throughout the sciences and to develop a hierarchy of data change metrics that allow the user to systematically increase the detail of their change analysis to narrow into aspects that are important for their specific application. In previous work, our team has developed DacMan,[4] a data change management tool for scientific datasets. DacMan uses file system information to detect data that has changed and provides a plug-in framework for custom domain-specific data change analyses. Our project's user research study[5] has identified user's perceptions on data change. Based on our experiences, our goal is to improve the understanding around data change analysis and provide a guiding framework that can help other domains. Specifically, in this article, we address the following questions that aim at helping the user to select suitable data change metrics and representations to understand its impact: (1) What are suitable data change metrics and how much time is available for computing them? (2) What are the user's expectations for data change presentation, including how have the data been changed (addition/deletion of data, numerical changes)? (3) What are the user's objectives and what will be done with the data change information? Understanding the impact of data change is critical to data analyses pipelines, and critically needed to efficiently manage the data pipelines for faster scientific discoveries.

In this article, we examine three diverse case studies to identify suitable data change metrics for different categories of data in order to answer questions above. We analyze (i) changes of time series and survey data that are caused by retrospective updates (QA/QC) of raw data and addition of newly observed data; and (ii) changes in X-ray scattering datasets (changes between frames). Data change of type (i) is important to researchers as results obtained based on an older data version may not be correct anymore. Data may be added because, as time passes, new observations are taken or data may be inserted when data records were "rediscovered," and inserted in what used to be gaps in the time series. Type (i) changes are also important to data providers. Data changes between versions must be quantified in order to assure high data quality. In some sciences such as cosmology, a new image processing software might be used and new observation data might arrive that must be processed. An analysis of the data changes is needed to ensure that the new data version provided to users is accurate and the processing software performs as expected without introducing biases. Type (ii) changes are important for experimentalists because the changes observed during an experiment will affect the setup of the follow-on experiment.

We investigate different approaches to quantifying the data change for the time series and x-ray data. We will analyze which data change metrics are the most suitable given the user's time constraints for the analysis and their targeted use of the analysis outcomes. The goal of this study is also to identify potential pitfalls if we rely on only a single metric (numerical or graphical presentation) for assessing the data change. Based on our experiences, we derive recommendations for a hierarchical data change analysis. We use test cases from real world applications arising in three diverse science domains including earth sciences (solar radiation computation using MODIS data), light source experiments (x-ray data from a crystallization process), and cosmology (sky survey data).

The remainder of this article is organized as follows. In Section 2, we present a summary of the related work and background information on our three test cases. In Section 3, we present the results of the data change analysis for each use case with different data change metrics. In Section 4, we discuss our results of the analysis for our guiding questions (1)–(3), and we provide a hierarchy of metrics for the data change analysis. Finally, in Section 5, we present the conclusions of our work.

2 | BACKGROUND

In this section, we briefly describe related work and the science applications and datasets that are used in our analyses.

2.1 | Related work and challenges

The metrics used for computing data change are usually domain specific and error-prone depending on the application, and the associated compute time may limit the choice of metrics to analyze the change. Tools that enable a systematic analysis of changes in the data do not exist. There are considerable challenges in developing a general tool that can automatically compute all changes between two data versions. These challenges include

the data formats used (HDF, FITS, EDF, etc.), which are usually specific to the science application; the encoding of data gaps (different scalars for fill values); data types (floats, binary, strings); different approaches for matching corresponding data entries (e.g., use of row numbers versus primary keys), but also differing needs for presenting the changes to users.

Identifying *file changes* from large directories is possible and DacMan[4] is a framework that has been developed for that purpose. DacMan identifies which files have been changed and it enables the user to plug in their own data change analysis methods. DacMan does not address the change in the data, and this is the focus of our work. We will present multiple data change analysis techniques (from computing statistics of changes to graphical change representations). DacMan can be used to identify the changed files that will be the input to our analyses and our analyses can be plug-ins in the DacMan framework, allowing us to scale the analyses.

Different types of metrics that measure differences between data points and data types have been developed for domain specific applications. These metrics include the string metrics such as the Levenshtein distance,[6] the Hamming distance[7] and other types of edit distances (see Reference 8 for a survey on string matching), which can be used to quantify the difference between two strings or sequences. The Wasserstein metric[9] measures the difference between two probability distributions. In image differencing, the Hutchinson metric[10] is widely used. The accuracy of different forecasting methods has been compared by the Makridakis competitions,[11] but these competitions have also been criticized and other methods for assessing forecast accuracy have been developed; see for example Reference 12. In contrast to difference metrics, similarity metrics quantify the similarity between two objects and its values are large when the objects are similar. Similarity metrics are widely used in bioinformatics in sequence alignment.[13] When comparing data values, for example, sample or prediction values, the root mean squared error (RMSE) is frequently used.[12] It is also used in many science areas, from environmental sciences to spectroscopy and even psychology. The mean absolute error is another metric that quantifies the difference between values, and it can have advantages over the RMSE in certain use cases.[14] The analysis of variance (ANOVA) is a collection of statistical models to compare datasets and it can be used to assess, for example, whether the means of two populations are equal. A useful method used for a comparison of datasets of different sizes is the Kolmogorov-Smirnov statistical test.[15,16]

2.2 | MODIS dataset for solar radiation computations

The MODIS[17] is an instrument aboard the TERRA and AQUA satellites. It collects the land and atmospheric data from the entire Earth's surface every one to two days in 36 spectral bands at three spatial resolutions (250, 500, and 1000 m). The data are transferred to ground stations where several levels of processing take place and various data products (e.g., atmosphere, land, ocean products) are computed and published on NASA's servers. New data are added as observations are received continuously and the data's scientific quality is periodically assessed with respect to their intended performance. Thus, several data collections exist, for example, collection 4, 5, 5.1, 6, 6.1. The data are provided in hierarchical data format (HDF[18]). Scientists use the data to study the Earth's dynamics and processes. MODIS data are widely used by the community.[19-23]

In our case study, we focus on the solar radiation at the Earth's surface which is computed based on selected MODIS products. The earth's solar radiation is an intermediate product needed for computing evapotranspiration[3]—an important part of the water cycle and responsible for cloud formation and precipitation. The required MODIS products are typically downloaded to a local machine to run the required processing. Although the individual files are relatively small (few MB's), the sheer number of files that must be downloaded, processed, and included in the radiation computation for multiple years adds up to several terabytes. For reference, for a single day, the whole process from data download to the completed radiation product can take up to 75 minutes, depending on the download speed.[20] Each data product contains ~188 files per day. Daily data in collection 6 are available starting from the year 2000 (with exceptions of few days of data outages). Computing the daily radiation product for almost 15 years takes for a *skillful user who has access to massive parallel processing* 1.5–2 months to complete. Therefore, as the data are updated and a new collection is published, being able to estimate the change in the data and the resulting radiation product will likely save a huge amount of time and resources if it indicates that the results will not change significantly. On the other hand, if the indication is that the changes are significant, not much is lost by these data change exploration (few hours) considering the resources that it takes to process the entire collection.

We use data products from MODIS collections 6 and 6.1 in our case study of investigating the influence of data change on the computed radiation products. These were the two most recent public data releases that contain the data products required for the radiation computations. We want to emphasize that as we began the study of data change, collections 5.1 and 6 were the most recent collections that contained the products we need for the radiation computations. While we were working on this study, collection 6.1 has been published, and NASA no longer provides data from collection 5.1 on the server. Note that even during the course of our study, MODIS data have been changed, namely by providing an updated data version, which caused us to redo our analysis, and by discontinuing a data product, which required us to update the radiation computation software.

The process of radiation computation starts with projecting the downloaded MODIS data into a sinusoidal tiling system since the data are collected in the swath space (additional details in Reference 20, figure 1). Next, individual scientific datasets are extracted from the MODIS products and used in the radiation computation. These products include MOD03 (geolocation), MOD04_L2 (atmospheric aerosol product), MOD06_L2 (cloud data), MOD07_L2 (atmospheric profile data product), and MCD43B3 (albedo product). During the course of conducting our study, NASA discontinued to provide product MCD43B3. Thus, we repeated our analysis with the product MCD43A3 which has a finer resolution than MCD43B3 and therefore required changes in the radiation code to accommodate this change. Scientific datasets that are used in the computations include, among others, daily cloud top temperature, cloud top pressure, surface temperature, and surface pressure.

2.3 | Sloan Digital Sky Survey dataset

The Sloan Digital Sky Survey (SDSS) is a multi-filter imaging and spectroscopic redshift survey that uses a 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. SDSS creates the most detailed three-dimensional maps of the Universe, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects.[24] The data can be accessed over the internet and is provided in FITS format.[25] There are numerous scientific uses of SDSS data, for example, cosmology measurements of dark energy and basic astronomy studies of stars, galaxies, and quasars. SDSS has enabled many science discoveries such as the development of cosmological models that help to describe the history and the future of the universe,[26,27] the study of the growth of black holes,[28] the analysis of galaxies,[29] and so on. Over 7700 peer-reviewed publications in astronomy and other sciences have used and were made possible by SDSS data. These publications have received more than 376,000 citations, which stresses the importance of the SDSS data to the scientific community.

The forms of data changes and arising challenges with SDSS include detector changes mid-survey that sometimes require a different type of data processing for one time period versus another; different software versions that are used for data processing; and the need to make updates to file formats and parameterizations while still supporting backward compatibility with previously released formats. The latest data are publicly released every year and the releases are cumulative, that is, the new release includes all the sky coverage of prior releases. Thus, for a new release, all raw data are reprocessed with the latest software version. The data releases are several GB large. In our use case, we focus on the spAllfiles that contain metadata such as photometry, classification, and redshifts. For the SDSS community, assessing data change between data releases, is mostly important for quality assurance and documentation purposes. Thus, analyzing data change will give us insights into the types of changes we encounter (addition of data, changes and additions of measured properties), as well as possible bugs in the new processing software or biases that were introduced during reprocessing.

Due to its high importance to the scientific community, it must be guaranteed that the new SDSS release is of high quality and that the data are correct and not biased. To this end, we have to develop and compare different metrics that reflect all changes comprehensively.

2.4 | X-Ray scattering data from the advanced light source

Synchrotron light sources produce electromagnetic radiation from storage rings and particle accelerators. They are used in research areas such as materials science, biology, physics, chemistry, and many others. The Advanced Light Source (ALS)[30] at Lawrence Berkeley National Laboratory is a Department of Energy funded synchrotron facility that generates bright beams of ultraviolet and soft X-ray light. Every year, more than 800 refereed journal articles are published by users and staff of the ALS (e.g., References 31-33). At many ALS end-stations, a high speed camera takes images of an experiment as conditions are changed. Scientists then analyze the images in order to design follow-up experiments. It is estimated that domain scientists routinely generate 10,000's to 100,000's of images in a few days of beamtime. The sheer amount of image data that is produced during a single experiment is often too much to analyze in real-time, yet the analysis of frame-to-frame data changes impacts the real-time setup of experiments. Moreover, a large fraction of the images is not even informative as the camera is usually started well before the target phenomenon is expected to appear and it is turned off well after. Thus, in some cases, up to 80% of the captured frames may not be of interest to the scientist. A real-time data change analysis tool that identifies automatically the subsequence of images that capture the target phenomenon would significantly increase the scientists' productivity and accelerate scientific discovery as it would allow the scientist to focus their attention on the important images only.

In this use case, we are interested in defining data change metrics for image sequences that are fast to compute in order to automatically identify the most meaningful image subsequence that captures when the observation of interest happens (temporal change). Being able to analyze the data change as the images are captured would significantly reduce the amount of data that must be transferred, stored, and analyzed by the scientist, saving time and resources, and resulting in more efficient science.

3 | RESULTS

In this section, we will analyze each dataset introduced in the previous section and describe the various metrics that we used to asses the data change. The goal of this analysis is to gain an understanding of which change metrics perform best for which purpose, and then derive recommendations of which metrics should be used in a given context.

