Earth Syst. Sci. Data, 10, 2275–2278, 2018
https://doi.org/10.5194/essd-10-2275-2018

Open Access Earth System Science Data

Editorial: Data publication – *ESSD* goals, practices and recommendations

David Carlson[1] **and Tomohiro Oda**[2,3]

[1]Co-Chief and Founding Editor (outgoing) of Earth System Science Data, Director (retired) of World Climate Research Programme, Bozeman, MT, USA
[2]Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA
[3]Goddard Earth Sciences Technology and Research, Universities Space Research Association, Columbia, MD, USA

Correspondence: David Carlson (ipy.djc@gmail.com)

Published: 13 December 2018

Abstract. *Earth System Science Data* (*ESSD*) provides a wide range of openly accessible, high-quality, well-documented and highly useful data products while ensuring recognition of and credit to data providers. As authors, reviewers, and editors of many *ESSD* publications, we encounter uncertainty about mechanisms and requirements for open access, about what constitutes a published data product, and about how one goes about submitting, evaluating or using *ESSD* products. With this short note, published during an important editorial transition, we use our combined experience to define guidelines, requirements and benefits of the *ESSD* processes.

1 Introduction

Earth System Science Data (*ESSD*), initiated during the International Polar Year 2007–2008 and supported from then until now with the skill and generosity of Copernicus Publications, exists to

a. induce and facilitate exchange of original data or data collections useful for advancing earth system research,

b. ensure free, open and persistent access to those data,

c. certify – through peer review and public discussion – the scientific quality of those data and the accuracy and utility of the data descriptions, and

d. provide full useful acknowledgement and credit to data providers.

Publishing data in *ESSD* entails more than just sharing data through mechanisms such as specifying the URL for a data repository in a journal article. Researchers can satisfy institutional or national data access requirements by using effective repository and access services offered by data archive centres; we do not dispute any data centre's claim to have "published" those data. However, to thoroughly understand and exploit data for subsequent research, users increasingly turn to *ESSD* or other data publication journals for extensive descriptions, validations and documentation. *ESSD* thus catalyses a positive feedback loop; users require more information, data providers develop better descriptions and more-extensive uncertainty analyses, and more data (archived at more data centres) become more useful to more users. Because much-used *ESSD* products very often reflect skilful compilations from many sources, users gain valuable information about those sources. With so much data potentially accessible, and users often at a loss about which version, which product or which source on which to base their work, it becomes convenient, necessary and even critical for *ESSD* to establish, sustain and communicate the highest standards for data quality and availability through its publication processes. With this short commentary, in the format of an annotated checklist, published on the occasion of retirement of the founding chief editors of *ESSD*, we – from our combined author, reviewer and editor viewpoint – describe basic requirements that we consider most important for publishing high-quality research data in *ESSD*. We offer our list as a guideline for incoming editors, for reviewers and for authors and data

providers. Although we anticipate that this list and summary may prove useful for a wider research community, we focus here specifically on *ESSD*. To extend and enhance its relevance, *ESSD* must define its role(s) and functions within and in service to our evolving earth system research environment.

2 Primary goal and purpose of *ESSD*

ESSD aspires to meet or exceed this expectation; careful work by data providers and a "stamp of approval" from *ESSD* will allow future researchers to use *ESSD*-published data with confidence. The journal expects that a future user, 5 or more years after date of publication, will find exactly the data, the tools and the recipe (description) that allow her or him to completely and reliably reproduce any figure from the original data description or accompanying research paper. Future users must know that they could, if necessary, start exactly where *ESSD* authors started and – presumably as a starting point for their own fresh analysis – reproduce exactly the same outcomes. Because data repositories can, unfortunately, cease operation, the combination of an *ESSD* data description with a valid permanent identifier should allow future readers of the *ESSD* publication to access the original data product, reproduce any data manipulations, and assess utility of that data – regardless of its age or current repository – for research applications.

3 Recommendations

To enable this overall goal of reliable durable data access, *ESSD* recommends that providers, reviewers and users adhere to the following recommendations.

