European Commission

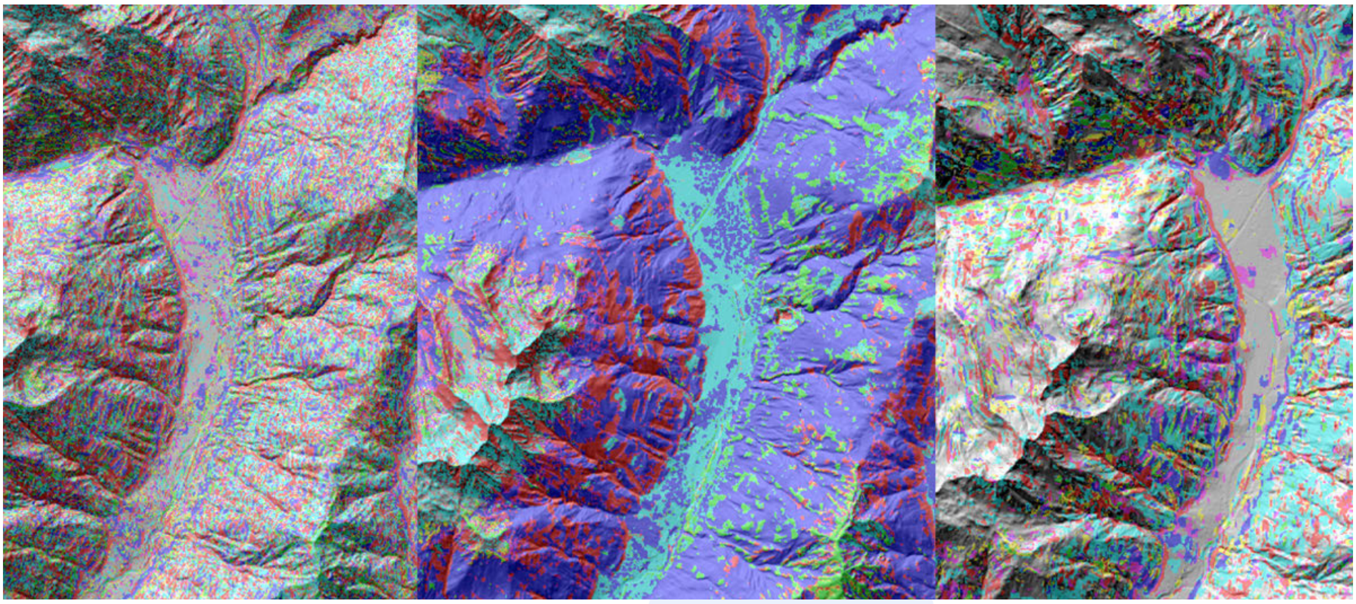
Joint Research Centre

The Digital Elevation Model Intercomparison Exercise - DEMIX Final Report

In-depth evaluation of six free&open global high-resolution DEMs using a novel, statistically sound framework highlighting the superiority of the Copernicus DEM

Bielski, C., Guth, P., Grohmann, C. H., López-Vázquez, C., Trevisani, S., Gesch, D., Riazanoff, S., Corseaux, A., Reuter, H.I., Strobl, P.

2026



Contact information
Name: Peter STROBL
Email: Peter.Strobl@ec.europa.eu

Joint Research Centre
https://joint-research-centre.ec.europa.eu

JRC134935
EUR 40689

PDF ISBN 978-92-68-38944-7 ISSN 1831-9424 doi:10.2760/3121340 KJ-01-26-158-EN-N

Luxembourg: Publications Office of the European Union, 2026

How to cite this report: Bielski, C., Guth, P., Grohmann, C., López-Vázquez, C., Trevisani, S. et al., *The Digital Elevation Model Intercomparison Exercise - DEMIX Final Report - In-depth evaluation of six free&open global high-resolution DEMs using a novel, statistically sound framework highlighting the superiority of the Copernicus DEM,* Publications Office of the European Union, Luxembourg, 2026, https://data.europa.eu/doi/10.2760/3121340, JRC134935.

Contents

Abstract

This report presents results of an intercomparison of Digital Elevation Models (DEM) carried out under the auspices of the Committee on Earth Observation Satellites (CEOS). A DEM is a fundamental geospatial input layer for a wide range of Earth Sciences. A variety of different products have become available over the past decade. However, selecting the best possible solution for a specific application is no longer a simple task and often users lack knowledge and resources to make informed choices. The Digital Elevation Model Intercomparison eXercise (DEMIX) set out to address this deficit. Through community wide calls for participation the initiative assembled a group of experts contributing on a voluntary basis over a period of three years. They clarified and harmonised the necessary terminology, developed an innovative criteria-based approach and combined it with a statistically sound ranking method, published practical guidelines and open-source tools which will remain accessible beyond the active phase of the project. This report presents an overview about the main results, providing additional findings and insights which go beyond the peer-reviewed publications resulting from the project. It further discusses the results from the perspective of the Copernicus programme and outlines why and how the comparison method developed could and should be applied to other geospatial products and tools.

Foreword

Scientist's need for their analysis to select a pertinent dataset. For this they look at its quality and suitability, something often called 'fitness for purpose'. Until recently, however, data were rather scarce and often what was available had to do the job. This has changed nowadays especially in the geospatial and earth observing sciences because of the technological advances in sensor development, public funding of space related missions over the last few decades, and a more free and open policy towards sharing information.

This report documents the journey of a group of scientists that found themselves charged with comparing a variety of different datasets all describing the same fundamental and truly 'environmental' parameter, the shape of the Earth's surface captured by what is called a Digital Elevation Model (DEM) in order to provide users with advice on which DEM product might be most recommendable for their purpose.

Under the leadership of the European Commission's Joint Research Centre (JRC) and mandated by the Committee on Earth Observation Satellites (CEOS) and its Working Group on Calibration and Validation (WGCV), the Terrain Mapping Sub-Group (TMSG) set up the DEM Intercomparison eXercise (DEMIX) to achieve this goal. From the start TMSG joined forces with the International Society for Geomorphometry in order to shift the focus away from the producer's interests and to follow an independent, science based, and user-oriented approach.

Several aspects of this exercise are particularly remarkable, if not unique, and set a precedence. Advantages are the application of a scientifically sound method, transparent and user-oriented guidance provision to professionals but also to non-experts. A central role in this approach plays a novel methodological framework implemented through state-of-the-art tools to intercompare a set of global DEM candidates based on pre-defined criteria, representative testing, and a statistically sound ranking approach. The latter leans heavily on methods established in statistics under the term 'wine contest' which is why the framework was coined the 'DEMIX wine contest'. The main rationale for this framework is to generate tailored recommendations regarding available DEM products that are not limited to one domain, geographic area, or landscape type but are flexible and customizable in relation to the specific needs and requirements of users and their applications.

DEMIX focused on comparing the majority of available free and open global DEMs with a spatial resolution of approximately 30m. Six such products were identified and subjected to comprehensive tests and comparisons against authoritative reference data of high and reliable quality, representing as much as possible the whole spectrum of landforms on Earth. While regional and morphology related differences exist, the results evidence clearly and reliably the overall superiority of the Copernicus DEM with respect to the other products.

Besides an in-depth discussion of the methodological framework of the 'DEMIX wine contest' and a detailed presentation of the performed tests, their results and analysis, this report contains a plethora of materials collected over almost three years during which DEMIX was performed. All of it is considered an important contribution to better and more harmonised characterisation and understanding of Digital Elevation Models assisting experts and non-expert users in making the right choices.

Acknowledgements

The authors would like to thank their families, their respective organisations, the Committee on Earth Observation Satellites, the International Society for Geomorphometry, and all DEMIX participants for their support in performing this voluntary exercise over several years.

Authors

Bielski, C., Guth, P. L., Grohmann, C. H., López-Vázquez, C., Trevisani, S., Gesch, D., Riazanoff, S. Corseaux, A.,Reuter, H. I., Strobl, P.

1. Introduction

Section Contributors: Bielski, C., Guth, P. L., Grohmann, C. H., López-Vázquez, C., Trevisani, S., Gesch, D., Riazanoff, S. Corseaux, A.,Reuter, H. I., Strobl, P.

This document encompasses the main outcomes and presents the conclusions based on work of the members of the Digital Elevation Model Intercomparison eXercise (DEMIX). This original work was an attempt to achieve the following goal: *Propose a scientifically sound procedure able to rank the available free and open global DEMs, taking into consideration the user's needs. The procedure should be replicable as much as possible, promoting the use of open science including the implementation of FAIR principles but otherwise considering privacy limitations when applicable.*

1.1. DEMIX Origins – The Digital Elevation Model Intercomparison eXercise

The shape of the Earth's surface, together with meteorological conditions and presence of water, both often influenced by it, determine the liveability on our planet. Knowledge about terrain and topography is therefore of fundamental importance for monitoring and understanding of almost all terrestrial ecosystems. Until about 20 years ago we had a quite varying, often coarse, and inaccurate knowledge of terrain at global scale. This changed when the Shuttle Radar Topography Mission, known as SRTM, for the first time delivered a worldwide consistent and fairly fine resolution Digital Elevation Model (DEM), at least for latitudes below 60 degrees.

About a decade after the emergence of the first consistent (quasi-)global DEM, several similar products entered the scene and through artefact removal, merging and re-processing, this multitude of DEMs is even further augmented. Due to increasing adoption of free and open data policies and globally operating Cloud based exploitation platforms, these data sets are now becoming available to large user communities worldwide.

Today we find ourselves in a situation, in which it is often difficult, even for experts, to decide which DEM might be the most accurate or appropriate for a certain application or region. At the same time the dependency on global, accurate, and high-resolution DEM information is larger than ever, given their fundamental role in many domains dealing with high quality geoinformation products. A prominent example of this dependency is the pre-processing of remotely sensed Earth Observation imagery, necessary to arrive at measurements which are accurately georeferenced, bio-geophysically interpretable and inter-comparable between sensors, be they on the ground, in the air, or in space. Global Earth Observation initiatives such as the United States (US) Landsat mission or the European Union's (EU) Copernicus programme therefore have a vital interest in using the best available DEM in their processing chains and in providing access to these data also for their user communities in downstream utilisation and value adding.