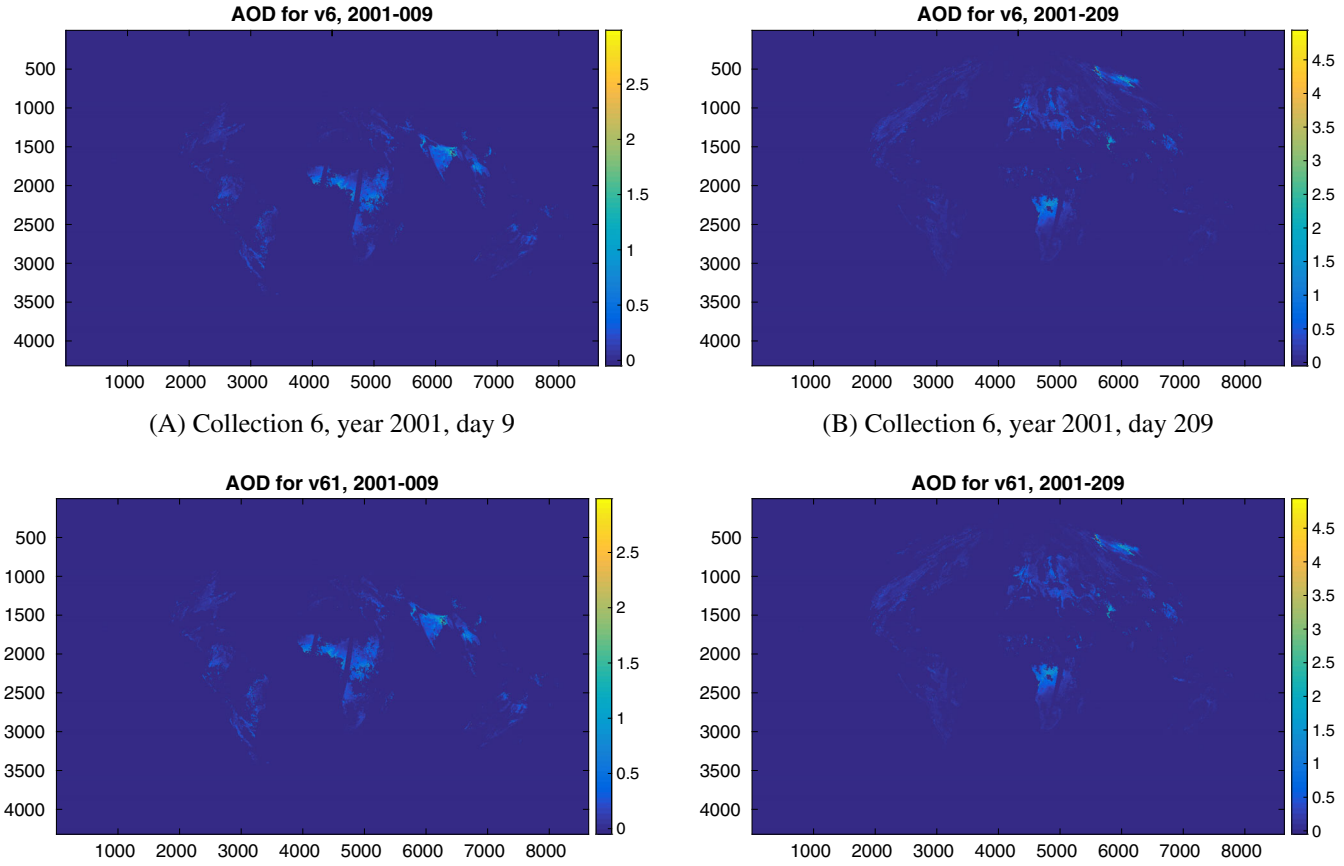
3.1 | MODIS dataset

We use MODIS collections 6 and 6.1 in this analysis together with the solar radiation products computed from these datasets. The Breathing Earth System Simulator is used[3] to compute the radiation product. It automatically downloads the required datasets form the server. Thus, every time a

new collection is published, at a minimum the URL of the collection needs to be updated in the software. In the case of switching to collection 6.1, additional parts of the code had to be changed to accommodate the fact that albedo product MCD43B3 was no longer provided and the new product MCD43A3 has a finer resolution.

In order to quantify the data change, we compute descriptive statistics between MODIS collections 6 and 6.1. These statistics include the percentage of the number of changed data entries (including entries that have fill values (Not a Number, NaN) in collection 6 and turned into a scalar in collection 6.1 and vice versa), and the mean, median, standard deviations, minimum, and maximum of the data and the data differences. We compute these data change statistics for four days in 2 years, namely Winter (day 9), Spring (day 97), Summer (day 209), and Fall (day 289) in 2001 and in 2013. The choice of days was made based on insights from our domain scientists. Seasonal data analyses is a common way scientists understand the data in this domain. Our goal with the choice of these days is to analyze if there is a correlation between the data differences and seasons and also to investigate if older year data have the same range of data change as recent year data. It is reasonable to expect that older year data experience updates (reprocessing) more often than recent year data. Moreover, the amount of data change (size of differences and number of changed entries) may potentially be higher for recent year data than for older year data due to potential sensor degradation.[2,34] We also use histograms to illustrate the differences between the data collections. This allows us to investigate if the data changes are skewed, which may indicate systematic updates in the data. We show with the help of map plots of the raw data the potential pitfalls of using only one metric to assess whether or not the data changed.

We examine two MODIS datasets more thoroughly, namely, *corrected optical depth land (AOD)* from MOD04_L2 and *cloud optical thickness (COT)* from MOD06_L2 which are used as an input for the solar radiation computation. Comparing the map plots of the AOD data for two days in 2001 for both collections shows that the differences are very subtle and hardly visible. If we only used these illustrations to judge the change of the data, we may be led to believe that the data did not change at all (Figure 1). Thus, we also use 3D histograms to analyze the data changes for the AOD product for all four days in 2001 and 2013. These histograms show how many data entries of a certain AOD value in collection 6 changed to another AOD value in collection 6.1. If no changes occurred, all bars in the histogram would lie along the diagonal from the front to the back corner of the plot (see Figures 2 and 3). For the AOD product, most data lie close to this diagonal for all days in both years, indicating that the data changes were relatively small. In particular, the data changes between collections 6 and 6.1 are smaller for the year 2001 than for 2013.



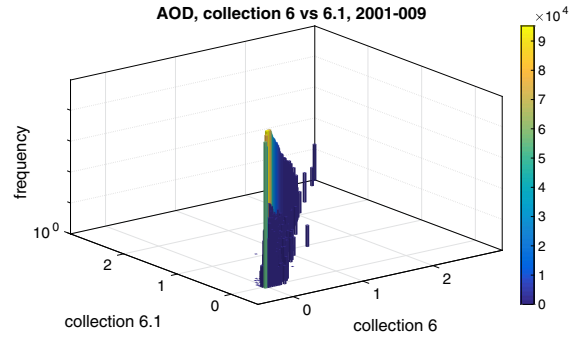
(A) Collection 6, year 2001, day 9

(B) Collection 6, year 2001, day 209

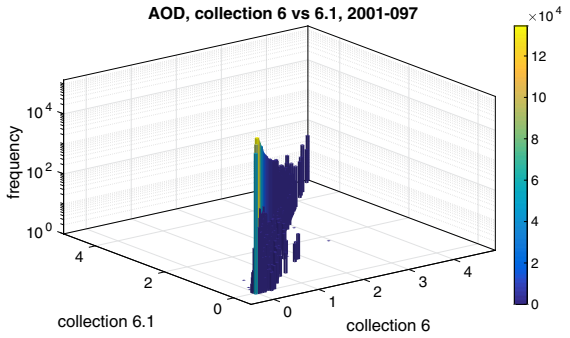
(C) Collection 6.1, year 2001, day 9

(D) Collection 6.1, year 2001, day 209

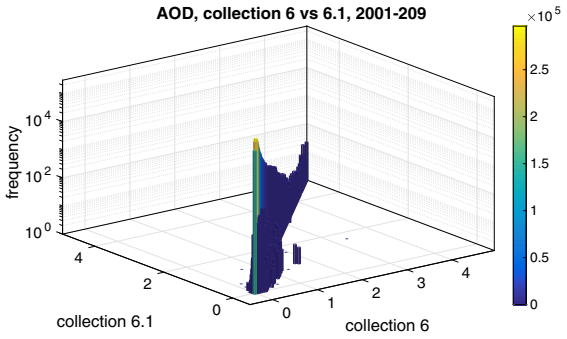
FIGURE 1 AOD data for two selected days in 2001 for both collections. Comparing the top left (right) figure to the bottom left (right) figure shows that the differences between the collection 6 and 6.1 data are hardly visible—potentially leading us to believe that the data did not change at all



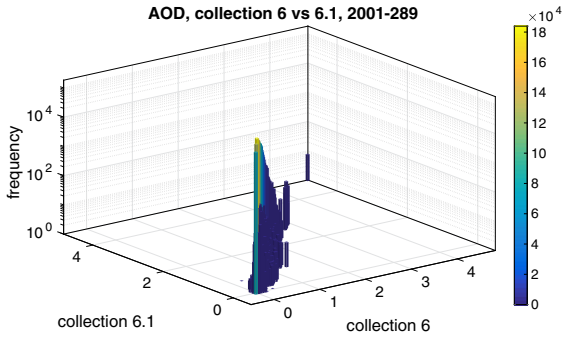
(A) Year 2001, day 9



(B) Year 2001, day 97

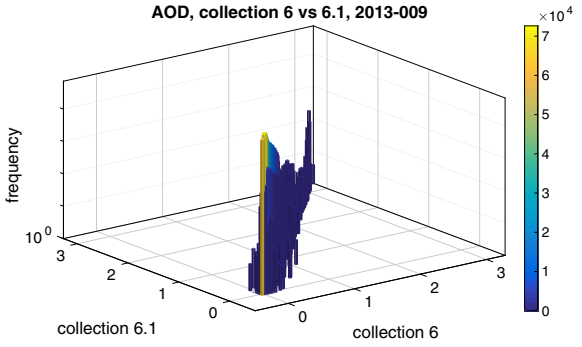


(C) Year 2001, day 209

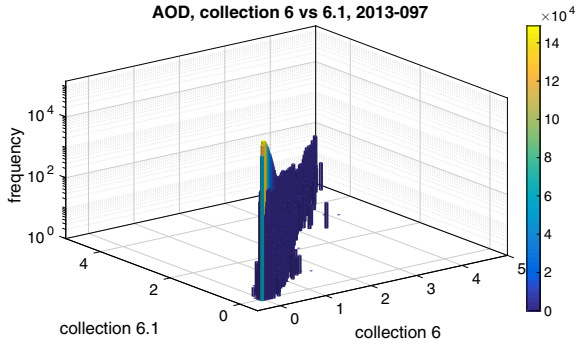


(D) Year 2001, day 289

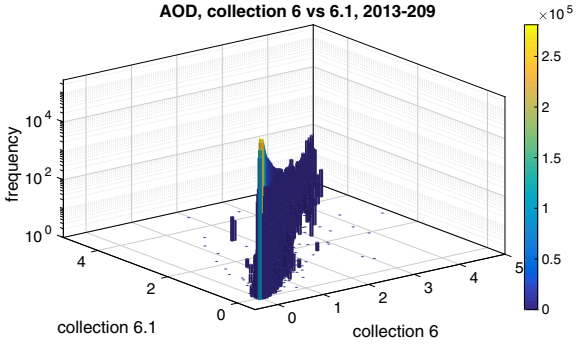
FIGURE 2 AOD data in collection 6 versus collection 6.1 for four days in 2001. The data all lie close to the diagonal from the front to the back corner, indicating only very small data changes



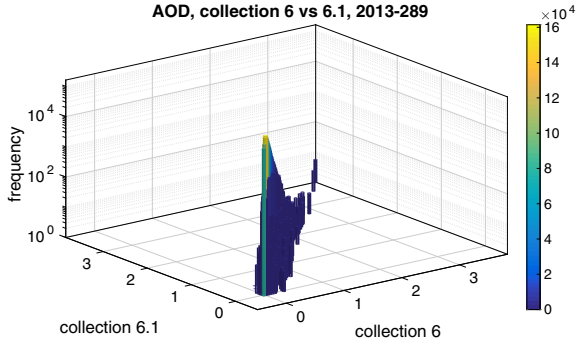
(A) Year 2013, day 9



(B) Year 2013, day 97



(C) Year 2013, day 209



(D) Year 2013, day 289

FIGURE 3 AOD data in collection 6 versus collection 6.1 for four days in 2013. The data lie close to the diagonal from the front to the back corner, indicating small data changes. Compared to 2001, however, the changes are larger (more off-diagonal data)

The statistics of the raw AOD data are similar for both collections (see Table 1). The percentage of NaN entries is computed based on the number of entries that are NaN over the total number of data entries, which is 6,226,810. Computing the statistics for the relative differences between the AOD data in collections 6 and 6.1, we find that for 2001, the data change was significantly smaller than for 2013 (less than 1% of data entries changed in 2001, more than 12% of the data entries changed for 2013). The mean and the median of the relative data change for 2001 are smaller than for 2013. In fact, the median data change from collection 6 to collection 6.1 was 0. For 2001, the number of data entries that were changed from NaN in collection 6 to a scalar value in collection 6.1 (and vice versa) were less than in 2013. The numerical values for the statistics are provided in Table 2. Our results indicate that data changes are smaller for "older years" (2001) and larger for "younger years" (2013).

Our analysis shows that for the AOD dataset, we have to consider all statistics in order to make an informed decision whether the data change is sufficiently large to be considered significant.

We conducted the same analysis we did for AOD also for the COT data. When comparing the map plots of the raw COT data for two days in 2001 and 2009 for both collections, we can barely see any differences. Again, this quick visual inspection may mislead us to believe that the data did not change from one collection to another (see Figure 4). When plotting the collection 6 versus the collection 6.1 data in 3D histograms, the data changes are much more prominent. If there were no changes, all data would lie on the diagonal from the front to the back corner. In contrast to the AOD data, the COT data are much more spread out and off-diagonal, indicating that many data entries changed. The data for 2013 changed significantly more than the data for 2001 because more values are off-diagonal for 2013. We also find that for COT many data lie around the edges, which indicates that data have been changed from their maximum value to a lower value and vice versa (see Figures 5 and 6).