3.1 Emphatic open access

For *ESSD*, easy free open access to data applies to data providers as well as users. Data providers must have easy access to no-cost mechanisms and services that will curate their data. Curation includes reliable long-term storage and backup, minting and maintenance of permanent identifiers, and appropriate metadata services that facilitate search, identification and download. Users, following identifier links embedded in an *ESSD* paper, should enjoy fast free reliable "two-click" access: one click to a relevant landing page and a second click to download. An ideal repository will include topical and geographic browsing. Users should not encounter registration steps, password requests, access agreements, or other log-in barriers or tracking mechanisms. From the start of open discussion, *ESSD* data products should exist in full public access without proprietary protection periods or other restrictions. Data publication as practised by *ESSD* depends on free bilateral unrestricted access. Most data repositories used by *ESSD* providers promote and support exactly this level of open access. New-to-*ESSD* data repositories can usually provide identical levels of access service. When authors lack information about or access to appropriate data centres, *ESSD* provides guidance and recommendations.

ESSD likewise insists on easily accessible non-proprietary databases, data products, data processing codes and other software tools necessary to process and use published data. Positive examples include Comma-separated value (.csv) files (which, if skilfully prepared, can contain abundant metadata), netCDF files (enabled by well-documented manuals and freely available netCDF libraries), MySQL databases, QGIS-compatible shapefiles, open-source script codes (R, Python), etc. Proprietary software products such as ArcGIS, MATLAB and Microsoft Access fail to support the open access and exchange necessary for *ESSD* data publication; products in these formats require conversion to non-proprietary formats for data sharing. Because researchers can generally use Excel and because many free translators exist, *ESSD* accepts Excel files as a special case.

3.2 Mandatory permanent identifiers

ESSD emerged synchronously with the application of digital object identifiers (DOI) to research data. The use of DOI for data identification and tracking and for version control remains critical to data publication processes; all *ESSD* data sets and data products must carry a DOI from the time of manuscript submission. Application of the DOI system to these products, whether flat files, databases or algorithms, serves to protect and inform both users and providers. Changes implemented as a consequence of the *ESSD* review process should in all cases result in a new DOI for the revised data product. The final published product will carry two DOI: one for the final data product as reviewed and perhaps revised and a second for the published description.

3.3 Accurate useful data descriptions including source attribution

An *ESSD* data description provides a unique complete recipe covering original data sources, data collection methods where applicable, tools and overall preparation of the data product. To help users – particularly users interested in, but perhaps unfamiliar with, the product – *ESSD* authors must provide accurate documentation of sources, algorithms, codes, models, etc., sufficiently to allow new users to develop subsequent or alternate analyses or conclusions. Ideally, detailed *ESSD* data descriptions prevent or at least minimise subsequent data misinterpretations or misuse. A good *ESSD* paper will include attribution tables that summarise data used, data sources (with URL) and journal citations so that readers and users can easily follow the same links to the same sources. Formats for all data links and attributions should follow current best practices (find examples and links to formal data citation principles under *ESSD*'s data policy, https://www.earth-system-science-data.net/about/data_policy.html) and include a full accurate cita-

tion in the paper's reference list. A carefully prepared rigorously reviewed description represents a strong value-added feature of data publication as practised by *ESSD*.

3.4 Inclusive lists of codes and tools

To meet the goal of providing complete data preparation documentation, *ESSD* data products and data descriptions should include all codes, libraries, statistical or interpolation routines, model versions, etc. For example, when authors develop or use processing schemes in R, they must provide the specific names and URLs of those R codes. When they have validated their product through use of or comparison with models, they must provide exact details of model configurations, reliable links to model versions, etc., sufficiently to allow readers and users to replicate the analysis. Often, *ESSD* authors provide a flow chart of sources, processing steps and outcomes, accompanied by a table listing sources with necessary details. Data providers, who typically carry this information informally, generally benefit from the effort needed to formally record and document these procedures.

3.5 Extensive validations

Authors will need to demonstrate, first to reviewers and later to a wide range of users, the validity and applicability of their datasets and data products. Exact mechanisms and options for validation will vary substantially among and across data products. Because *ESSD* serves to ensure the suitability of published data for future research, each *ESSD* paper should demonstrate skill and utility of the submitted data product by some form of comparison to prior products, alternate data sources, similar products at different time or space resolution, model outcomes, initial short records of recent sensors, etc. For some data products, full validation with independent source materials may prove scientifically or technologically infeasible; comparisons to prior or alternate products may not offer quantitative validation. Community-wide compilations such as global budgets might only allow validation of specific components or comparisons to earlier versions. In all cases, however, authors must have made and reported best efforts at intercomparison and validation.