The European Commission (EC), responsible for the Copernicus programme, had identified this need and tasked the European Space Agency (ESA) to procure a state-of-the-art global DEM for program wide use. This "Copernicus DEM" plays a central role in the reference data framework that determines the quality and traceability of all Copernicus products and services. Considering that Copernicus is among the largest providers of Earth Observation imagery and products globally, substantial user attention and requests for access were expected. In addition, a new DEM or an update to an existing one raises questions about its quality, consistency, and comparability with respect to other products and the de-facto standards such as the initially mentioned SRTM based suite of DEMs. The multitude of technologies used to produce DEMs and the differences in scope and scale often render this a difficult task, which is further complicated by the fact that neither quality nor consistency of a DEM are exhaustively and unambiguously defined and easily measurable. And while it is obvious that a systematic benchmarking of several global DEMs against each other will be of great use, the parameters, methods, and algorithms to be applied in such an exercise were all but obvious at that time.

To address this need, the Joint Research Centre (JRC), EC's in-house science service, teamed up with the International Society for Geomorphometry (ISG) in organising a workshop in January 2019 dedicated to the benchmarking of global DEMs. The list of invitees was composed of leading experts and major stakeholders including providers of global DEM data and application scientists. One of the recommendations of this workshop was to bring the task to an international level and call for global participation. A respective proposal was presented to the Committee on Earth Observation Satellites (CEOS) Working Group on Calibration and Validation (WGCV) at its 2019 meeting. WGCV agreed to re-activate its Terrain Mapping Subgroup (TMSG) and mandated it to set-up and conduct together with the ISG the Digital Elevation Model Intercomparison eXercise (DEMIX).

After an open call for participation publicised by CEOS-WGCV and ISG, a group of just under 50 volunteers was formed in early 2020 and started working under the restrictions of the Covid Pandemic entirely remotely. While initially planned to take only about 18 months the magnitude of the task, the diversity of the group and the difficulties of working exclusively remotely protracted the work until summer 2023. By then the scientific results and peer-reviewed articles produced were consolidated to be presented at the Geomorphometry 2023 conference and summarised in this report.

1.2. Set-up and Structure of DEMIX

Within the scope of the Committee on Earth Observation Satellites (CEOS) (https://ceos.org/), there are several Working Groups which focus on different areas related to the measurement of objects and processes from space based remote sensing techniques. The CEOS mission is "to ensure international coordination of civil space-based Earth observation programs and promote the exchange of data to optimize societal benefit and inform decision making for securing a prosperous and sustainable future for humankind". This mission is achieved through the international collaboration of specialists, practitioners, and organisations with similar goals. It is supported by space agencies from around the world to provide leadership and generate new and innovative ideas and initiatives to sustain the CEOS activities.

The work presented in this document stems from the CEOS Working Group on Calibration & Validation (WGCV). One of the important goals of this CEOS WG is to "ensure long-term confidence in the accuracy and quality of Earth Observation data and products". In particular, the group focused on validation aspects related to global DEMs which were produced using different Earth Observation space-based techniques. CEOS defines validation in the following manner: the process of assessing, by independent means, the quality of the data products derived from EO system outputs.

In order to provide an added value with respect to the various existing but mostly regional or local DEM comparison studies, DEMIX sought to have a global outreach to include all one arc second DEMs with quasi-global coverage and are available under a free and open data policy (FAIR). This includes:

1. SRTM v3 (Farr et al. 2007) to assess its adequacy and the need for a re-evaluation of the results from the many studies that have used it;
2. ASTER-GDEM v3, (METI, NASA) (Abrams, 2020);

3. ALOS (AW3D30) v3.2 (JAXA, free and open version of the Japanese ALOS based global DEM) (Tadono, 2016) ;
4. NASADEM (Crippen et al., 2016) a reprocessed version of SRTM;
5. CopDEM (EC/ESA, free and open version of WorldDEM™, the commercial version of TanDEM-X produced by Airbus Inc.). Copernicus DEM - Global and European Digital Elevation Model (CopDEM); https://doi.org/10.5270/ESA-c5d3d65
6. FABDEM (Hawker et al. 2022).

The Digital Elevation Model Intercomparison eXercise (DEMIX) Working Group (WG) took up this challenge focusing on global DEMs. Our WG addresses the need for standardizing the ways that different data sources are combined to ensure interoperability among existing and future Earth Observing systems. Furthermore, a method was developed for the intercomparison of global DEMs to help users and practitioners rank global DEMs based on specific criteria that are most relevant to their requirements. The DEMIX WG is made up of three sub-groups which each tackled different aspects related to the defined exercise through the exchange of technical information and expertise and joint experiments.

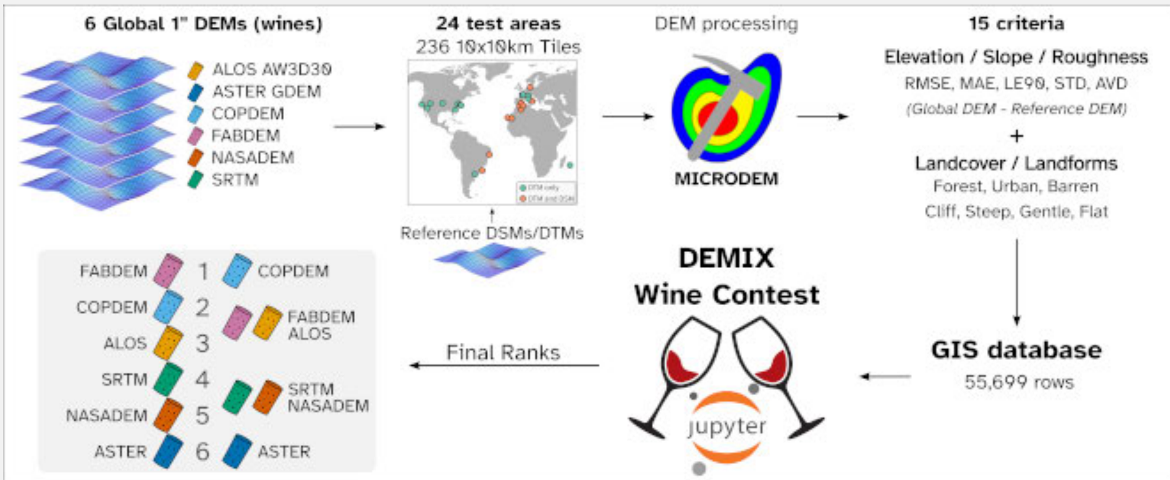
Established in 2020, the first DEMIX sub-group (SG1) focused on the needs of terminology and the analytical basis for the intercomparison of global DEMs. Chaired by experts in the field, Peter Guth and Igor Florinsky, the SG1 focused on defining common terminology and analytical basis in order to facilitate the cooperation of experts from different domains, and from around the world, to the intercomparison of global DEMs. Global DEM users come from a wide range of scientific and technical fields and therefore it was imperative to bring everyone into the same context. This was achieved by a thorough review of existing terminology and its use in literature in order to propose a consistent set of terms which also matches with the analytical basis of geomorphometry. This work culminated in a peer reviewed open publication (Guth et al. 2021c) that continues to support the DEMIX WG. The SG1 continues to provide valuable input into the DEMIX WG through the review of SG2 developments ensuring coherence and consistency with the agreed set of definitions. Some of the work attained in SG1 are presented in the next section.

The DEMIX subgroup 2 (SG2) is entitled Algorithms and Software because the subgroup focuses on the development of the intercomparison of global DEMs. The SG2 is chaired by Conrad Bielski and Carlos Grohmann who worked with the members to develop a strategy for the intercomparison of global DEMs. The goal of identifying methods and implementations for the quantitative and/or qualitative intercomparison of DEM products was realised through the implementation of the wine contest. The wine contest delivered on the identified requirements of openness, reproducibility, adaptability, and statistical rigour. This was achieved through the use of open data and the development of open-source implementations. Most importantly, the SG2 was able to deliver significant global DEM intercomparison results which are being shared with the scientific community for discussion and further development. The outcomes of the SG2 work are presented in detail in this report.

Box 1. The Wine Contest

The DEMIX members during the development of the method presented in this report adopted the name 'wine contest' in order to use a catchy name for the statistical analysis envisioned for the intercomparison of global DEMs. Such a name and the associated analogy is not new as many statistical works take advantage of this analogy including Olkin, I., Lou, Y., Stokes, L. and Cao, J., 2015. Analyses of wine-tasting data: A tutorial. Journal of Wine Economics, 10(1), pp.4-30. http://dx.doi.org/10.1017/jwe.2014.26 - "Many studies produce data consisting of rankings or ratings of n objects by m judges…" . Another example is from Granato, Daniel, Verônica Maria de Araújo Calado, and Basil Jarvis. "Observations on the use of statistical methods in food science and technology." Food Research International 55 (2014): 137-149. http://dx.doi.org/10.1016/j.foodres.2013.10.024 where the following is stated: The Friedman (1937, 1939) test is used to determine differences in central location (median) for analysis of trials with one-way repeated measures having three or more dependent samples and is also based on ranking of data. Dependent samples might include, for instance, the assessment scores of k taste panelists who are judging n independent samples of wine — are the scores given by each panelist consistent or is there a significant difference between panelists and, if so, does this differ between the wine samples?".

For more details on the applying the 'wine contest' to the intercomaprison of global DEMs, read on!



Source: Carlos Grohmann

The third DEMIX subgroup is 3 (SG3) chaired by Serge Riazanoff. This ongoing activity is developing online tools which should make it easier for DEM users and practitioners to take advantage of the developed wine contest method. A section in this report is dedicated to this work.