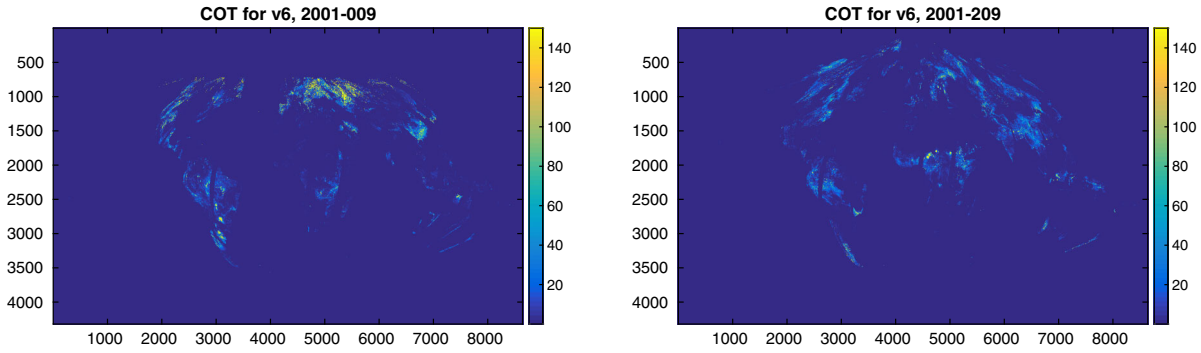
TABLE 1 Statistics for AOD data (Y-DOY-C= year - day of year - collection number; #NaN = number of data entries for which no observations exist; std = standard deviation of observations)

Y-DOY-C	%NaN	Max of obs's	Min of obs's	Mean of obs's	Median of obs's	Std of obs's
2001-009-6	83.3	2.9870	−0.0500	0.2056	0.1570	0.2093
2001-097-6	84.0	4.9170	−0.0500	0.2832	0.1700	0.3562
2001-209-6	71.83	4.9390	−0.0500	0.2023	0.1120	0.2956
2001-289-6	79.80	4.8300	−0.0500	0.2049	0.1510	0.2152
2001-009-6.1	83.3	2.9870	−0.0500	0.2052	0.1560	0.2093
2001-097-6.1	84.01	4.9180	−0.0500	0.2826	0.1700	0.3559
2001-209-6.1	71.83	4.9390	−0.0500	0.2016	0.1110	0.2951
2001-289-6.1	79.80	4.8300	−0.0500	0.2044	0.1500	0.2150
2013-009-6	85.57	3.2790	−0.0500	0.2335	0.1580	0.2732
2013-097-6	80.04	4.9910	−0.0500	0.2517	0.1630	0.3453
2013-209-6	70.13	4.9980	−0.0500	0.2404	0.1390	0.3433
2013-289-6	78.95	4.5740	−0.0500	0.1867	0.1350	0.1995
2013-009-6.1	85.52	3.2650	−0.0500	0.2306	0.1550	0.2710
2013-097-6.1	80.03	4.9860	−0.0500	0.2505	0.1620	0.3443
2013-209-6.1	70.09	4.9930	−0.0500	0.2386	0.1380	0.3420
2013-289-6.1	78.81	4.0500	−0.0500	0.1856	0.1350	0.1971

Note: Valid AOD data range is −0.05 to 5.0.

TABLE 2 Statistics for differences between AOD data in collection 6 and 6.1 (Y-DOY= year - day of year; %changed entries = number of changed entries over number of total data entries; #(NaN-6 →Sc-6.1) = number of NaN data entries in collection 6 that were changes to scalars in collection 6.1; #(Sc-6→ NaN-6.1) = number of scalar data entries in collection 6 that were changes to NaN entries in collection 6.1; rel diff's = relative differences)

Y-DOY	%Changed entries	#(NaN-6 →Sc-6.1)	#(Sc-6→ NaN-6.1)	Mean rel diff's	Median rel diff's	Stand. dev rel diff's
2001-009	0.44	128	246	0.0140	0	0.4637
2001-097	0.45	118	246	0.0107	0	0.4989
2001-209	0.79	266	356	0.0097	0	0.4342
2001-289	0.56	176	297	0.0176	0	0.6478
2013-009	12.23	5478	2031	0.0474	0.0108	0.3118
2013-097	12.24	4301	3616	0.0349	0.0035	0.5308
2013-209	12.18	4922	2712	0.0255	0	0.4617
2013-289	15.28	10314	1466	0.0407	0.0074	0.4279



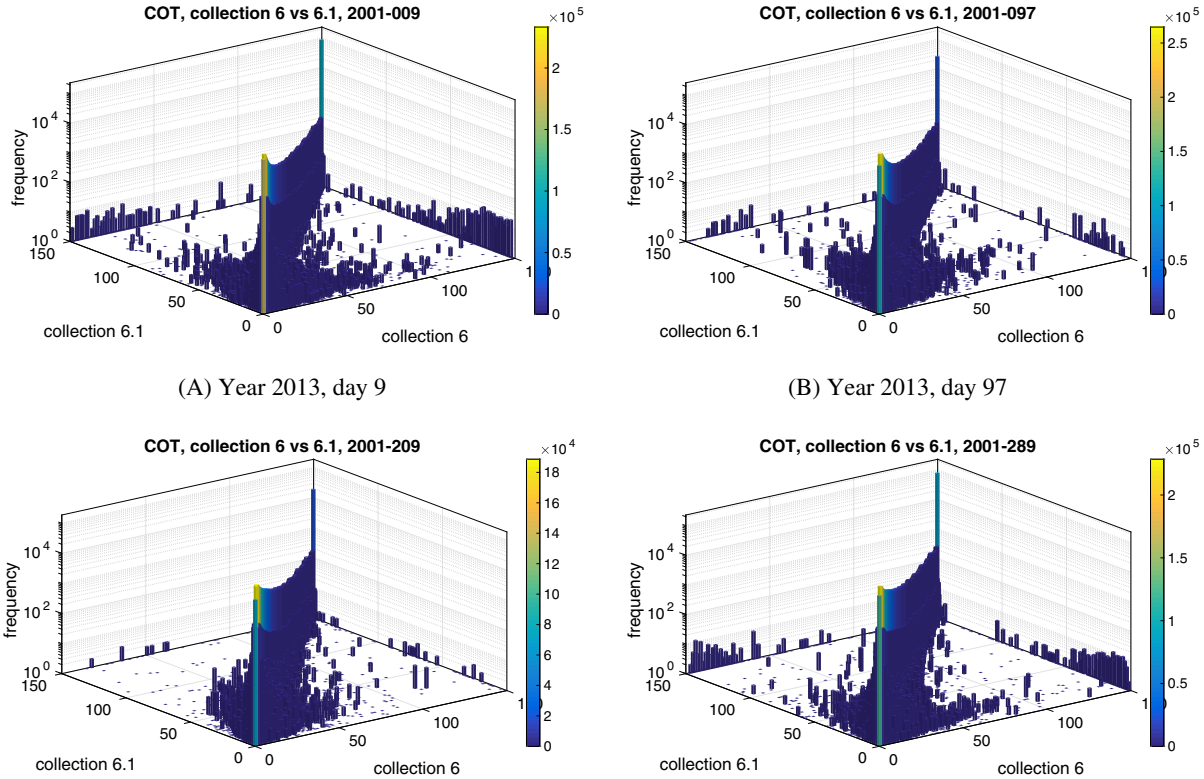
(A) Collection 6, year 2001, day 9

(B) Collection 6, year 2001, day 209

(C) Collection 6.1, year 2001, day 9

(D) Collection 6.1, year 2001, day 209

FIGURE 4 COT data for two selected days in 2001 for both collections. Comparing the top left (right) figure to the bottom left (right) figure shows that the differences between the collection 6 and 6.1 data are hardly visible—potentially leading us to believe that the data did not change



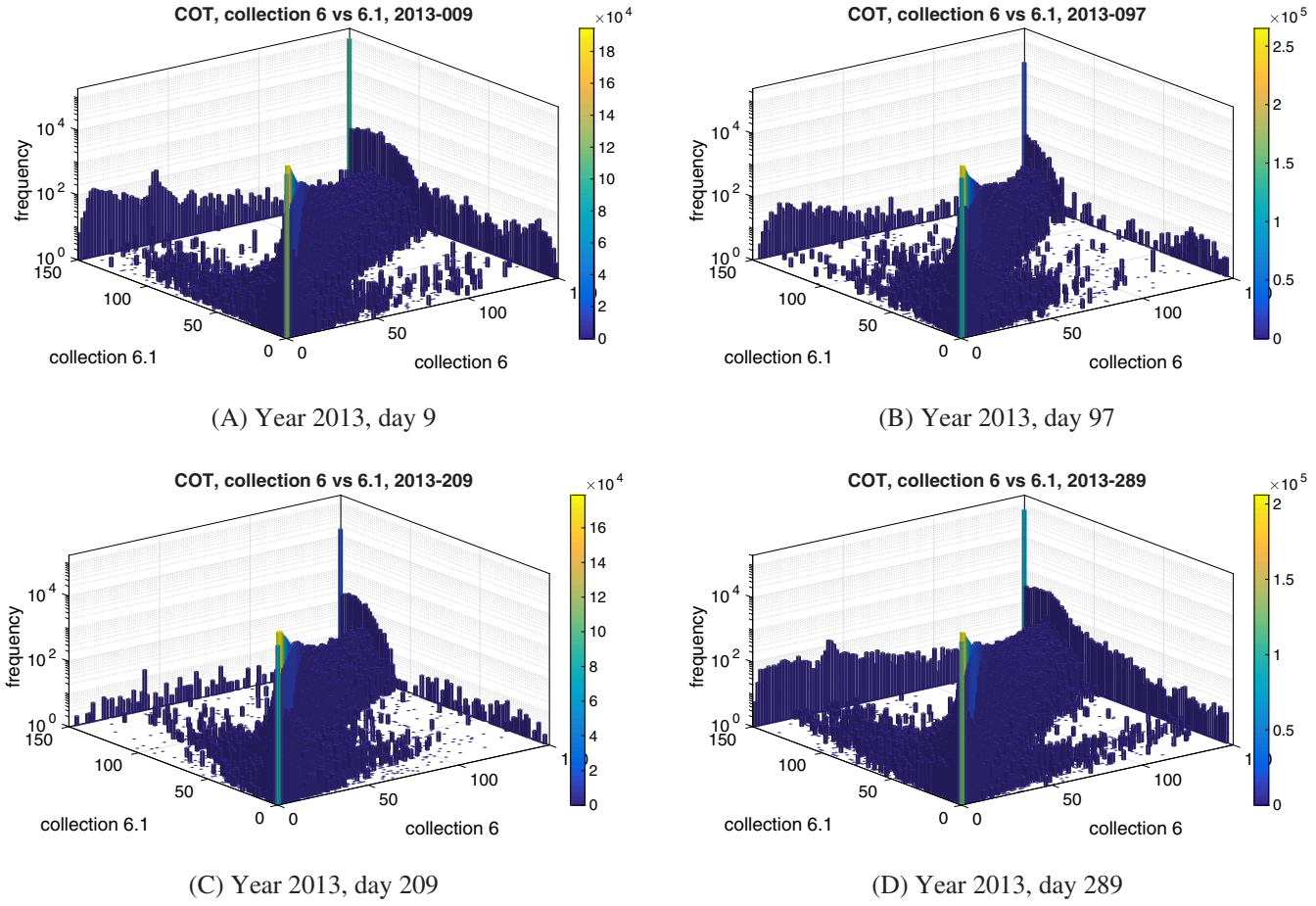
(A) Year 2013, day 9

(B) Year 2013, day 97

(C) Year 2013, day 209

(D) Year 2013, day 289

FIGURE 5 COT data in collection 6 versus collection 6.1 for 4 days in 2001. The data are more spread out and off-diagonal than for the AOD dataset, indicating more data changes



(A) Year 2013, day 9

(B) Year 2013, day 97

(C) Year 2013, day 209

(D) Year 2013, day 289

FIGURE 6 COT data in collection 6 versus collection 6.1 for 4 days in 2013. The data are significantly spread out and many data points lie off the diagonal (from front corner to back corner). This shows that many data entries were changed, in particular there were more changes for the 2013 data than for the 2001 data shown in Figure 5

We computed the statistics for the raw COT data and the relative differences. The differences between the statistics for the raw data are small and by themselves they do not indicate any major changes (Table 3). However, when we analyze the statistics of the relative changes, we see that they are generally larger for COT than for AOD with up to 5% differences in means. Compared to the data changes for AOD, the percentage of data entries that changed in the COT data from collection 6 to 6.1 is significantly larger and so is the number of data entries that changed from NaN in collection 6 to a scalar in collection 6.1 and vice versa. In contrast to the AOD data, the percentage of changed data entries is about equal for 2001 and 2013, but the magnitude of the change is larger for 2013 than for 2001. The statistics are summarized in Table 4.