3.6 Explicit uncertainty accounting and analysis

Each *ESSD* data set, database, data product or data processing algorithm contains and perhaps induces uncertainties. *ESSD* products will also carry uncertainties inherited from source data. Authors must explicitly and extensively describe and document those uncertainties. Exact expressions of and standards for uncertainty will vary depending on types and sources of data, but as a service to and courtesy to subsequent users, every *ESSD* data product must include uncertainty documentation. Authors may need to rely on and cite their own expert judgement, but such conclusions must appear explicitly within an overall uncertainty assessment in terms of percent, standard deviation, or other accepted metrics. Future users, including modellers, require careful, explicit and quantitative uncertainty analysis that will allow them to choose or avoid subsequent use based on documented uncertainties of the *ESSD* product.

3.7 User guidance in a data availability section

Authors must describe access to their data product(s) in an explicit data availability section (another value-added feature of *ESSD*). This section must list current primary and alternate data repository links, explain any versions, include links to open-access source files, etc. Where a user will encounter multiple files, authors must explain the contents and expected uses of each file; the availability section should not point to an FTP site full of raw text (.txt) files. When, for convenience, authors provide smaller-size (e.g. monthly) files as surrogates for larger higher-resolution (e.g. daily) files available offline, the data availability section should provide explicit descriptions and access guidance for both the small and large files. "Contact the author" does not represent useful or appropriate guidance for data availability. As mentioned above, all links to third-party data sources should appear in the reference list in citable accessible formats; the authors may wish to add notations, guidance, version information, etc. in this section. Data availability sections should also describe plans and schedules for future updates, when applicable. All *ESSD* papers should have included their specific data link as the final sentence of the abstract and should repeat those links accompanied by all necessary explanation and assistance in the data availability section.

3.8 Interest and utility

Ideally, and to justify efforts of reviewers, editors and publication staff, an *ESSD* data product will prove interesting and useful to a wide range of users. Authors should know that, to ensure that *ESSD* products enable substantial advances in future research, editors must apply dual criteria in all cases; does the data as submitted demonstrate sufficient quality and will the data product interest a sufficient number of users? Clearly, a small data set collected over a short time at a single location generally does not qualify (e.g. an emission factor measured for a short time at one location might represent a significant data-gathering effort but with limited impact), while a community compilation of a global product covering 6 or 7 decades should qualify. Those end points include a wide range of plausible intermediary products. Over its short time of existence *ESSD* has tried to adopt an inclusive and expansive view of potential data impact. By close adherence to the guidelines above, data providers will help *ESSD* editors assess interest and utility.

4 Discussion and challenges

Very rarely will an *ESSD* paper meet all the above criteria with highest-level compliance. As reviewers and editors will know, proper effective evaluation should encourage authors to modify their initial submission in directions and with amendments that allow the data product as eventually published to more closely meet the full range of recommendations.

ESSD has proven – not least by the emergence of notable competitors in the data publishing realm – the utility of data publication as a service to data providers and data users. Challenges remain. Technical details and standards described here, including data formats and permanent identifiers, will evolve; *ESSD* must anticipate and accommodate those changes. Models, including visualisation environments, have grown as a pervasive part of earth science research. For many properties of the earth system, a high-resolution model forced with reanalyses, carefully validated against a long time series of sparsely distributed observations, represents a necessary approach. Perhaps not in *ESSD* but somewhere, by some discussion and evaluation processes similar to those implemented by *ESSD*, those model and observational compilations deserve documentation, evaluation, distribution and credit to authors just as much as "pure" observational datasets presented in *ESSD*. For many data compilations, databases offer a positive combination of accessibility and promulgation of metadata standards leading to useful search, browse, and upload and download possibilities. A description of those databases seems to represent a valid and useful function of *ESSD*, but authors often seem reluctant or confused about meeting the DOI-labelled data snapshot requirement of *ESSD*. How does one capture, document and preserve for posterity an effective record of a rapidly evolving database? Given the increasing size and scope of earth system data, particularly in our satellite era, researchers will and must develop and apply machine learning tools and techniques. If we agree that those tools and the data products that emerge should achieve the same level of open-access documentation and availability as other *ESSD* data products, does *ESSD* – now or in the future – represent the obvious and appropriate "home" for such data products? We believe *ESSD* will play a very positive role in the evolution and improvement of data exchange, publication and archive services.