We believe that the DEMIX WG has achieved a significant milestone with the publication of the outputs from the SG2. As evidenced through the publication of this report together with the peer reviewed article presenting the DEMIX wine contest applied to the intercomparison of six global DEMs, users of global DEMs can use the tools and outcomes to compare and select the most appropriate digital elevation model for their applications. Furthermore, work demonstrates the superiority of the Copernicus DEM product over the other global DEM products compared.

The DEMIX activities will continue in the short term to develop more online tools through SG3 and expand the range of criterion applicable to the DEM user community. These may also be proposed in the future within the intergovernmental Group on Earth Observations (GEO) and Global Earth Observation System of Systems (GEOSS) frameworks.

1.3. List of DEMIX Affiliations

Over the course of the DEMIX, many experts and practitioners have provided valuable inputs and discussions which we would like to acknowledge here.

Table 1. Members and their affiliations who provided inputs to DEMIX activities

First Name	Last Name	Affiliation
Alynne	Affonso	University of São Paulo, Institute of Energy and Environment Spatial Analysis and Modelling Lab (SPAMLab), Brazil
Clement	Albinet	European Space Agency (ESA)
Giuseppe	Amatulli	School of Forestry & Environmental Studies, Center for Research Computing, Yale University, New Haven, USA
Francisco Javier	Ariza-López	University of Jaén, Department of Cartographic, Geodetic and Photogrammetric Engineering, Paraje de las Lagunillas S/N, Spain
Jonathan	Bamber	University of Bristol, UK
Santhi Sree	Basavaraju	National Remote Sensing Center, ISRO, Hyderabad, India
Conrad	Bielski	EOXPLORE UG, Germany
Bodo	Bookhagen	Institute of Geosciences, University of Potsdam, Germany
Claudia Cristina	Carabajal	Hexagon US Federal @ NASA/Goddard Space Flight Center, USA
Beata	Csatho	University at Buffalo, Buffalo, NY, USA
Axel	Corseaux	Visioterra, France
Jeffrey	Danielson	U.S. Geological Survey, Earth Resources Observation & Science (EROS) Center, USA
Luca	De Felice	European Commission, Joint Research Centre (JRC), Ispra, Italy
Christian	Depraetere	Institut de Recherche pour le Développement (IRD), Maison de la Télédétection, Montpellier, France
Ernest	Fahrland	Airbus Intelligence
Lang	Feng	China Shanghai Aerospace Electronic Technology Institute, China
Igor	Florinsky	Institute of Mathematical Problems of Biology, the Keldysh Institute of Applied Mathematics, Russian Academy of Sciences
Shannon	Franks	U.S. Geological Survey (USGS), USA
Ferran	Gascon	European Space Agency (ESA)
Dean	Gesch	U.S. Geological Survey (USGS), USA
David	Goodrich	USDA-ARS, Southwest Watershed Research Center, Tucson, AZ
Carlos Henrique	Grohmann	Institute of Astronomy, Geophysics and Atmospheric Sciences, Universidade de São Paulo, Brazil
Kévin	Gross	Visioterra, France
Peter	Guth	US Naval Academy (Retired), USA
Adrian	Guzman	Mexican Space Agency, Mexico

Laurence	Hawker	University of Bristol, UK
Tomislav	Hengl	OpenGeoHub Foundation, Netherlands
Virginia	Herrera Cruz	Airbus Defence and Space GmbH, Germany
Michelle	Hofton	Department of Geographical Sciences, University of Maryland College Park, USA
Martin	Huber	German Aerospace Center (DLR), Earth Observation Center (EOC), German Remote Sensing Data Center (DFD), Land Surface Dynamics (DFD-LAX), Germany
Zhenhong	Li	Newcastle University, UK
Carlos	López-Vázquez	Laboratorio LatinGEO IGM+ORT, Universidad ORT Uruguay, Uruguay
Huaxing	LU	Geomatics S&T college, Nanjing Tech University, China
Kevin	Macdonald	Loxcel Geomatics (business), Canada
Rodolfo	Méndez Baillo	Institute of Land Surveying of the Faculty of Engineering, Universidad de la Republica, Uruguay
José Luis	Mesa-Mingorance	Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría. Universidad de Jaén, Spain
Stefan	Mühlbauer	Airbus Defence and Space GmbH, Germany
Jan-Peter	Muller	Mullard Space Science Laboratory, Dept. of Space and Climate Physics, University College London, Holmbury St Mary, Surrey, UK
Diego	Nascimento	Universidade Federal de Goiás, Brazil
Enrique	Nicolás	INTA (National Aerospace Institute of Spain) - Earth Observation and Atmosphere Department - User Segment and Exploitation Area, Spain
Leandro	Parente	OpenGeoHub Foundation, Netherlands
Paolo	Pasquali	Sarmap SA, Switzerland
Luntadila	Paulo	Fathom, UK
Ben	Purinton	University of Potsdam, Germany
Cheng-Zhi	Qin	Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences, China
Juan Francisco	Reinoso-Gordo	Department of Architectural and Engineering Graphic Expression, School of Civil Engineering, School of Architecture, University of Granada, Spain
Raj	Rengarajan	U.S. Geological Survey (USGS), USA
Hannes	Reuter	EUROSTAT – Statistical Office of the European Union
Serge	Riazanoff	VisioTerra, France
Paola	Rizzoli	Microwaves and Radar Institute, German Aerospace Center – DLR, Germany
Jie	Shan	School of Civil Engineering, Purdue University, USA
David	Small	University of Zurich, Switzerland
Fabiana	Soares	UFSCAR Campus Sorocaba, Brazil

Pierre	Soille	European Commission, Joint Research Centre (JRC), Ispra, Italy
Jason	Stoker	US Geological Survey, USA
Peter	Strobl	European Commission, Joint Research Centre (JRC), Ispra, Italy
Koyel	Sur	Punjab Remote Sensing Centre, India
Takeo	Tadono	Japan Aerospace Exploration Agency (JAXA), Japan
Sebastiano	Trevisani	University Iuav of Venice, Venice, Italy
Andreas	Uttenthaler	GAF AG, Germany
Adriaan	Van Niekerk	Centre for Geographical Analysis, Stellenbosch University, South Africa
Birgit	Wessel	German Remote Sensing Data Center, German Aerospace Center, Germany

Source: CEOS-WGCV-TMSG

1.4. How to Read this Document

Due to the size and breadth of the work achieved by the DEMIX, this document encompasses many technical and practical outcomes which can interest a variety of readers. Therefore, the report is divided into different sections which each focus on relevant topics of DEMIX and the intercomparison of global DEMs. The authors suggest choosing different paths depending on the requirements of the reader.

Path A – I am only interested in the DEM comparison outcomes and conclusions:

1. Check out the final DEMIX wine contest rankings in **Figure 13**;
2. Read about the latest DEMIX developments (Section 12)
3. Read the Conclusions and Recommendations Section (Section 13)

Path B – I want to understand what method was applied to rank the global DEMs:

1. Section 3 provides details about the preparation of the data and the sampling applied;
2. The DEMIX Wine Contest method is presented in Section 4 ;
3. Manuals for the software tools applied to prepare the data for the intercomparison and computing the necessary statistics are found in Section 5 and 6 respectively.

Path C – I want to learn about potentially applying the method to other domains for quantitative and qualitative comparison:

1. Start with the Section 4 where the DEMIX wine contest is described;
2. Read Section 9 to better understand the framework and statistics involved;
3. Section 10 provides a discussion on several topics related to the implementation of the DEMIX wine contest.
4. Section 7 presents a qualitative wine contest implementation and presents alternative visualisations of the DEMIX results.

Otherwise, please choose what sections interest you the most based on the following section descriptions.

Section 1 – This section introduces the DEMIX and provides an overview of the membership that assisted in the discussion, development, and implementation of the methods.

Section 2 – This section presents information about the Digital Elevation Model (DEM) including terminology and definitions used. Provides a technical introduction for readers not familiar with this type of environmental and geospatial information.

Section 3 – This section provides details related to the preparation of the data and the manner in which samples were prepared are described.

Section 4 – This section explains how the DEMIX wine contest was implemented and the results of the intercomparison.

Section 5 – This section presents the MICRODEM software used in the preparation of the data used for the intercomparison and provides details for users to take advantage of it for testing and/or reproducing the tables.

Section 6 – This section presents how the statistical analysis applied for the intercomparison of the six global DEMs was implemented in a Jupyter Notebook and provides the user with details to run the code on their own.

Section 7 – This section shows how the DEMIX group took advantage of the wealth of data produced for the wine contest and provides other visualisations to support the rankings. Furthermore, a qualitative example is demonstrated using the DEMIX wine contest.

Section 8 – This section presents work that supports the result of the intercomparison including references to other work and further analysis that was not presented in any other publications.

Section 9 – This section delivers more details related to the wine contest framework presented as part of the DEMIX intercomparison are presented in order to demonstrate the potential not only to expand the potential for comparing other DEM characteristics but also to extend to other types of information.

Section 10 – This section discusses associated technical ideas linked to the implementation of the wine contest.

Section 11 – This section presents auxiliary information on work that was carried out in the frame of the DEMIX.

Section 12 – This section presents the latest developments related to DEMIX beyond the report that discuss novel outputs and updated references.

Section 13 – This final section presents the conclusions and recommendations associated with the intercomparison of the six global DEMs.

The references and acronyms are provided at the end of the report.