For both datasets, COT and AOD, it holds that the data change for 2013 is generally larger than for 2001. The domain scientists suggest that, one explanation could be that older data (from earlier years) have gone through several more iterations of quality assurance and updates, and thus less effected by the most recent data changes. Another explanation is related to aging measurement devices aboard the satellites.[34] With time the calibration accuracy decreases, and therefore younger year data may need more correction than older year data.[2]

Finally, we investigate how the changes of the AOD and COT data propagate to the shortwave radiation product for all 4 days of both years. The computational expense associated with computing the data product is non-trivial as outlined in the introduction. Approaches such as sensitivity analyses and machine learning to identify sensitivities and correlations between changes in the datasets and the data product would require us to recompute the data product for significantly more than 8 days in order to be applicable and allow for reliable conclusions. Since our goal with this analyses was to provide a quick answer, we do not consider these methods here.

We use map plots of the solar radiation data product for a first visual inspection of the data change. Similar to the COT and the AOD data, the differences between using collection 6 and 6.1 data are hardly visible (see Figure 7). We also compute the statistics for the radiation product for all 4 days in both years. The statistics computed from the data in collection 6 and 6.1, respectively, are very similar often with changes at the second or third decimal point. The percentage of NaN values is significantly lower than for the AOD and COT products, which is related to the gap-filling

TABLE 3 Statistics for COT data (Y-DOY-C= year - day of year - collection number; #NaN = number of data entries for which no observations exist; std = standard deviation of observations)

Y-DOY-C	%NaN	Max of obs's	Min of obs's	Mean of obs's	Median of obs's	Std of obs's
2001-009-6	69.54	150	0.010	20.3619	7.4800	34.5357
2001-097-6	63.24	150	0.010	14.5610	7.7800	21.6587
2001-209-6	65.96	150	0.050	16.6098	9.7300	21.5503
2001-289-6	61.93	150	0.030	21.0181	10.1700	31.1224
2001-009-6.1	69.48	150	0.010	20.2895	7.4600	34.4651
2001-097-6.1	63.16	150	0.010	14.5208	7.7400	21.6270
2001-209-6.1	66.00	150	0.040	16.6050	9.7200	21.5472
2001-289-6.1	61.81	150	0.040	20.9111	10.1200	31.0146
2013-009-6	70.05	150	0.020	22.4792	8.8900	36.2403
2013-097-6	66.66	150	0.010	14.0876	7.2200	21.4178
2013-209-6	66.88	150	0.050	16.6051	9.4900	22.0414
2013-289-6	65.28	150	0.030	22.0401	10.020	32.9924
2013-009-6.1	70.28	150	0.030	22.0645	8.8900	35.6925
2013-097-6.1	66.83	150	0.010	14.0394	7.3000	21.1999
2013-209-6.1	67.18	150	0.030	16.2557	9.5100	21.1507
2013-289-6.1	65.52	150	0.020	21.4681	9.9900	32.0994

Note: Valid COT data range is 0 to 150.

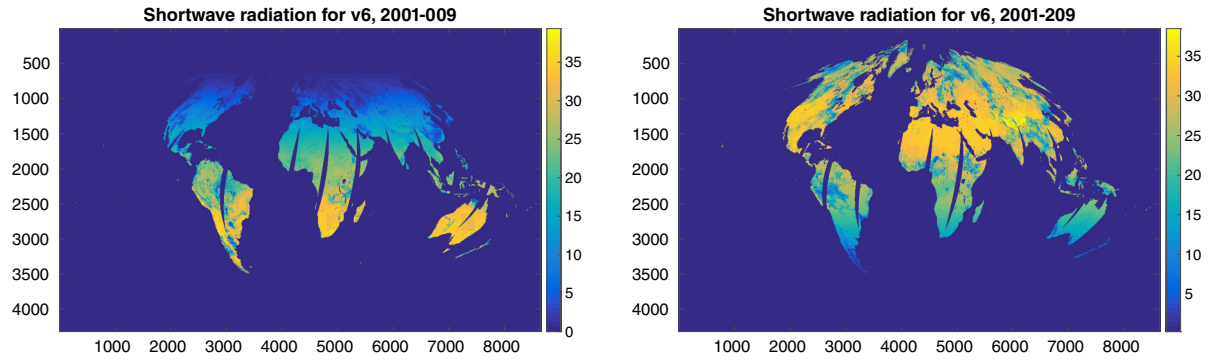
Y-DOY	%Changed entries	#(NaN-6 →Sc-6.1)	#(Sc-6→ NaN-6.1)	Mean rel diff's	Median rel diff's	Stand. dev rel diff's
2001-009	23.56	33826	29801	0.0346	0.0054	0.3672
2001-097	28.54	30940	25752	0.0273	0.0054	0.1283
2001-209	32.01	34203	36671	0.0309	0.0070	0.1369
2001-289	33.17	32134	25125	0.0263	0.0057	0.2423
2013-009	26.08	22863	37458	0.0566	0.0257	0.3626
2013-097	30.04	41840	52506	0.0461	0.0144	0.1907
2013-209	32.74	35906	54322	0.0574	0.0238	0.2069
2013-289	32.83	30694	45284	0.0637	0.0333	0.3359

TABLE 4 Statistics for differences between COT data in collection 6 and 6.1 (Y-DOY= year - day of year; %changed entries = number of changed entries over number of total data entries; #(NaN-6 →Sc-6.1) = number of NaN data entries in collection 6 that were changes to scalars in collection 6.1; #(Sc-6→ NaN-6.1) = number of scalar data entries in collection 6 that were changes to NaN entries in collection 6.1; rel diff's = relative differences)

method in the radiation computation code (see Table 5). The statistics computed for the relative differences between the radiation product using collection 6 and 6.1 show a similar pattern as for AOD and COT, that is, the changes are generally larger for the 2013 days than for the 2001 days, which is a direct result from the larger data changes observed for AOD and COT for 2013. Considering the numerical values of the data change, there does not appear to be a large difference in the radiation product when using data from collection 6 versus collection 6.1. This indicates that the method for computing the radiation product is to some extend robust to changes in the AOD and COT data (see Table 6), which can be explained by the averaging method used in the radiation computation.

The 3D-histograms for the computed solar radiation product using collection 6 versus 6.1 show that the majority of the data lie on the diagonal from the front to the back corner of the plots, indicating that the majority of the computed radiation data did not change. Similar to the AOD and COT data, the histograms have more spread around the diagonal for 2013 than for 2001, reflecting that more data changed for 2013 (see Figures 8 and 9).

Our analysis shows that in order to thoroughly analyze the data change, we should not only rely on a single metric, but instead use numerical metrics as well as image representations of the data change (such as histograms) in order to obtain a complete understanding of the magnitude of the data change and its associated importance.



(A) Collection 6, year 2001, day 9

(B) Collection 6, year 2001, day 209

(C) Collection 6.1, year 2001, day 9

(D) Collection 6.1, year 2001, day 209

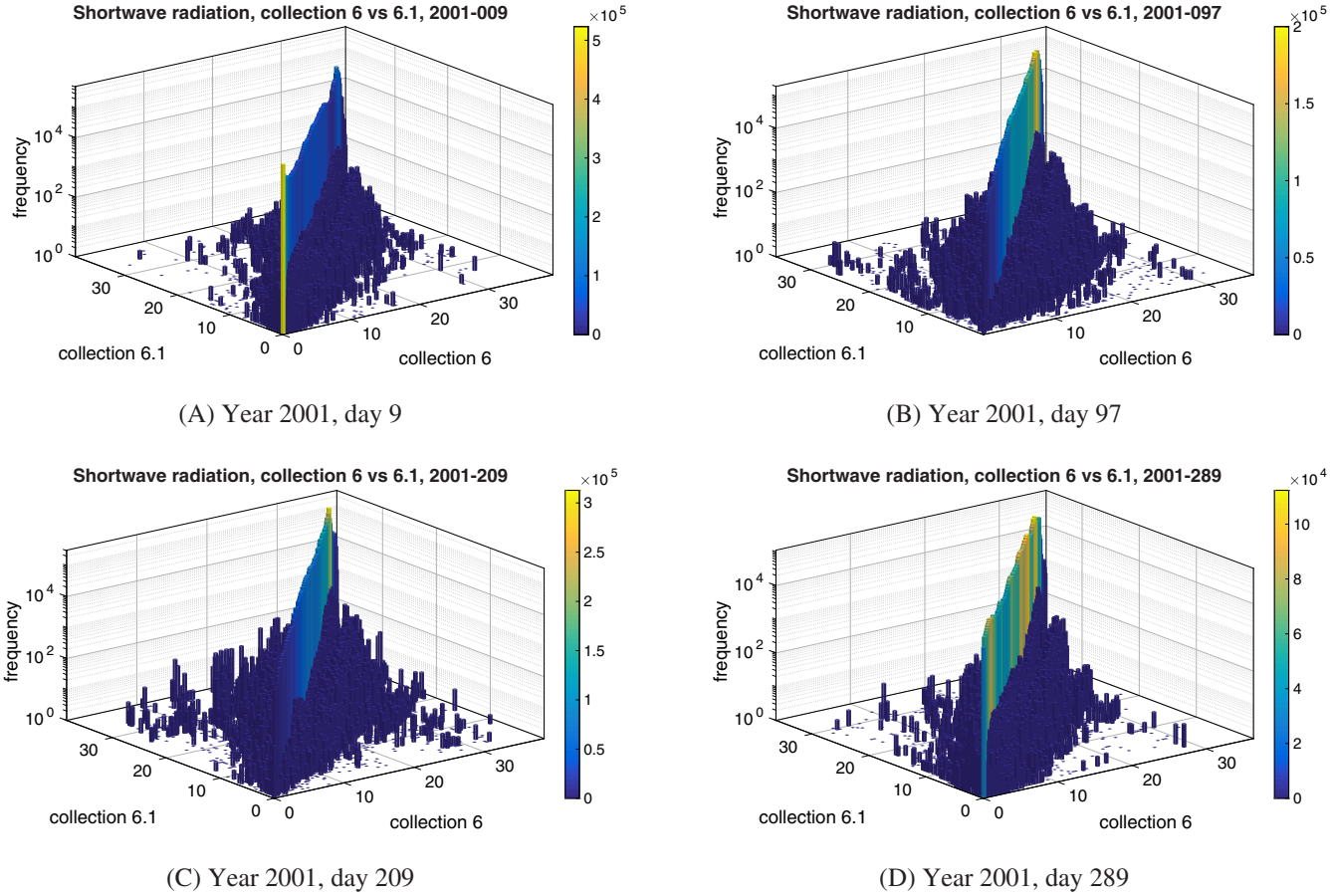
FIGURE 7 Map plots of the Earth's shortwave radiation computed with MODIS collection 6 and 6.1 for days 9 and 209 in 2001. Compare the top left (right) figure to the bottom left (right) figure. Differences in the radiation are hardly visible and might lead us to believe that there were no significant data changes

TABLE 5 Statistics for the computed shortwave radiation product (Y-DOY-C= year - day of year - collection number; #NaN = number of data entries for which no observations exist; std = standard deviation of observations)

Y-DOY-C	%NaN	Max of obs's	Min of obs's	Mean of obs's	Median of obs's	Std of obs's
2001-009-6	6.03	39.3624	0	14.3914	13.6020	11.2966
2001-097-6	6.89	35.8608	0.8284	22.2442	23.7041	7.0608
2001-209-6	6.99	38.4844	0.2616	25.3939	27.5399	7.7546
2001-289-6	7.22	36.2679	0	17.2768	18.7849	9.4521
2001-009-6.1	6.03	39.3624	0	14.3907	13.6042	11.2947
2001-097-6.1	6.89	35.8608	0.8284	22.2440	23.7018	7.0592
2001-209-6.1	6.99	38.4844	0.2616	25.3996	27.5408	7.7526
2001-289-6.1	7.22	36.2679	0	17.2749	18.7834	9.4483
2013-009-6	7.36	38.3859	0	13.6957	12.9156	10.9101
2013-097-6	6.20	35.4871	0.4070	21.9179	23.4254	6.8706
2013-209-6	6.40	38.2856	0.2093	24.6230	26.1414	8.0837
2013-289-6	6.14	36.5684	0	17.9968	19.2849	9.7748
2013-009-6.1	7.36	38.3859	0	13.7241	12.9778	10.9067
2013-097-6.1	6.19	35.5315	0.4070	21.9257	23.4258	6.8528
2013-209-6.1	6.40	38.2856	0.2093	24.6664	26.1599	8.0380
2013-289-6.1	6.14	36.5684	0	18.0289	19.3243	9.7536

Y-DOY	%Changed entries	#(NaN-6 →Sc-6.1)	#(Sc-6→ NaN-6.1)	Mean rel diff's	Median rel diff's	Stand. dev rel diff's
2001-009	2.44	162	244	0.0030	0	0.0424
2001-097	2.51	166	147	0.0029	0	0.0364
2001-209	2.97	245	134	0.0033	0	0.0370
2001-289	2.48	176	192	0.0033	0	0.0522
2013-009	5.28	164	171	0.0101	0	0.1272
2013-097	7.25	633	491	0.0074	0	0.0790
2013-209	8.20	1231	1184	0.0098	0	0.1016
2013-289	7.32	147	159	0.0125	0	0.1502

TABLE 6 Statistics for differences between radiation products computed with collection 6 and 6.1 (Y-DOY= year - day of year; %changed entries = number of changed entries over number of total data entries; #(NaN-6 →Sc-6.1) = number of NaN data entries in collection 6 that were changes to scalars in collection 6.1; #(Sc-6→ NaN-6.1) = number of scalar data entries in collection 6 that were changes to NaN entries in collection 6.1; rel diff's = relative differences)



(A) Year 2001, day 9

(B) Year 2001, day 97

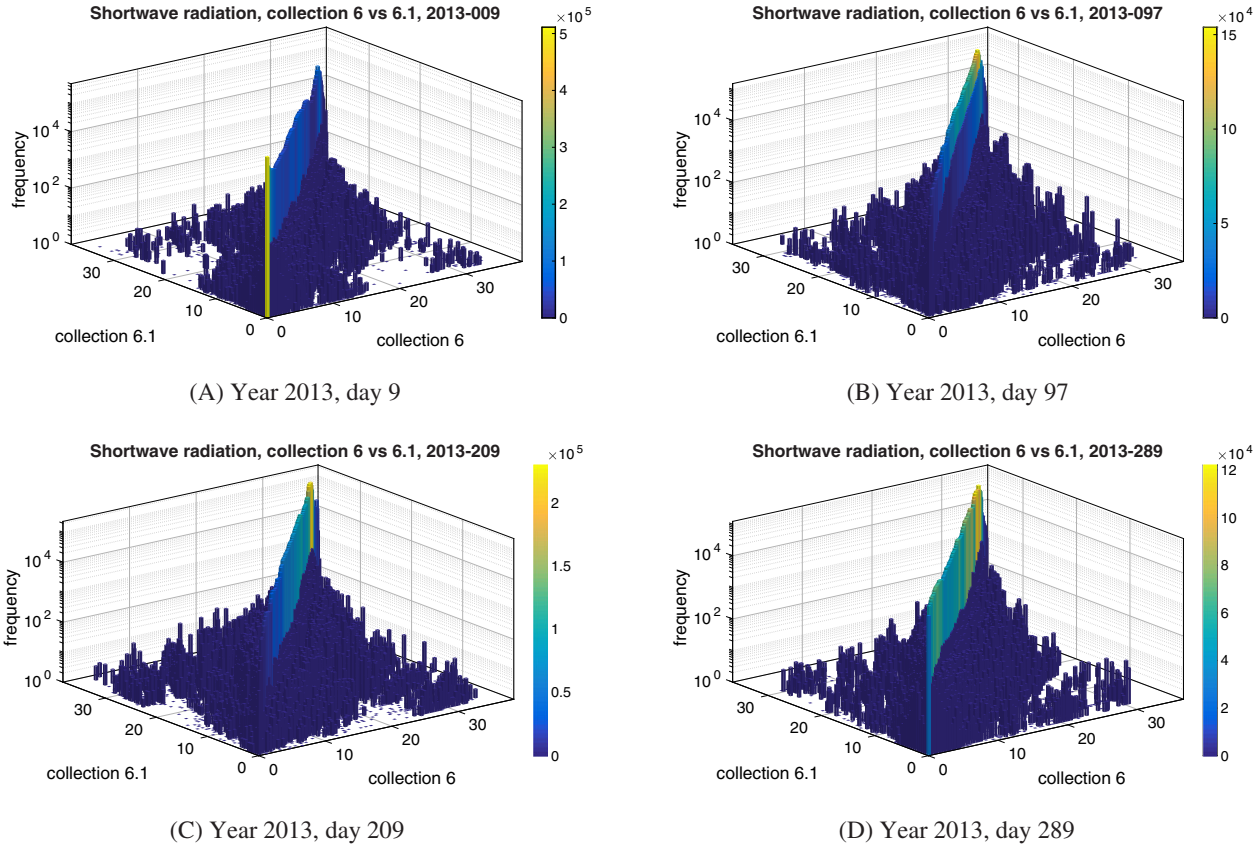
(C) Year 2001, day 209

(D) Year 2001, day 289

FIGURE 8 Solar radiation computed for 2001 with collection 6 and 6.1, respectively. The location of the bars indicates how a radiation value computed based on collection 6 changed when using data from collection 6.1 (and vice versa). If there was no data change at all, all bars would lie on the diagonal from the front corner to the back corner of the plot

3.2 | Sloan Digital Sky Survey dataset

Our data change analysis for the SDSS uses the two data realeases of SDSS-III, namely data release 11 (DR11) and data release 12 (DR12),[35] in particular the `spAll` files. These files are summaries of metadata such as photometry, classification, and redshifts and they contain all measurements, object properties, the dates of observation, and so on. DR11 contains data through summer 2013 and DR12 is the final release of SDSS-III and runs through July 2014. The data are arranged in tables in FITS format, where each row corresponds to an observed object and each column corresponds to the measured property. We use the python package astropy.io[36] to load the tables. Each time a new version of the data is released, the raw data are used and newly processed.



(A) Year 2013, day 9

(B) Year 2013, day 97

(C) Year 2013, day 209

(D) Year 2013, day 289

FIGURE 9 Solar radiation computed for 2013 with collection 6 and 6.1, respectively. The location of the bars indicates how a radiation value computed based on collection 6 changed when using data from collection 6.1 (and vice versa). If there was no data change at all, all bars would lie on the diagonal from the front corner to the back corner of the plot

Comparing the size of the tables in DR11 and DR12, we find that DR11 has 231 data columns and 2,085,000 rows (observations), whereas DR12 contains 236 columns and 2,512,000 rows. The columns added in DR12 are `EBOSS_TARGET0`, `EBOSS_TARGET1`, `EBOSS_TARGET2`, `EBOSS_TARGET_ID`, and `THING_ID_TARGETING`. A new column may be added, for example, if an additional measurement or object property previously not reported becomes of interest. Since DR12 contains an additional year of observations, it has more rows as more objects (more patches of sky) were observed. However, the additional rows in DR12 are not simply appended to the data table in DR11, but rather they are inserted, that is, the row number of an observed object is to some extent arbitrary. In order to analyze how the data changed from DR11 to DR12, we compare only those objects and properties that appear in both data releases. We use the columns `PLATE`, `MJD`, and `FIBERID` as primary keys to match the objects from both releases. Therefore, the joint table has 231 columns and 2,085,000 rows. Of these columns, 18 contain non-numerical data (strings), and 213 contain numerical values. The changes between DR11 and DR12 are mainly attributed to using a new software version for reprocessing the data.

In order to compare the non-numerical data, we count the number of rows for which an entry in DR11 was changed in DR12. For example, the data entry `v5_6_5` in DR11 in columns `RUN1D` and `RUN2D` was changed to `v5_7_0` in DR12 for all 2,085,000 rows in both columns (this is the version of the processing software). The data did not change at all for eight of the remaining non-numerical data columns (including `PROGRAMNAME`, `CHUNK`, and `PLATEQUALITY`). This simple metric allows the data publisher to assess if the intended changes (e.g., update of processing software) were made for all objects.

For the numerical data, we compute descriptive statistics such as the minimum, maximum, mean, and median of the difference between the DR11 and the DR12 data in order to assess the data changes. We also compute the number of outliers and the fraction of outliers over the total number of observations. We define as outliers all values that lie outside of 1.96 standard deviations of the mean difference. The results of these simple statistics show that 150 of the numerical data columns did not change at all. Thus, we have to investigate further data change only for the remaining 63 numerical data columns.

Except for column `SPECOBJID`, all columns with changed numerical data contain outliers. `SPECOBJID` is the ID of the optical spectroscopic objects, including encoding the version number of the input data processing that was used. All entries are shifted by the same value, −97,280, indicating a consistent change of input processing version number for all objects from DR11 to DR12. This change was made purposefully and the fact that the data change analysis reflects this consistent change for all entries assures the data publisher that the changes were made successfully. Furthermore, 49 of the 63 numerical columns have zero median change and mostly only a very small fraction of outliers (less than 1% of the data

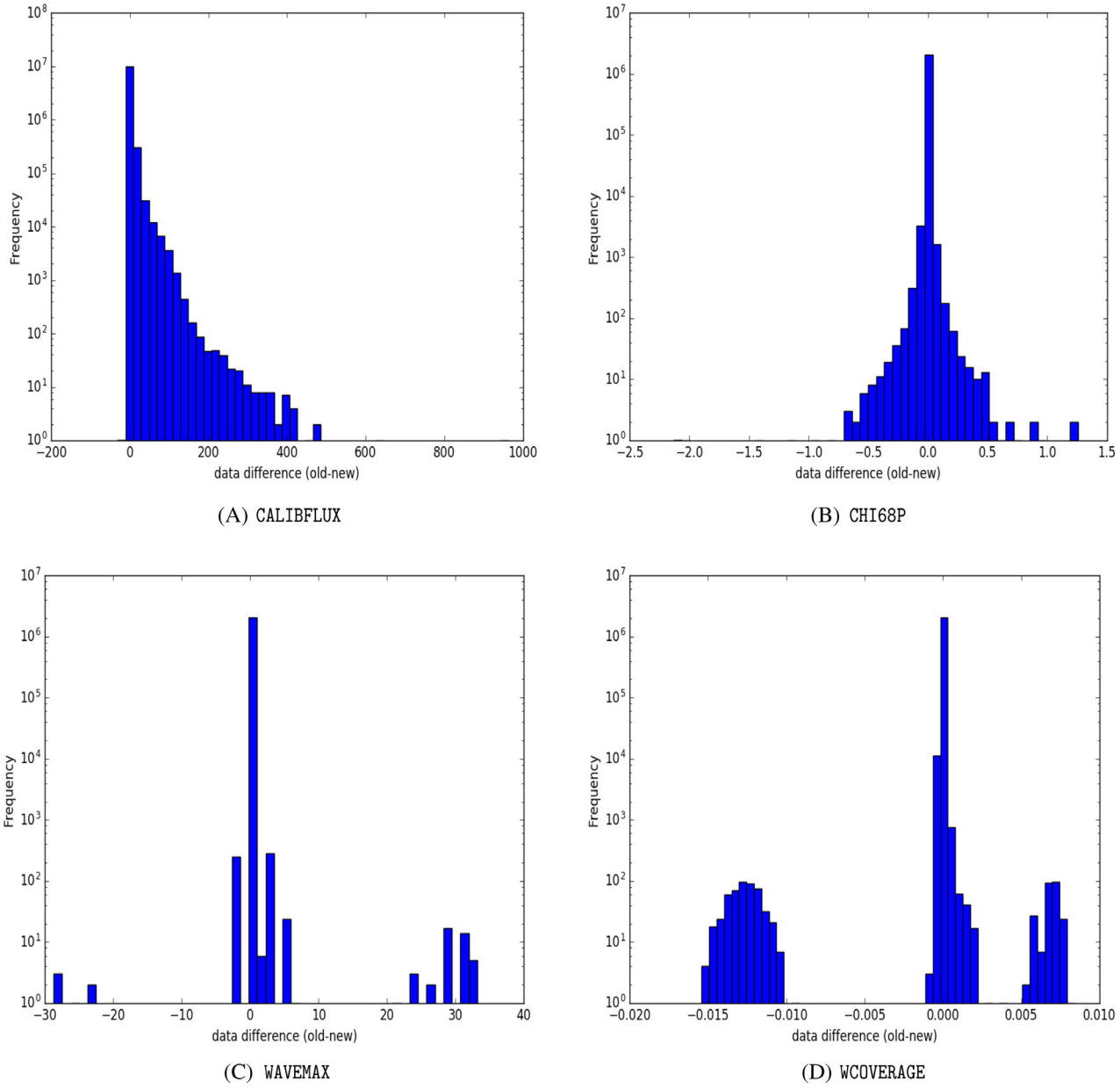


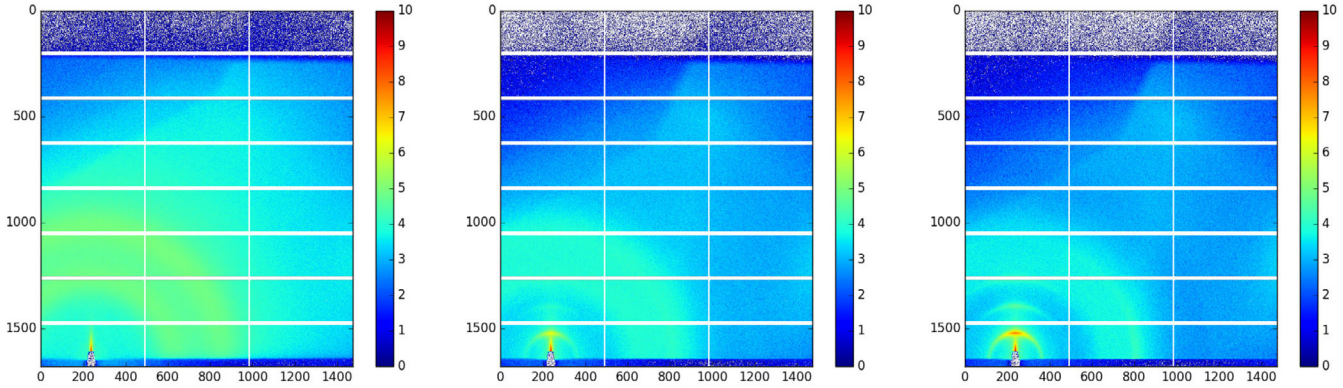
FIGURE 10 Histograms of the differences between DR11 and DR12 data. Different data columns have different distributions of data changes

differences are outliers). When outliers are detected between the data versions, it has to be ensured that they are scientifically sound and not caused by any biases or processing software bugs.