2. Digital Topography Related Terminology

Section contributors: Conrad Bielski, Peter Guth, Sebastiano Trevisani, Carlos Grohmann, Peter Strobl

2.1. Introduction

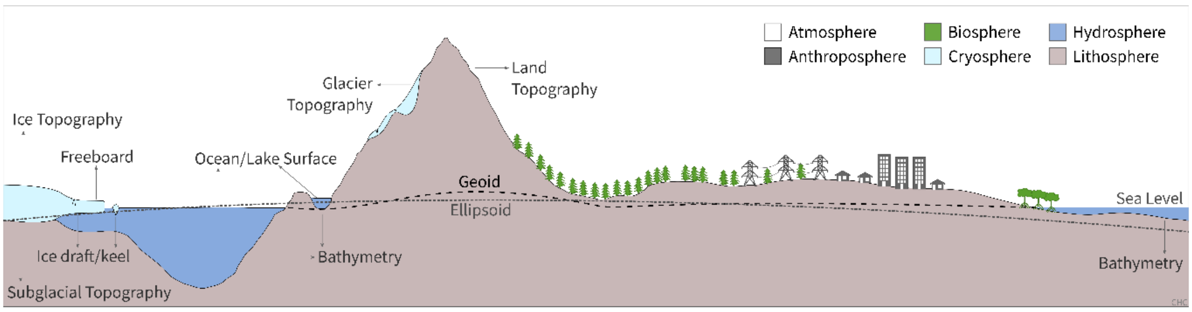
Information about the surface of the earth is applied across a wide range of topics and themes because this data is easily integrated into geographic information systems (GIS) and any geospatial service that requires it. From climate science to archaeology, to environmental science and routing applications, digital topography is being used. However, across all these different disciplines and applications the understanding of the underlying data and information varies significantly. Therefore, this section will provide specialist terminology relevant to digital elevation models (DEMs) which were adopted within the DEMIX and to help the reader better understand the developed framework of the DEMIX wine contest. The terminology presented here is specific to digital topography and should be considered authoritative when applied across other domains and applications which take advantage of this valuable geospatial information.

2.2. Digital Elevation Models: Terminology and Definitions

The first effort of the DEMIX group was writing a review paper on DEM terminology (Guth et al., 2021). While this work was general in scope and not focused on DEM intercomparison, it laid out a number of points required to understand the challenges in intercomparison.

The earth can be regarded as a series of "spheres", and a DEM can generally be regarded as the top of the lithosphere or the bottom of the atmosphere (**Figure 1**). The creator of a DEM must deal with the other spheres between the lithosphere and the atmosphere. Depending on the use of the DEM, users might want the other spheres removed; for example, a DEM like ETOPO delivers a seamless global surface without the water. However, removing the water is not easy. Sonar mapping is slow and expensive, and much of the best bathymetry is generalized and derived from measuring the marine geoid. Glaciers, vegetation, and buildings present similar challenges. We might like to see them removed but actually making that happen might not be possible or easy.

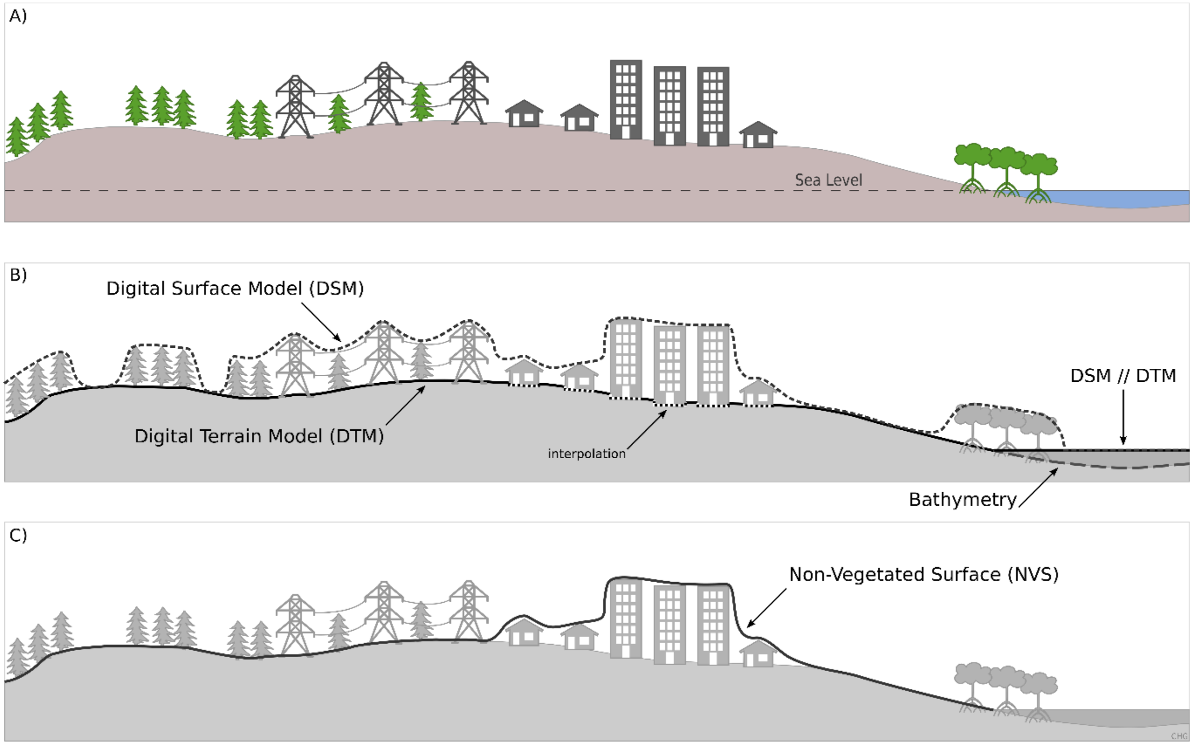
Figure 1. Spheres of the earth, and how they relate to DEM creation



Source: Guth et al. (2021)

A DEM is a general term for a digital representation of elevations (or height) of a topographic surface in the form of a georectified grid (Guth et al. 2021c). If the DEM depicts the lower surface of the atmosphere, it is a digital surface model (DSM) or first return surface if collected by lidar (**Figure 2** B). A digital terrain model (DTM, or bare earth DEM) depicts the top of the lithosphere, with some variation on the treatment of ice and water. For high resolution current lidar DEMs, algorithms can easily remove trees and create a realistic ground surface, but buildings prove much more difficult (see section 2.4 below). This leads to the non-vegetated surface (NVS), easily created from lidar point clouds and preferable for some users compared to the building ghosts in a DTM.

Figure 2. Types of DEM based on surface/terrain of interest.: from Guth et al. (2021)

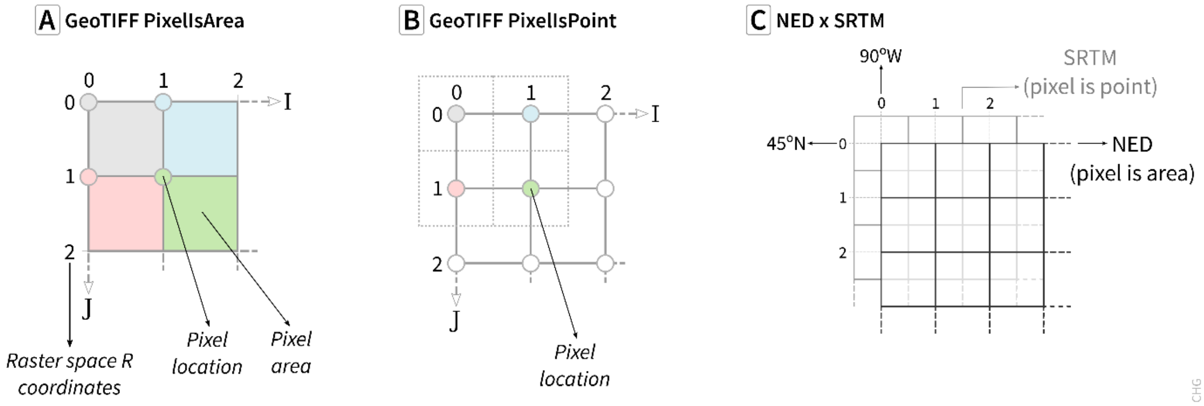


Source: Guth et al. (2021)

The collection sensors used by the global DEMs determine what surface they can collect, and they have limited penetration through foliage. They are thus closer to a DSM than a DTM, but the penetration will vary depending on whether the sensor uses optical or radar energy, and the radar band used. Radar penetrates foliage more than optical sensors. Of the six DEMs we consider, only FABDEM comes close to being a DTM because it attempted to remove the vegetation and buildings from CopDEM.

The earliest widespread DEMs were the US Defence Department's digital terrain elevation data (DTED) and the US Geological Surveys ASCII format DEMs, each with its own format. Both talked about the elevations as representing "points", but the collection methods did not allow them to actually represent the elevation at a specific point. As DEMs later adopted the Geotiff format, used for all 6 DEMs considered here, two representations were available for the pixel geometry, using tag 1025. Based on common usage in the remote sensing community, the pixel is area geometry used the coordinates for a corner of the pixel. Continuing the usage for DTED and the USGS, pixel is point representation uses the centroid of the pixel. Applied to one second global DEMs, the points given in the Geotiff file have a one half pixel offset between ALOS (pixel is area) and the other five DEMs. Since all of these DEMs use a region approximately the size of the pixel, the actual pixels areas only overlap for ¼ of the area (**Figure 3** C). While some users would like to regard tag 1025 to indicate the sampling strategy used to create the DEM, the Geotiff standard does not mention this, and the actual usage for these DEMs does not reflect the fact that all are using the posted elevations to represent an area.