In order to enable a visual data change analysis, we also present the data differences between DR11 and DR12 in the form of histograms. These histograms show that the differences do not follow the same distribution for all data columns (see, for example, Figure 10). The histograms and the computed descriptive statistics (not shown) provide a high-level summary of the data changes caused by using a new processing software for the domain scientists. The statistics help us to figure out which data columns to focus further analyses on. The histograms can give us additional visual cues of possible problems in the reprocessed data.

3.3 | X-ray scattering dataset

In this use case, we analyze the temporal changes of two series of scattering images taken during experiments at the ALS. For each experiment, the camera captures 300 frames. The first sequence captures a crystallization process in which dendrites grow. In some cases, the changes are obvious to the human eye as shown in Figure 11. In contrast, the changes in the second sequence are much more subtle; a dark shadow moves through the

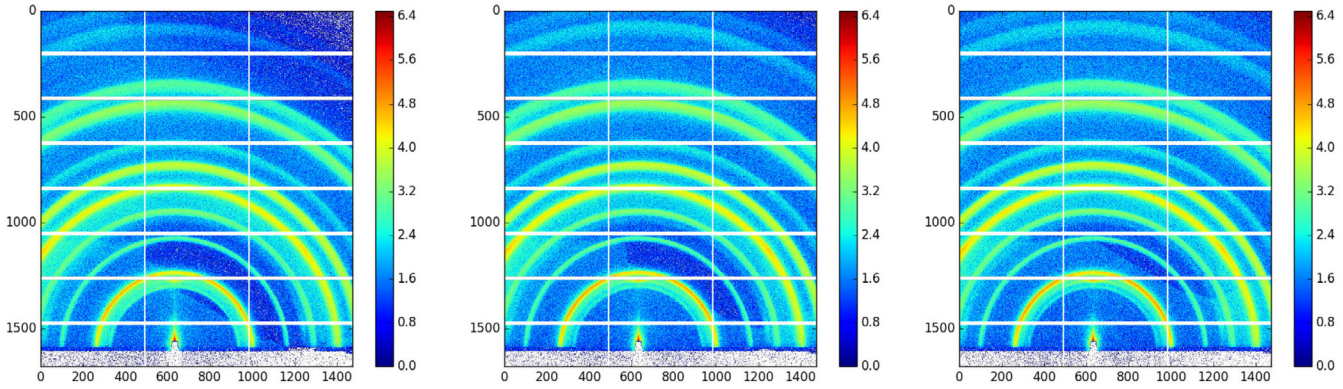


(A) Frame 1/300 – before crystallization

(B) Begin of dendrite growth (frame 112/300 – crystallization in progress)

(C) Frame 300/300 – crystallization has stopped

FIGURE 11 Scattering image sequence 1 for crystallization experiment: dendrites grow. The data are in logarithmic scale



(A) Frame 1/300

(B) Frame 100/300

(C) Frame 300/300

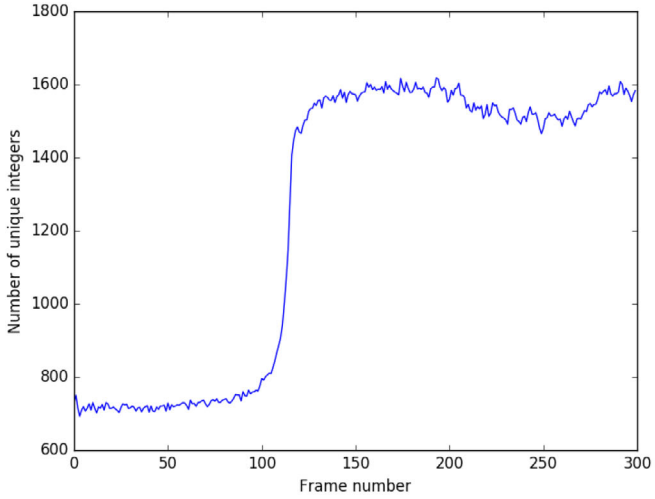
FIGURE 12 Scattering image sequence 2: a shadow moves through the background. The data are in logarithmic scale

background (snapshots are shown in Figure 12). However, each set can have hundreds of images and it is not possible for a human to figure this out. The goal is to automatically identify the subsequence of images during which the change occurs (dendrites grow, the shadow moves) which will enable the scientist to save time visually analyzing the images.

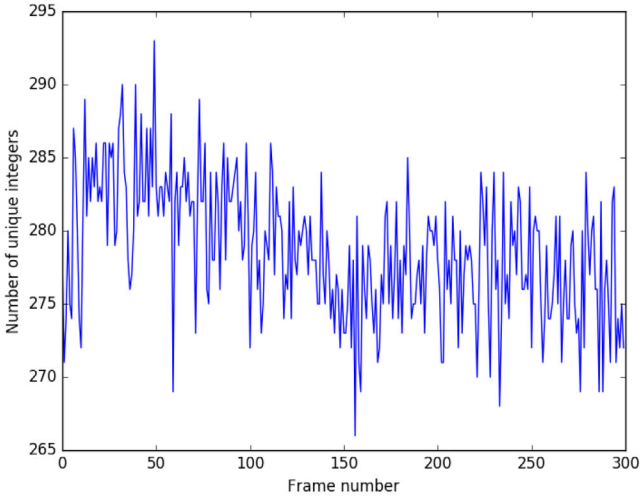
The scattering images are provided in EDF (European data format[37]) and are collected in HDF5 files that are about 3GB. Both cases contain 300 images. We import the extracted EDF files in Python with the fabio package.[38] We also directly work with the HDF5 files using the h5py package.[18] All images are represented by a matrix of integers of size (1679, 1475).

The goal of our analysis is to find the subsequence of images that contain the most important temporal changes, that is, in which frames do the dendrites grow (use case 1) and in which frames is the shadow moving through the background (use case 2)? We must identify statistics that compute the changes fast since light source users need real-time feedback to make adjustments for follow-up experiments. Since these change statistics will influence follow-up experiments and potentially be used to determine which images to store, we have to ensure that we do not fail to identify changes, and thus avoid false negatives (not detecting change if there is change). Wrongly identifying change where there is none (false positives) is, however, acceptable.

The first change measure we use is a simple *count of unique integers* in each frame. Figure 13 shows this measure for both test cases. There is a clear difference between the results for case 1 and case 2. For case 1 (left panel), we observe a large jump in the number of unique integers between frames 90 and 130. The frames in which the jump is observed also coincides with our visual inspection of the 300 frames and crystallization detection. On the contrary, counting the number of unique integers for case 2 gives us non-informative results (see Figure 13(B)). For case 2, the number of unique integers ranges between 265 and 295, whereas for the case 1, the numbers range between 700 and 1600. Also a visual inspection of all images for case 2 reveals that there is no event that causes an intensity increase as in case 1. Thus, we conclude that counting the number of unique integers is a useful and fast change measure for observations similar to case 1 in which obvious changes are expected to be observed (Figure 14).



(A) X-ray scattering data sequence 1 (dendrite growth).



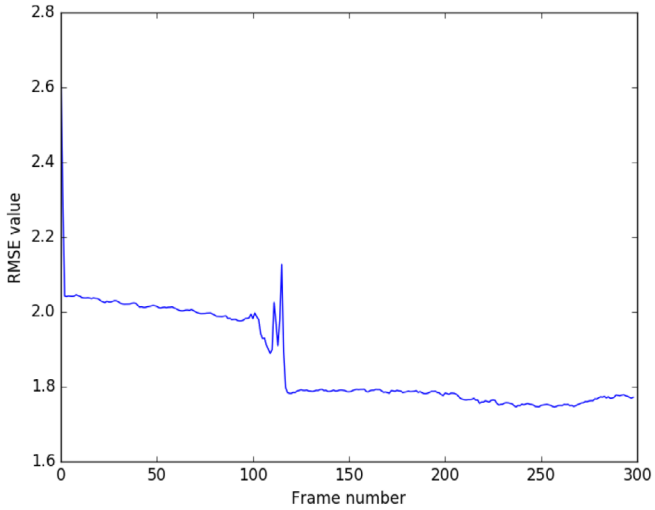
(B) X-ray scattering data sequence 2 (moving shadow).

FIGURE 13 Number of unique integers versus frame number. The obvious jump (left frame) indicates when significant changes (such as crystallization) are observed. For subtler changes (right frame) the number of unique integers as change measure is not informative

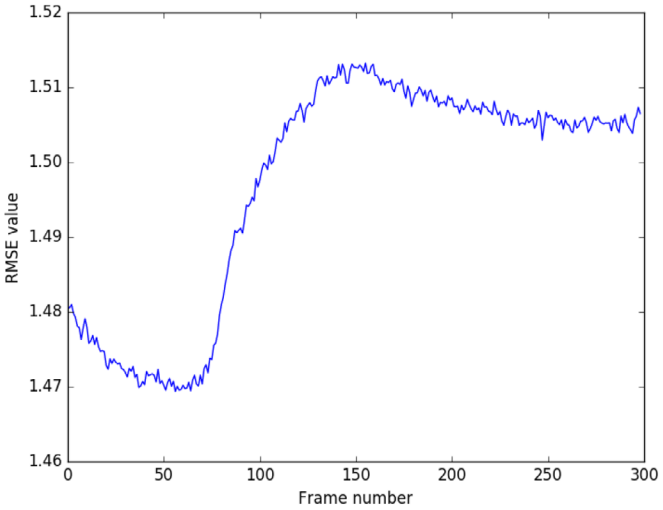
Our second measure is the RMSE between consecutive frames:

$$RMSE(A,B) = \sqrt{\frac{\sum_{i=1}^{N_r}\sum_{j=1}^{N_c}(A_{ij}-B_{ij})^2}{N_r \times N_c}}, \tag{1}$$

where A is a matrix representation of the Ith frame and B is a matrix representation of the $(I+1)$th frame. The subscript ij denotes the ijth element of each matrix. N_r denotes the number of rows and N_c is the number of columns of each frame (here: $N_r = 1679$, $N_c = 1475$). We compute the frame-to-frame RMSEs for both cases and we observe an obvious change in the RMSE values for both cases (approximately between frames 90 and 120 for case 1, and between 70 and 150 for case 2) which coincide with our visual inspection of the image sequences (Figure 14).

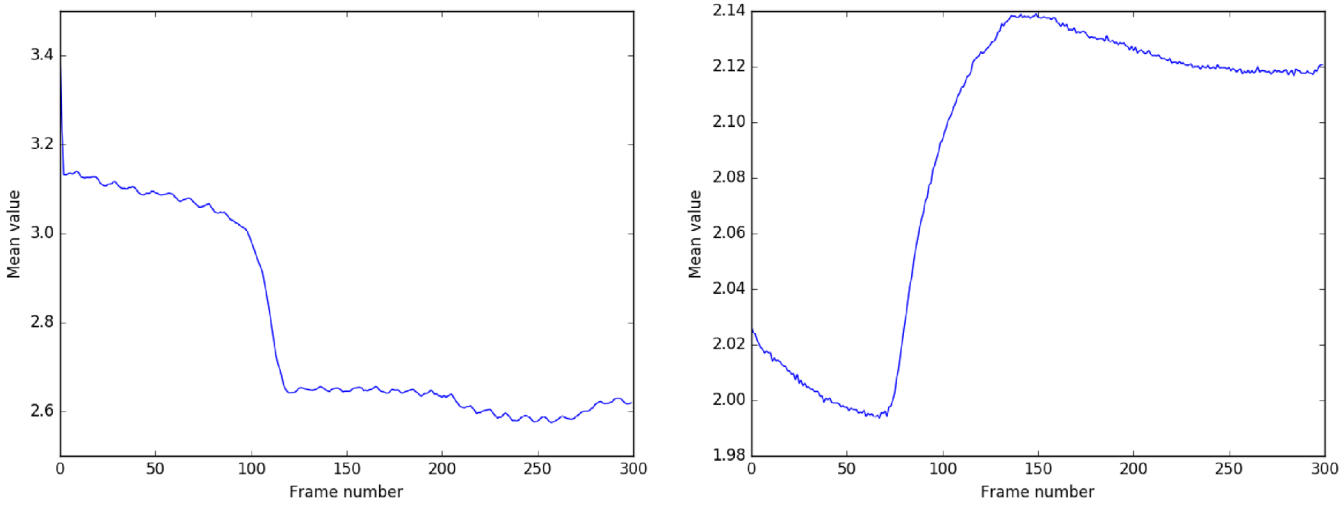


(A) Image sequence 1 (dendrite growth).



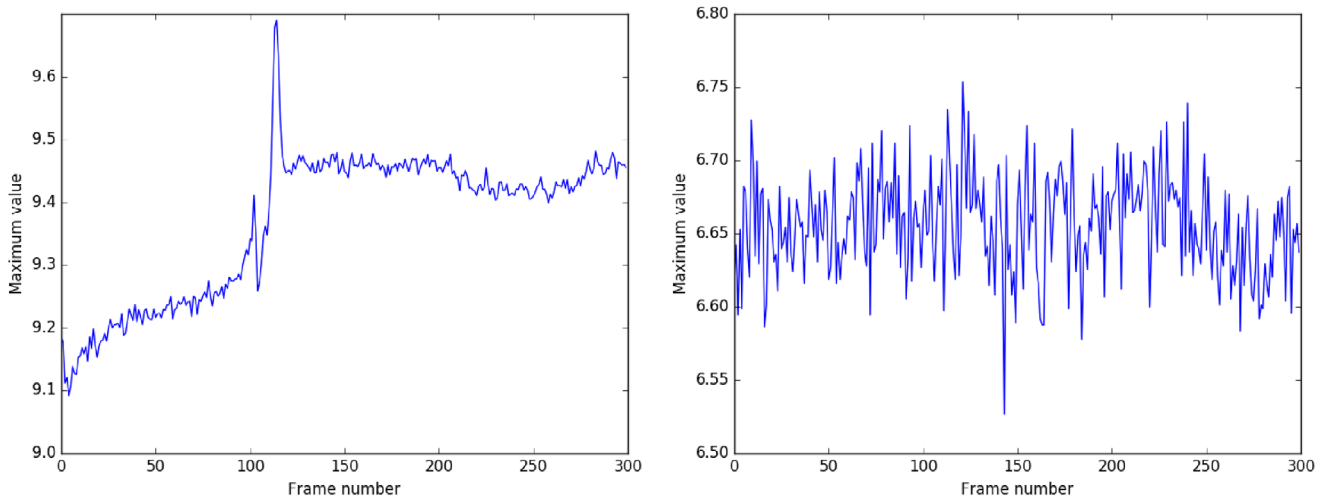
(B) Image sequence 2 (moving shadow).

FIGURE 14 Root mean squared error between consecutive frames. For both test cases, the changes indicated by the RMSE values coincide with the true changes



(A) Image sequence 1 (dendrite growth). (B) Image sequence 2 (moving shadow).

FIGURE 15 Mean value over each frame versus frame number. This metric indicates change for obvious and subtle changes



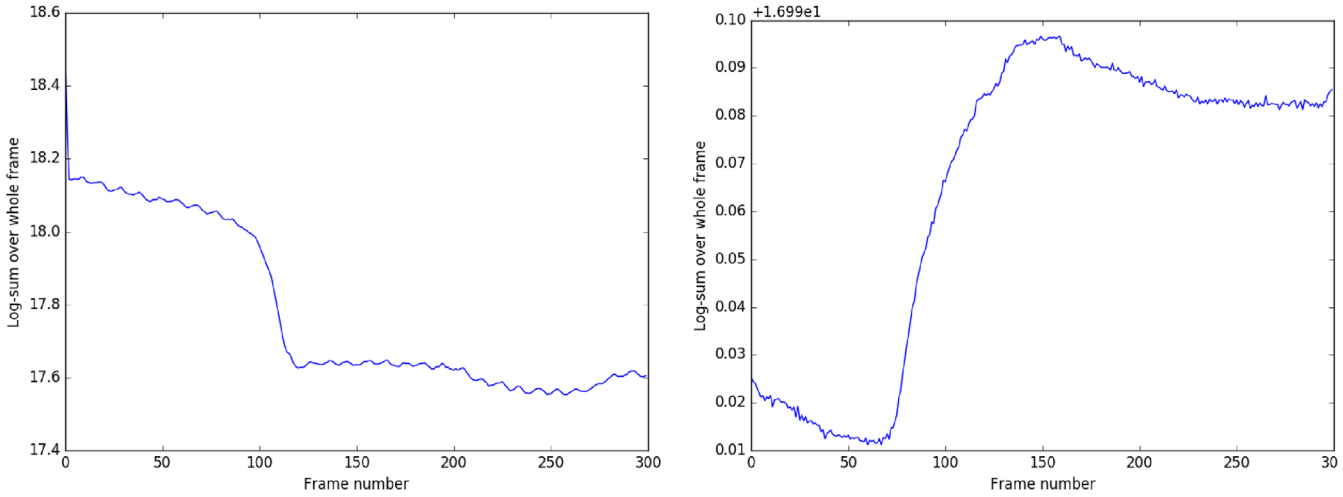
(A) Image sequence 1 (dendrite growth). (B) Image sequence 2 (moving shadow).

FIGURE 16 Maximum value over each frame versus frame number. This metric indicates change only for obvious changes

We also calculate other descriptive statistics such as the mean and maximum numbers for each frame, as well as the sum of the data over the whole frame. The results are shown in Figures 15 (mean), 16 (max), and 17 (frame sum). The figures corresponding to case 1 (left panels) clearly show when a significant change occurs, which is approximately between frames 90 and 130. The maximum values are significantly larger during crystallization (Figure 16(A)), and the sum of the data over each frame also confirms that the important frames are between 90 and 130 (Figure 17(A)). We conclude that for image sequences in which significant changes are expected, all of the change measures (number of unique integers, RMSE, mean, maximum, and frame sum) are informative for detecting the change.

For case 2, only the mean value and the frame sum are informative change indicators (between frames 75 and 150, Figures 15(B), 16(B), and 17(B)). Case 2 does not contain a process such as crystallization that significantly changes the intensity of the pixels. Thus, we cannot expect the maximum value over each frame to change as significantly as for case 1. Figure 16(B) confirms this expectation. From this analysis, we conclude that for image sequences for which we expect only subtle frame-to-frame changes, the most informative change measures are RMSE (Figure 14(B)), mean values (Figure 15(B)), and the frame sum (Figure 17(B)). The maximum value and the number of unique integers per frame are not informative.

Each of the five change measures, when applied to a series of images, can be summarized in a single graph that the scientist has to look at in order to identify at what point in time (which frames) the change occurs. Thus, instead of inspecting all 300 frames, it is sufficient to focus on ~40 (case 1) and ~75 (case 2), respectively, frames for which change actually happens. Since some of the change measures perform well only for one of the use cases while other measures are uninformative, we recommend computing all change measures for any given scattering dataset. This guarantees that some subset of the measures shows the information we are looking for.



(A) Image sequence 1 (dendrite growth). (B) Image sequence 2 (moving shadow).

FIGURE 17 Sum over all data of each frame versus frame number. This metric indicates change for obvious and subtle changes

4 | DISCUSSION

In the previous section, we analyzed various data change metrics for datasets encountered in diverse science applications related to earth sciences, cosmology, and materials science. We used different types of analyses, ranging from quantitative statistical measures to graphical representations to capture the data change and present it.

For the scientist interested in analyzing data change in their specific applications, we recommend using a hierarchical approach. We believe that the general structure of the hierarchy will remain the same across datasets, but the specific features of the hierarchical structure will be different for different datasets. Thus, we should strive for a combination of several complementary metrics in order to obtain a comprehensive understanding of the data change. Generally, we recommend computing as many metrics for as many expertly-chosen data subsets as the time frame for data change analysis allows. Based on our case studies, we provide a discussion of our guiding questions outlined in the beginning of this article, for developing a hierarchy for data change analyses that we recommend following when assessing change for a new dataset.

4.1 | Time limit for data change analysis

One consideration for choosing an appropriate data change metric is the time available to obtain the analysis results (Figure 18). For example, computing the statistics for data change (RMSE, mean, minimum, maximum change, etc.) is computationally inexpensive (once the data are locally stored) and can be done for many types of datasets, yet the existence of fill-values and data gaps needs to be given special consideration. Based on the use cases that we analyzed in the previous sections, we have to distinguish between real-time (online) analysis and offline analysis. In real-time analysis, we have to detect the data change fast. For example, the scientists at light sources or other experimental facilities need immediate feedback about their experiments to monitor their data and verify that their experiments are going correctly. Scientists are interested in looking only at the images that contain the event of interest. These images have to be identified as they are being captured. Lengthy procedures of writing the data to storage and then retrieving the information later for analysis is not practical. Thus, we only have a very low amount of time for conducting the data change analysis, which significantly limits the number and the type of change metrics that can be used. Fast-to-calculate metrics are the only practical option. In the ALS use case we found that computing statistics (RMSE, maximum, frame sum, etc.) was fast and representative of most data changes.

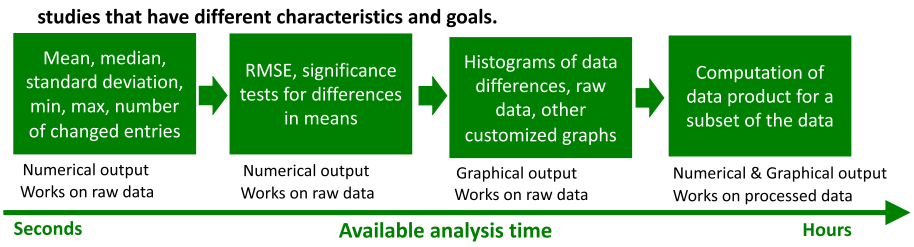


FIGURE 18 Hierarchical data analysis. Our figure shows the different metrics that can be used for data change analyses in relation to available analysis time

On the other hand, if we are facing a situation similar to the MODIS data, where a complete recomputation of the shortwave radiation product would require 1.5–2 months, investing a few hours or even a day on the data change analysis is acceptable. Since the outcomes of the change analysis may cause a high computational expense, many different metrics should be computed in order to gain a comprehensive understanding of the data changes and their impact on the data product. Compared to multiple months of data processing, the compute overhead of a thorough change analysis is negligible. In these cases, we recommend computing a subset of the data product with the updated data (in the MODIS case, we could afford the solar radiation computation for a few representative days) and use the data change analysis to evaluate the corresponding change in the data product. Similarly, in the SDSS data case, there was no time limit for the change analysis. The goal in SDSS is to publish correct data that do not contain biases or unreasonable systematic changes. Thus, the goal is to ensure data quality, and therefore the amount of time spent on the data change analysis is irrelevant. For both SDSS and MODIS, we found that presenting data change in the form of statistics and histograms allowed the most comprehensive analysis.

In summary, when choosing the metrics for analyzing the data change, there will be a tradeoff between the computational cost (amount of time) for the analysis and the thoroughness of the analysis. If only very little time is available, then choosing the most representative metrics that can be computed within very short time is highly important.

4.2 | Data change presentation

Several types of data change presentations should always be considered for the analysis. Our results show that a combination of numerical metrics and graphical representations is beneficial. Although images are relatively quick to look at, it still requires a certain level of user experience to understand data change, objectively analyze it, and derive detailed insights and unbiased conclusions. As we have seen for the two ALS datasets, the graphical representation of the number of unique integers per frame carries a clear message about data change only for one case whereas for the other case, the message was not that clear. One must be careful to not try and interpret too much information into a single figure as it may not necessarily carry all the information one is looking for. Thus, more than one change metric is often necessary.

For the MODIS use case, we saw that graphical representations of the data (old versus new) as well as several statistics about the data were necessary to identify changes. For the AOD data, for example, the 3D histograms showed that many of the data lie on the diagonal (indicator that values did not change). The statistics that showed the minimum, maximum, mean, and so on, of the old and the new data versions confirmed this.

For problems such as the SDSS data where scientists are concerned about the possibility of introducing systematic errors into the new data release by using a new processing software, numerical values that reflect which data columns changed allow a first high level view. Histograms that show the difference between the old and new data are insightful. Scientists generally have an expectation about the shape of the data change distribution and the presence of outliers. Inspecting 2D histograms that clearly show the modes of the distributions and their spread may help identify bugs or verify features in the processing software that was used to generate an updated dataset.

4.3 | Hierarchy of data change analysis

Different analysis time frames require different types of data change analysis. In an real-time online data change detection, we may not have the time to compute all change statistics. The ranges of the changes are also less important as the question is whether or not changes occurred. If there is change, then a visual analysis of the relevant frames will follow in order to verify that the experiments proceeded as planned, and, if needed, to prepare follow up experiments. The data change detection serves in this case as a tool to save time during image processing and further data collection.

In the offline analysis, when significantly more time and computer resources can be invested in the change analysis of a datasets, a good starting point for change detection is the computation of simple statistics. These statistics could include the mean, the median, the maximum, and the minimum of both the old and the new data. Histograms that show the distributions of old and new data may also be insightful. For time series data, a good metric to look at is the number of data entries and whether the entries were appended or inserted, which may indicate that old data records have been "rediscovered." For datasets that have fill values or NaN's to indicate gaps, counting the number of these entries in both datasets may be insightful in order to gain an understanding of the type of data change that occurred. If the only change is that NaN's were converted to numbers, this could be an indicator of a new processing software that does gap filling or the inclusion of data records that were previously not there, and thus further analyses should be conducted. Regarding fill values, care must be taken in case these values change between data versions.

The second level of data change analysis should include RMSE computations and visual presentations of absolute differences in data entries. In order to compute these metrics, one has to carefully match entries from the old and the new datasets (in case of data insertions as in the SDSS use case). Similarly, as we cannot compare NaN values to real values, a preprocessing step is involved that excludes entries from both data sets when at least one of them is labeled NaN. Similar to NaN entries, fill values must be excluded as they are real numbers and would lead to incorrect data change computations.