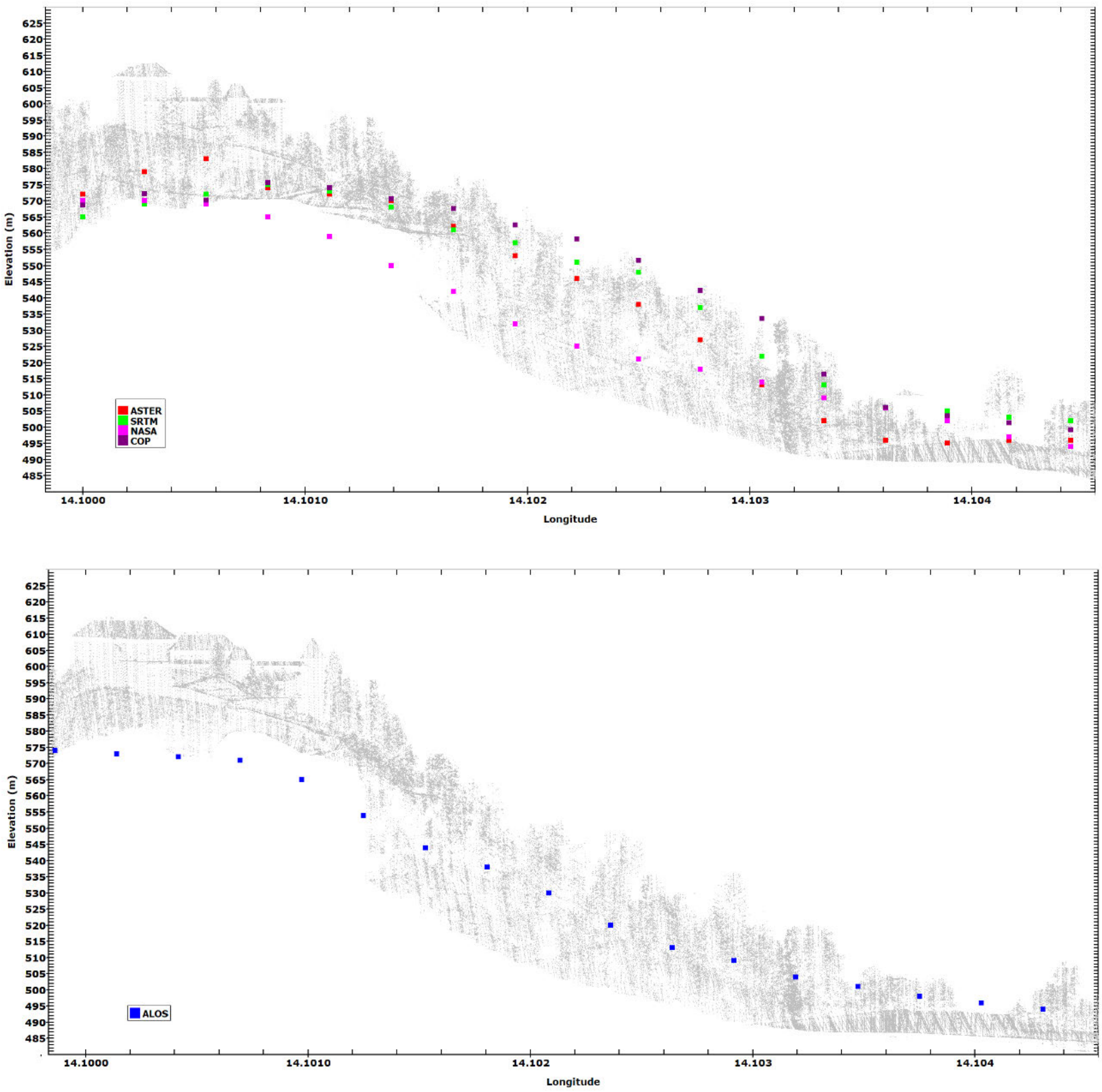
Figure 3. GeoTIFF pixel representation used for global DEMs.



Source: Guth et al. (2021)

Figure 4 shows a one pixel wide slice through a lidar point cloud in Bled, Slovenia. Lidar surveys typically have 2–20 returns per square meter, or 1,500–10,000 points corresponding to each elevation posting in a 1" DEM. The global DEM elevations should lie within the envelope of the lidar returns but often lie above or below all the sampled points in the cloud. Note that the lower panel for ALOS, compared to the upper panel for the other DEMs, has a half-pixel north-south shift for the lidar points, seen by the differences in the light grey lidars returns, and a half-pixel east -west shift in the 1" elevations seen in the coloured symbols. The ALOS DEM cannot be directly superimposed on the other DEMs, which led to important design constraints for DEMIX.

Figure 4. One pixel wide through a lidar point cloud and the locations of the global DEM pixels. This figure was created before the release of FABDEM.



Source: Guth et al. (2021)

2.3. Link to Terminology of Neighbouring Domains & Standards

The INSPIRE Directive came into force in May 2007 establishing an infrastructure for spatial information in Europe to support Community environmental policies. INSPIRE is based on the infrastructures for spatial information established and operated by the Member States of the European Union. The Directive addresses 34 spatial data themes needed for environmental applications, with key components specified through technical implementing rules. This makes INSPIRE a unique example of a legislative “regional” approach. While the DEMIX work included the current Copernicus DEM as one of the elevation models under consideration, it was integrated into the intercomparison as a global dataset and not just an EU dataset. Therefore, INSPIRE requirements were not taken into account because they were out of scope for the exercise. The underlying data intercomparison does provide valuable information that relates to many of the environmental themes and the terminology must be understandable and inter-operable with the global community working in these domains.

Of the 34 spatial data themes identified within the INSPIRE Directive, there are three which are relevant to the DEMIX work. From Annex 1, both the coordinate reference systems and geographical grid systems themes are applicable whereas from Annex 2, the elevation theme is relevant.

From the perspective of the DEMIX framework, the coordinate reference system (CRS) is of paramount importance because it defines the manner in which the spatial information is referenced. For geospatial data and information to be shared and intercompared this information is not only needed but must also be applied correctly if any transformations must be made. For this reason, the theme is part of the INSPIRE Annex 1 (URI http://inspire.ec.europa.eu/theme/rs) because it is a fundamental piece of information that uniquely references the geospatial data. The coordinate reference systems theme also refers to 3D space including a height parameter based on a geodetic horizontal and vertical datum. For interoperability between geospatial datasets where the height is important, this information plays an important role.

This particular theme has been developed for geographic information in general because it is based on the ISO 19111 specification which is not applicable to precise positioning requirements. These would require additional parameters within the specification. The INSPIRE theme is sufficient for referencing the global DEMs used in the intercomparison but may not necessarily be applicable to the reference data. As geospatial measurements become more accurate and their spatial resolution improves, this theme will need to be revisited and updated.

The second Annex 1 theme related to the use of DEMs is the geographical grid system which relates to the manner in which geospatial measurements are stored in a grid format. The INSPIRE definition is the following: "Harmonised multi-resolution grid with a common point of origin and standardised location and size of grid cells." which is a subset of the DEMIX definition of the (Geo)Rectified grid: "Grid for which there is an affine transformation between the grid coordinates and the coordinates of an external coordinate reference system (http://inspire.ec.europa.eu/theme/gg). If the coordinate reference system is related to the Earth by a datum, the grid is a georectified grid." (Guth et al. 2021c). Within the INSPIRE context, the grid is defined for the Pan-European context with a standardised location and size of grid cells, as well as projection: EPSG 3035 (https://epsg.io/3035). While the Copernicus DEM is made available in an INSPIRE compliant format, it was not used within the context of this study because the intercomparison was global and the INSPIRE ETRS89/LAEA projection would not scale appropriately. For the DEMIX intercomparison, the team decided to keep the DEM producers original geographical grid system in order to not introduce any potential alterations during grid transformations. Harmonization between reference data and global DEMs was achieved by converting the reference data to the target global DEM grid system.

The INSPIRE Annex 2 theme Elevation (http://inspire.ec.europa.eu/theme/el) is directly linked to the work performed within DEMIX. This theme is defined in the following manner: "Digital elevation models for land, ice and ocean surface. Includes terrestrial elevation, bathymetry and shoreline.". This elevation INSPIRE theme also includes the purpose of a Digital Elevation Model which is to provide an elevation property with reference to a specific origin. While the general understanding of what a DEM is similar to the DEMIX understanding, i.e. digital topography (Guth et al. 2021c), the wording of particular definitions of DEM related terms do differ. The global intercompared DEMs focused on land which could include ice found on the land surface. However, the DEM products compared did not include any information about the ocean surface, bathymetry or shorelines. Such elevation datasets (if available) could be intercompared within the DEMIX wine contest framework but were not directly investigated at this time.

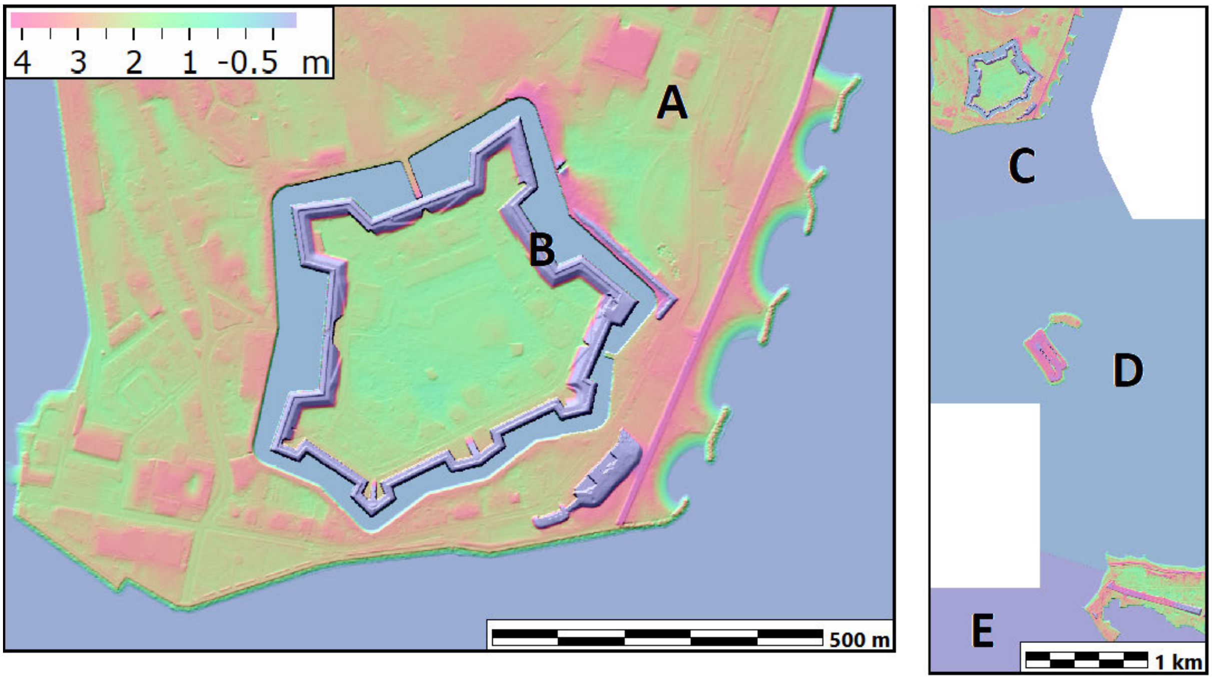
The INSPIRE definitions have been integrated into legal acts such as the INSPIRE Directive and Implementing Rules which places their definitions into a different scope than purely academic. From a policy perspective, the DEMIX wine contest framework could be adapted to help implementing organisations choose the most relevant datasets for policy monitoring and reporting requirements especially given a set of rules against which an intercomparison can be made.

2.4. Challenges in DTM Creation

In general, a DTM user expects that it represents the elevation of terrain after filtering out vegetation and anthropic structures and smoothly extending the topographic surface through any resulting gaps using the remaining neighbouring ground elevation points. For airborne lidar DEMs this process is generally obtained by 1) identification of non-ground features; 2) filtering out of these features; 3) reconstruction of the surface by means of interpolation. A wide range of algorithms and tuning parameters can be used for the process, and these are often proprietary to the vendor collecting and processing the data. There is general agreement, for example, that buildings and structures such as bridges and viaducts should be removed, since they inherently create a sharp discontinuity in the surface. However, other features, such as roads, embankments, concrete channels, and check dams are not completely filtered out and are generally represented in a high resolution DTM.

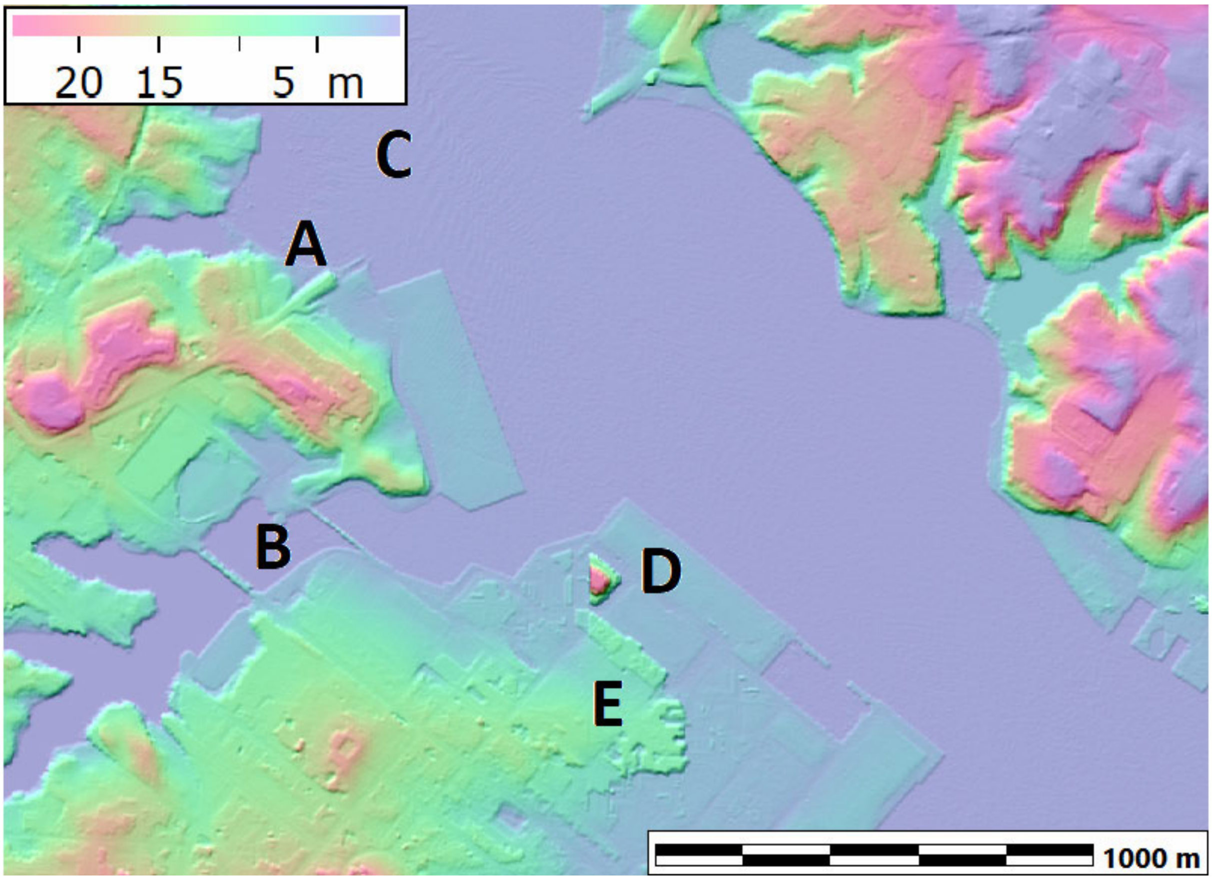
Guth (2018a, 2018b) discussed what a DTM should show, and gave examples of the challenges presented by castles, or any large buildings. **Figure 5** to **Figure 7** give examples of the anomalies that can be encountered.

Figure 5. A 1 m DTM of Fort Monroe with the following anomalies: (A) ghosts of buildings and roads clearly present; (B) "castle" not removed as a building; (C, D, D) water shown with three different elevations, from north to south -0.75 m, -0.66 m, and -0.90 m. The DTM does a very good job of edits to put the most at constant elevation, where the source lidar did not do a good job.



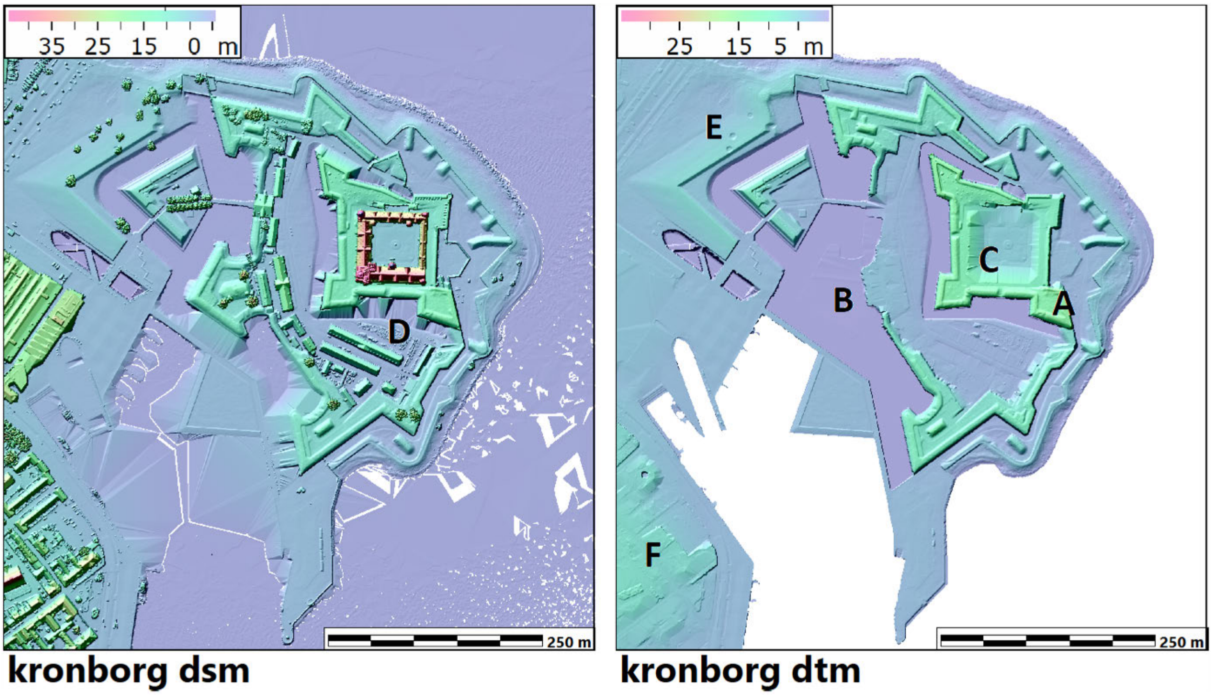
Source: Guth et al. (2021)

Figure 6. 1/9 arc second DTM showing Rickover Hall, with the following anomalies: (A) bridge abutment ends abruptly; (B) two smaller bridges not removed, which creates dams when trying to use the DTM for coastal flooding models; (C) boat wakes and waves present in some of the water; (D) only half of the large building removed;(E) many ghosts of buildings and roads remain.



Source: Guth et al. (2021)

Figure 7. 1 m DSM and DTM for Kronborg Castle. Anomalies in the DTM: (A) most of the castle remains, but (B) part of the castle transformed to water; (C) triangulation anomalies present inside the castle, similar to those (D) in the moat in the DSM; (E) pits in the DTM from removal of some trees; and (F) many ghosts of buildings and roads remain.