The next level of data change analysis could then include the computation of data product changes if applicable. For the MODIS case, this meant to compute the solar radiation product for representative days of the year which then allowed us to gain insight into how data changes translate into data product changes. This could further lead to sensitivity studies for datasets where many more data-product pairs can be computed within reasonable time, that is, one could investigate the relationship of the range of the (input) data change and the corresponding range of the data product change. This might allow predictions for the influence of future data changes.

4.4 | Objectives of data change analysis

The goals or/and objectives of what is to be derived from the data change analysis should also play a role in the decision about the metrics to compute and how far the hierarchy tree of change metrics should be traversed. Again, it is important to differentiate between online and offline analysis. In the online analysis, the goal could be as simple as to detect if at all and when a change happened, and a more detailed analysis may not be necessary because a scientist will visually inspect the change-relevant images to verify the correctness of experiments. On the other hand, in the offline analysis, the details of the data changes are important and the goal may be to study their influence on the final data product.

Similarly, there should be a difference between data user and data supplier. When data suppliers publish a new version of their data, the quality and accuracy of the data are of highest importance as hundreds or thousands of data users will be impacted. Thus, the data supplier should take into account all applicable data change metrics, from numerical indicators to visual representations of the old versus the new collection. From the data user point of view, it is of interest how data changes will impact their data products. If data products are used for making policy decisions that impact a large number of people and companies (e.g., water management decisions), the data change analysis should be as detailed as possible in order to capture if the new data will yield different outcomes.

4.5 | Challenges of developing general data change analysis methods and software

Throughout our study, we have encountered multiple difficulties that make it challenging to develop a general data change analysis tool. *The use of different file formats* in different science areas is one challenge we encountered in our analysis. Scientific data are recorded in different (often very science specific) formats (HDF for MODIS, EDF for the ALS, FITS for SDSS) and different software libraries are needed to read these datasets. It is difficult to create data change analysis software that is robust to all kinds of data formats. It is necessary to explore a standard data model that can be used for the analyses and plugins that allow users to convert from specific formats to the data models. This will also allow the use of the methods developed in this article and otherwise with possibly diverse data sources.

For some scientific applications, only *subsets of the data* are needed in the computation of a data product of interest. For example, the NASA-provided MODIS data files contain many more datasets than needed in the computation of the solar radiation product. Thus, instead of computing the differences for all datasets in MOD04_L2, we only have to compare the values for `Corrected_Optical_Depth_Land`and not, for example, for `Surface_Reflectance_Land` or `Mass_Concentration_Land` as the latter two datasets do not enter the solar radiation computation. Also, different fill values are used for different data products for times when no measurements were available. These fill values are often given as very large or very small numerical values and must be identified by the user because an automatic analysis tool will not be able to distinguish a true measurement from a fill value as both are just scalars. Again, a preprocessing step of the data by the user is necessary.

Another challenge arises from *matching data in the old and new datasets.* As we have seen for the SDSS data, new observations are not only appended but also inserted between previous observations and a unique combination of keys (columns) is necessary to match and compare the data. Similarly, time series data must be matched such that observations are compared correctly. In time series, data are not only added for the days between pulls, but sometimes also inserted or deleted.

Different data types (continuous, binary, integer, NaN, fill values) can be encountered and must be treated correctly. Specific data types may allow for more specific types of change analysis. In the ALS datasets, we only deal with integer values, which also allows us to define a change metric that is based on the number of unique integers per frame. For MODIS data, we generally deal with continuous and fill values. Fill values must be defined by the user so as to not corrupt the change metric computation.

Developing a software that is agnostic to the data type and file format will be challenging without any burden on the user's side. Ideally, we would like to have a software that we can give two datasets and that will return all the differences that exist. But the above mentioned hurdles will have to probably be addressed by the users themselves. At this point it becomes important to identify which "part" of the data enters the computation of their data products. For example, if only summary statistics (such as mean values) are used in the computation of the data product, computations that investigate point by point differences are not needed. Thus, it seems more promising to define an input data format for the software that the user has to adhere to.

Generalized metrics for data change analysis are hardly possible and an *interactive software design* might be necessary. In a hierarchical data change analysis the computation of the different statistics at the lowest level could be the only metrics that are general change indicators. Beyond these metrics, it seems most useful to have the user in the loop in order to identify which metrics will make sense for their specific objective at the next level.

5 | CONCLUSIONS

Scientific datasets often undergo periodic updates as new observational data become available or new processing methods and software are developed. For the scientist, it is important to understand the impact of data changes on their workflows and science results. In many cases, the cost of recomputation of data products is high and scientists would prefer to not reprocess the data if the changes are insignificant or irrelevant.

In this article, we presented a novel hierarchical approach to the quantitative evaluation of the data changes for real-time (online) and offline analyses. We illustrated the application of this approach using three use cases with different categories of data—satellite data, cosmological data, and imaging data. In particular, we explored the application of different types of data change metrics to assess the changes in the MODIS data (used in Earth sciences), SDSS data (used in cosmology), and x-ray scattering data from the ALS. We computed simple statistics of the data changes (mean, median, etc.), used graphical displays of the data versions and their differences (histograms), and derived application-specific change metrics to exploit special characteristics of datasets. We found that the data change metrics should be aligned with the available time for the change analysis: the online data change analysis has to be fast and it may suffice to indicate if change happened; the offline analysis should be comprehensive and multiple metrics should be used to accurately analyze what changed and how it changed. If the goal is to narrow down the scientist's focus on a subset of the data for further visual analysis (as in the ALS data case), then using few fast-to-compute metrics is sufficient. On the other hand, if the outcome of the data change analysis would imply months of additional computations, or if a problem with the data quality would impact a wide network of users, then all possible change metrics should be taken into account to ensure a comprehensive understanding of the changes and that the conclusions drawn from the change analysis are correct.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy, Office of Science and Office of Advanced Scientific Computing Research (ASCR) under Contract No. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

DATA AVAILABILITY STATEMENT

A subset of the data that support the findings of this study are available in the supplementary material of this article.

ORCID

Juliane Müller https://orcid.org/0000-0001-8627-1992

REFERENCES

1. Paine D, Ramakrishnan L. *Surfacing Data Change in Scientific Work*. Berkeley, CA: Lawrence Berkeley National Laboratory; 2018.
2. Zhang Y, Song C, Band LE, Sun G, Li J. Reanalysis of global terrestrial vegetation trends from MODIS products: browning or greening? *Remote Sens Environ*. 2017;191:145-155.
3. Jiang C, Ryu Y. Multi-scale evaluation of global gross primary productivity andevapotranspiration products derived from Breathing Earth System Simulator (BESS). *Remote Sens Environ*. 2016;186:528-547.
4. Ghoshal D, Ramakrishnan L, Agarwal D. DAC-MAN: data change management for scientific datasets on HPC systems; 2018.
5. Paine D, Ramakrishnan L. Surfacing data change in scientific work. Paper presented at: Proceedings of the International Conference on Information; 2019:15-26; Springer, New York, NY.
6. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. *Sov Phys Dokl*. 1966;10:707-710.
7. Hamming RW. Error detecting and error correcting codes. *Bell Syst Tech J*. 1950;29(2):147-160.
8. Navarro G. A guided tour to approximate string matching. *ACM Comput Surv*. 2001;33(1):31-88.
9. Dobrushin RL. Definition of a system of random variables by conditional distributions ((In Russian)). *Teor Verojatnost i Primenen*. 1970;15:469-497.
10. Drakopoulos V, Nikolaou NP. Efficient computation of the Hutchinson metric between digitized images. *IEEE Trans Image Process*. 2004;13(12):1581-1588.
11. Makridakis S, Hibon M. The M-3 competition: results, conclusions, and implications. *Int J Forecast*. 2000;16(4):451-476.
12. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecast*. 2006;22(4):679-688.
13. Mount DM. *Bioinformatics: Sequence and Genome Analysis*. 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2004.
14. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*. 2005;30:79-82.
15. Durbin J. *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia: Society for Industrial and Applied Mathematics; 1973.
16. Marsaglia G, Tsang WW, Wang J. Evaluating Kolmogorov's distribution. *J Stat Softw*. 2003;8(18):1-4. https://doi.org/10.18637/jss.v008.i18.
17. NASA MODIS website; https://modis.gsfc.nasa.gov/.
18. Collette A. *Python and HDF5: Unlocking Scientific Data*. Sebastopol: O'Reilly Media Inc; 2013.
19. Ryu Y, Baldocchi DD, Kobayashi H, et al. Integration of MODIS land and atmosphere products with a coupled-process model to estimate gross primary productivity and evapotranspiration from 1 km to global scales. *Global Biogeochem Cycles*. 2011;25:GB4017. https://doi.org/10.1029/2011GB004053.
20. Hendrix V, Ramakrishnan L, Ryu Y, van Ingen C, Jackson KR, Agarwal D. CAMP: community access MODIS pipeline. *Futur Gener Comput Syst*. 2014;36:418-429.
21. Zhang X, Friedl MA, Schaaf CB, et al. Monitoring vegetation phenology using MODIS. *Remote Sens Environ*. 2003;84(3):471-475.

22. Gao F, Anderson MC, Zhang X, et al. Toward mapping crop progress at field scales through fusion of landsat and MODIS imagery. *Remote Sens Environ*. 2017;188:9-25.
23. Mertes CM, Schneider A, Sulla-Menashec D, Tatem AJ, Tan B. Detecting change in urban areas at continental scales with MODIS data. *Remote Sens Environ*. 2015;158:331-347.
24. Sloan Digital Sky Survey website http://www.sdss.org/.
25. Ponz J, Thompson R, Munoz J. The FITS image extension. *Astronomy Astrophys Suppl Ser*. 1994;105:53-55.
26. Fagioli M, Riebartsch J, Nicola A, et al. Forward modeling of spectroscopic galaxy surveys: application to SDSS. *J Cosmol Astropart Phys*. 2018;2018(11):015-015.
27. Zhang J, An R, Luo W, Li Z, Liao S, Wang B. The first constraint from SDSS galaxy–galaxy weak lensing measurements on interacting dark energy models. *Astrophys J Lett* 2019;875:L11.
28. Bañados E, Venemans BP, Mazzucchelli C, et al. An 800-million-solar-mass black hole in a significantly neutral Universe at a redshift of 7.5. *Nature*. 2017;553:473.
29. Li X.-D, Park C, Sabiu CG, et al. Cosmological constraints from the redshift dependence of the Alcock–Paczynski effect: application to the SDSS-III boss DR12 galaxies. *Astrophys J*. 2016;832:103.
30. Advanced light source website https://als.lbl.gov/.
31. Kilcoyne ALD, Tyliszczak T, Steele WF, et al. Interferometer-controlled scanning transmission X-ray microscopes at the advanced light source. *J Synchrotron Radiat*. 2003;10:125-136.
32. Le Gros MA, McDermott G, Cinquin BP, et al. Biological soft X-ray tomography on beamline 2.1 at the advanced light source. *J Synchrotron Radiat*. 2014;21:1370-1377.
33. Gupta S, Celestre R, Petzold CJ, Chance MR, Ralston C. Development of a microsecond X-ray protein footprinting facility at the advanced light source. *J Synchrotron Radiat*. 2014;21:690-699.
34. Jiang C, Ryu Y, Fang H, Myneni R, Claverie M, Zhu Z. Inconsistencies of interannual variability and trends in long-term satellite leaf area index products. *Glob Chang Biol*. 2017;23(10):4133-4146.
35. Alam S, Albareti FD, Prieto CA, et al. The eleventh and twelfth data releases of the Sloan digital sky survey: final data from SDSS-III. *Astrophys J Suppl Ser*. 2015;219(12):27.
36. Robitaille TP, Tollerud EJ, Greenfield P, et al. Astropy: a community Python package for astronomy. *Astron Astrophys*. 2013;558:A33.
37. Kemp B, Värri A, Rosa AC, Nielsen KD, Gade J. A simple format for exchange of digitized polygraphic recordings. *Electroencephalogr Clin Neurophysiol*. 1992;82(5):391-393.
38. Knudsen EB, Sørensen HO, Wright JP, Goret G, Kieffer J. Fabio: easy access to two-dimensional X-ray detector images in python. *J Appl Crystallogr*. 2013;46(2):537-539.

How to cite this article: Müller J, Faybishenko B, Agarwal D, et al. Assessing data change in scientific datasets. *Concurrency Computat Pract Exper*. 2021;e6245. https://doi.org/10.1002/cpe.6245

APPENDIX A. DATA SOURCES

Below are the references for the data and processing software we used in this study.

SDSS

The two SDSS data files used in our study are available at:
https://portal.nersc.gov/project/cosmo/data/sdss/dr11/boss/spectro/redux/v5_6_5/spAll-v5_6_5.fits, 7.8 GB;
https://portal.nersc.gov/project/cosmo/data/sdss/dr12/boss/spectro/redux/v5_7_0/spAll-v5_7_0.fits, 9.2 GB; The documentation is available at https://www.sdss.org/dr11/ and https://www.sdss.org/dr12/.

MODIS

The MODIS collection history is available at
https://modis-images.gsfc.nasa.gov/products_C006update.html,
and the data are downloadable from
https://modis.gsfc.nasa.gov.
The BESS processing software for solar radiation computation is available at
http://environment.snu.ac.kr/bess_flux/.