Source: Guth et al. (2021)

3. DEMIX Intercomparison Preparation

Section contributors: Peter Guth, Carlos Grohmann, Peter Strobl

3.1. Definitions

The following terms are defined in order to facilitate the presentation:

- Test area: a contiguous region that contains one or more DEMIX tiles, each with no more than 25% data voids.
- DEMIX tile: nominal 10 x 10 km region for statistical comparisons (Guth et al., 2023a).
- Source DEM: high resolution data from a national mapping agency or equivalent, with 1-5 m grid spacing generally on a UTM or similar grid.
- Reference DEM: aggregated source DEM (DTM and separate DSM if source available) at 1 arc second grid spacing, and both pixel-is-point and pixel-is-area sampling to match the candidate DEMs.
- Candidate or test DEM: one of the six global DEMs at one arc second spacing.

3.2. Comparison Philosophy

We chose to make the minimal changes to the candidate DEMs, so that we could fairly compare them to what a 1 arc second should show. A 1 sec grid will show much less detail, and much lower slopes, than a 1 m DEM. We aggregated the source DEMs to match the resolution and pixel representation of the candidate global DEMs. The only change to the global DEMs was to move each to the EGM2008 vertical datum, which had no impact on horizontal coordinates. The goal is to produce DEMs where simple map algebra shows the difference between the reference and candidate DEMs.

Most comparisons used elevation metrics, despite the fact that many DEM users need derived parameters like slope which magnify errors in the elevation values, and that for many operations relative and not absolute elevation differences are more important.

3.3. 1 Arc second global DEMS

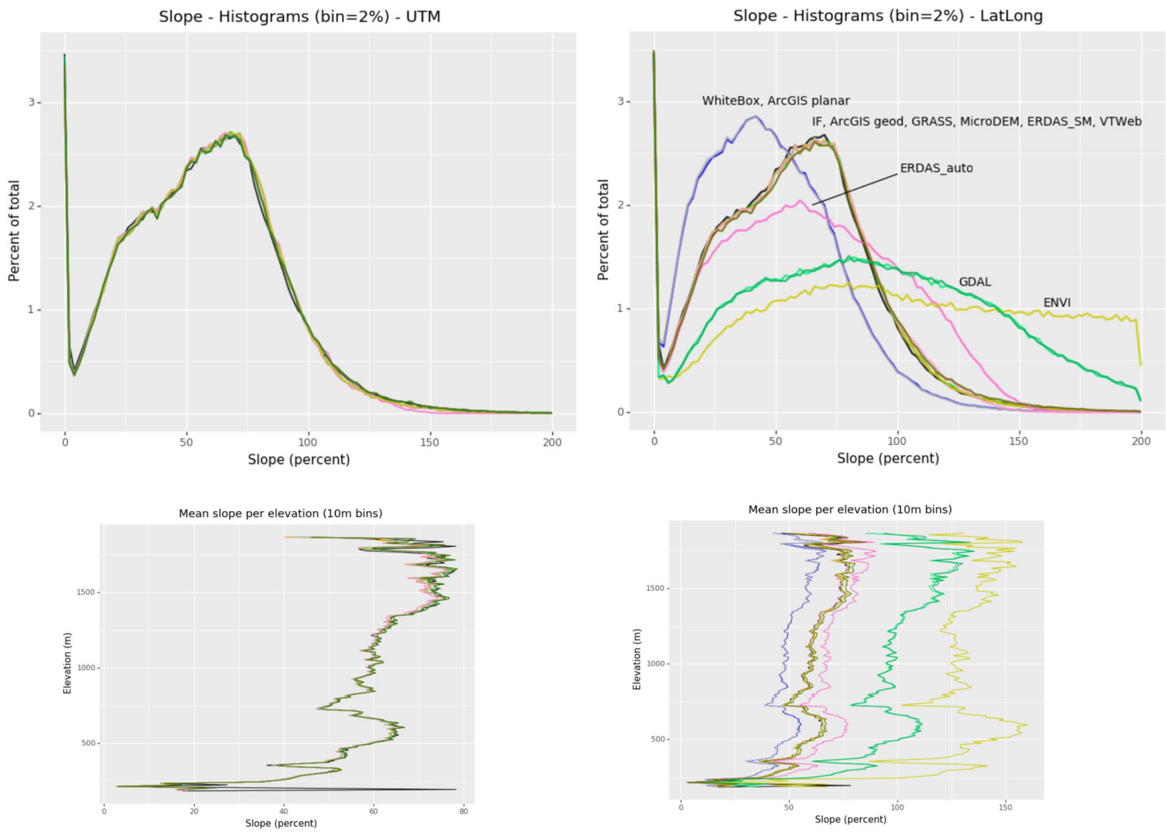
Currently six DEMs cover most of the world at one arc second resolution. Although often called 30 m DEMs, all use geographic coordinates. The pixels are approximately 30 m square at the equator but become increasingly rectangular as the horizontal spacing decreases approaching the poles. The DEMs are delivered as Geotiff files but use different pixel representations: ALOS uses pixel-is-area, which has a ½ pixel shift compared to the others. SRTM, NASADEM, CopDEM, FABDEM, and ASTER all have the same pixel-is-point geometry. The geometries only determine how software interprets the location of the values in the DEM grids, and not the underlying sampling strategy whereby each elevation is based on an area approximately equal to the 1 arc second spacing.

3.4. 1 Arc Second Global DEMs

All of the global DEMs use geographic coordinates. Guth and Kane (2021) showed that slope and similar metrics can be computed in geographic coordinates, without introducing reprojection artefacts, and software like MICRODEM, Grass, ArcGIS and WhiteboxTools can work with geographic coordinates.

Strobl and others (2021, their Figure 2 and Table 1) reported on the ability for GIS software to correctly handle slope computations for arc second DEMs. Our **Figure 8** below shows the results that led to those results using slope histograms, and elevation versus slope plots, using 9 software packages and 12 algorithms. There are different numbers of points in the two grids, and at different locations, but these two graphs do not affect the results. For a UTM grid, the results are virtually identical, and any differences probably result from numerical imprecision or how they handle whether any results can be computed along the DEM edge or where there is missing data in the 3x3 window. For the grid in geographic coordinates, the results show that several of the programs produce results that are very close to the UTM grid, and that a number of the others produce drastically different results. These results led to modification of WhiteboxTools, which now correctly handles the geographic coordinates along with MICRODEM and ArcGIS with its geodetic computation methods.

Figure 8. Slope histograms (top row) and elevation-slope plots (bottom row) for UTM (left column) and geographic (right column) DEMs for the same area, which were aggregated from high resolution lidar data.



Source: Strobl et al. (2021)

3.5. A Short History of Comparing Global DEMs

The Shuttle Radar Topography Mission (SRTM) marked the first project to map most of world's topography from space at relatively high resolution using a consistent methodology (Farr et al., 2007). SRTM flew in 2000 and released edited DEMs starting in 2004. Previously DEMs had been created from stereo photogrammetry, either directly from the imagery or indirectly from the contours on topographic maps, covered only a single country and were often not freely available (Gesch et al., 1999). SRTM produced and released a globally consistent DEM for latitudes lower than 60◦, initially at 1 arc second (1", ∼ 30 m at the Equator) grid spacing in the United States and at 3" (∼ 90 m) everywhere else, which starting in 2014 was also made available globally at 1" spacing. The U.S. Geological Survey (USGS) held a workshop in 2005, and the published papers set the standard for assessing SRTM using reference data and determining the quality of the DEM (Gesch et al., 2006). The approaches used either sparsely located control points (Rodriguez et al., 2006; Hofton et al., 2006), or comparison to reference DEMs using all points in the grid, comparing the elevations or including derived geomorphometric parameters such as slope (Guth, 2006). These workshop papers also emphasized that results vary with different land cover and terrain slope categories (Carabajal and Harding, 2005; Hofton et al., 2006).

The release of the 1" ASTER GDEM (Abrams et al., 2010) led to a large number of papers comparing SRTM and ASTER to each other and to reference data, and assessing whether SRTM or ASTER was "better" (Hayakawa et al 2008, Reuter et al 2009; Guth, 2011; Slater et al., 2011), generally finding SRTM was superior, even though at that point the 1" SRTM was still not available globally. A new milestone was achieved with the release of the TanDEM-X (Wessel, 2016; Rizzoli et al., 2017), a global DSM at 0.4" latitude pixel spacing and 0.4" to 4" in longitude. TanDEM-X data set remained limited to scientific use but produced a number of comparisons to SRTM and ASTER and ALOS (e.g. Grohmann, 2018). TanDEM-X is also the basis for the 1" and 3" DEMs of the same name and the basis for the commercial WorldDEM and therefore also for the Copernicus DEM (CopDEM – Strobl, 2020). Sequential releases of ALOS World 3D (AW3D30) (Tadono et al., 2016), NASADEM (Crippen et al., 2016), TanDEM-X and CopDEM led to additional comparisons of the freely available 1" DEMs trying to establish which was "best".

Most studies compare DEMs against each other or with reference data (such as ICESat/GNSS point elevations or lidar datasets) for a limited number of sites usually located in one country, and applying a single criterion (i.e., elevation differences), typically based upon popular statistics such as mean, median, standard deviation and root mean square error (RMSE) (Guth 2006; Hofton et al., 2006; Rodriguez et al., 2006; Grohmann 2018; Zhang 2019; González-Moradas and Viveen 2020; Gdulová et al. 2020; Guth and Geoffroy 2021; Ebinne et al. 2022; González-Moradas et al. 2023). Other criteria include comparisons of contour lines (Grohmann 2018, Ebinne et al. 2022) drainage network (González-Moradas & Viveen 2020, González-Moradas et al. 2023), and topographic profiles (Grohmann 2018, Guth and Geoffroy 2021, Ebinne et al. 2022).

In a recent extensive literature review López-Vázquez and Ariza-López (2023) considered all the papers after 2002 that perform a comparison between DEM, being at least one of them of global coverage. It should be acknowledged that aside from those considered in the review, there exist many comparisons between local DEM, as well as a number of papers involved not in a comparison but just in the accuracy evaluation of a DEM. The review considered 313 papers, denoted as the corpus hereinafter. What follows is a brief summary.

The typical DEM comparison paper analyses the accuracy of two or more DEM using sparse GNSS reference data over a single region. The most popular metric of success is the RMSE. The most popular parameter to evaluate is the elevation accuracy, but there are other options (slope, aspect, etc; see Zingaro 2021 or Nadi 2020). The region might vary wildly in size: it can include as much as the whole of Australia (Rexer and Hirt 2014), or as few as just part of a university campus (Hasan 2019).

The reference data type was also of diverse nature, from the more modern LiDAR clouds up to legacy 1:100.000 official topographic maps. In many cases the accuracy of the reference data was not mentioned at all, and it is unclear if they are substantially better than those of the candidate DEMs as required. Most studies applied a single criterion. The criteria used is typically of deterministic type, but there are examples of subjective ones. The most popular are based upon statistics such as standard deviation and root mean square error (Guth 2006; Hofton et al., 2006; Rodriguez et al., 2006; Grohmann 2018; Zhang 2019; González-Moradas and Viveen 2020; Gdulová et al. 2020; Guth and Geoffroy 2021; Ebinne et al. 2022; González-Moradas et al. 2023). Other criteria include comparisons of contour lines (Grohmann 2018, Ebinne et al. 2022) drainage network (González-Moradas and Viveen 2020, González-Moradas et al. 2023), and topographic profiles (Grohmann 2018, Guth and Geoffroy 2021, Ebinne et al. 2022).

Due to its early appearance ASTER-GDEM and STRM (90 m and 30 m versions) are the most popular ones in the corpus. Nearly 35% of the papers involve a comparison of four or more GDEM. Nearly 93% of the papers analyse test sites from a single country. Only 3% describe a comparison at the global scope, heavily connected to the type of reference data considered. Becek (2014) used elevations of 96 runways from diverse aerodromes. Grohmann (2016) used as reference data the height and location of high peaks. ICESat data now offers an opportunity to expand the list.

The literature review shows that there was an ample variety of test sites, covering many different landscapes and land uses. When more than one location in the same country is used, typically they are selected in order to be representative of a different terrain morphology. After choosing a metric (RMSE, for example) they achieve a ranking in each test site. If the ranking agrees the conclusion is pretty simple and a winner is established.

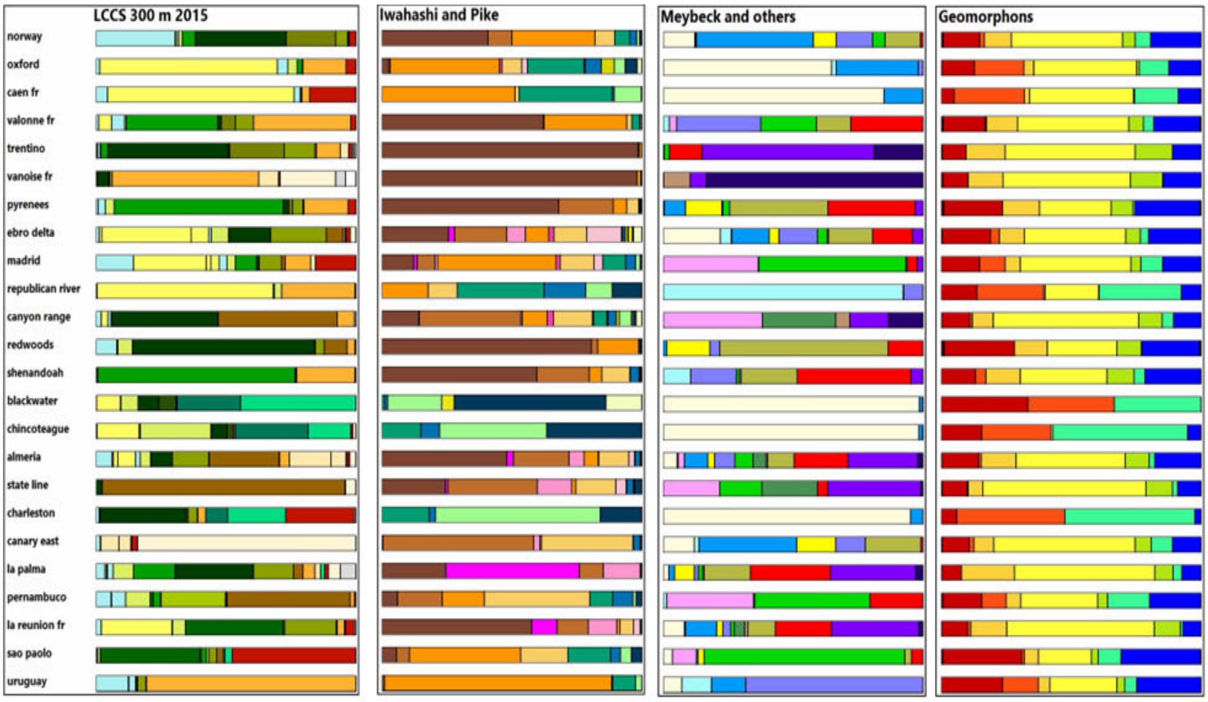
What was barely considered is what to say about the preferred DEM once the rankings differ among the test sites. The conclusions were explicitly limited to the cases under consideration, without any claim that the results are representative of those at a larger scale. Some authors (González-Moradas and Viveen 2020, Dobre et al. 2021, Zylshal et al. 2021, among others) produced an opinion table like the one to be described in this report, but they fail to provide any cue about how to handle discrepancies within it. Rawat et al. (2022) circumvented the problem by creating a weighted linear combination of the parameters and finding the ranking after sorting its values. As an ad-hoc procedure, it is unable to establish a confidence level for it.

Setting aside for the moment some exceptional cases (ICESat, GEDI, etc. among others) there is typically no reference information to use at a global scale. Thus, a sampling strategy is needed. If we have a handful of test sites where reference data is available, we can proceed as if they were selected at random in order to attain a global conclusion. In principle such hypotheses will allow establishing a confidence level as a function of the number of test sites involved, but the situation was not considered in the reviewed literature about DEMs.

3.6. Selection of Test Areas

Figure 9 presents the results from 4 classifications for each of the 24 test areas. The left column displays the land cover classes using the United Nations Land Cover Classification System with 38 categories (Buchhorn et al., 2020, Di Gregorio et al., 2000). The colour denotes the class whereas the width of the bar denotes the percentage of that particular cover class. The right three columns of **Figure 9** present three geomorphometric classifications, from left to right Iwahashi and Pike (2007); Meybeck et al. (2001), and geomorphons (Jasiewicz and Stepinski, 2013). These geomorphometric classes range between 7 to 16 categories each. Furthermore, Figure 9 reveals the challenge in trying to neatly characterize a 100 km^2 tile because no two DEMIX tiles, even adjacent ones, are alike from their surface characteristics. Our tile selection clearly represents a large range of different landscapes and provides a much better diversity of surface characteristics than previous efforts comparing global DEMs. We argue that our results are thus broadly representative of the entire earth.

Figure 9. Tile classifications for the 24 DEMIX test areas tiles in our sample, for one landcover classification and 3 geomorphometric landform classifications. Areas are arranged from north to south.



Source: Bielski et al. (2024)

3.7. The DEMIX Tile System

For each test area, one or more DEMIX tiles were extracted. A DEMIX tile covers approximately 10 km x 10 km defined on a geographic latitude/longitude grid. This grid covers the entire globe and was developed in the context of DEMIX to allow anyone to use a consistent sampling. The DEMIX tiles provide a global tessellation at a scale fine enough to produce locally significant results, coarse enough to keep the total number of tiles manageable, and nearly equal in size to allow for statistical aggregation and comparative analysis. Technical details, naming convention and grid definition files can be found in (Guth et al., 2023a).

3.8. How DEMIX Improves the Comparison of Global DEMs

The work differs quite significantly from previous studies in a number of ways and most importantly in the total combination of factors that were applied for the intercomparison and analysis of the candidate global DEMs. The following list compares and contrasts the DEMIX wine contest method to previous work:

1. Computed distributions of the differences between the reference and global DEMs for elevation, slope, and roughness. Many earlier studies just compare the distributions of the DEM with respect to the distribution of reference elevation data. Furthermore, the derived geomorphometric parameters in this study increase the effect of differences among the DEMs and the derived characteristics are often more important than the original DEM elevations for many applications.

2. Integrated five parameters for each difference distribution – most previous studies integrate fewer parameters, typically only RMSE and MAE (e.g., Grohmann, 2018). While the chosen seven parameters are correlated, all the relationships are included in the database and show how little difference the choice makes.

3. In this instance only 1” DEMs were considered; changes in DEM spacing impact DEM characteristics making reliable intercomparisons across multiple resolutions difficult. There are enough good quality global 1” DEMs applicable to land-based processes that one should only consider 3” DEMs such as MERIT (Yamazaki et al., 2017) when lower spatial resolution is explicitly required or if the user feels that MERIT resembles a better DTM than FABDEM aggregated to 3”.

4. Created reference 1” DEMs from the fine spatial resolution source reference DEMs, based on open data from national/local mapping agencies. These DEMs have significantly better spatial resolution and detail compared to the target 1” resolution. However, the choice to use 1” reference DEMs requires accepting this resolution and any limitations it imposes on the subsequent analysis.

5. No re-interpolation of the candidate global 1” DEMs was made; the original open data as downloaded from their respective sources was used in the contest. Interpolation cannot improve a DEM or create additional elevations that were not captured in the original data.

6. All computations were performed in native geographic coordinates (Guth and Kane, 2021). While there are differences in the statistics between the rectangular 1” pixels and square 30 m pixels, interpolating to UTM coordinates adds its own changes to computations done on the DEM which we avoided.

7. Explicitly created reference DEMs to consider the half pixel offset between ALOS and CopDEM (and the other lower quality DEMs) for the intercomparison, a consequence of the decision not to resample candidate DEMs. This issue is generally not discussed in the literature (or addressed by interpolation) although Purinton and Bookhagen (2021) designed their comparisons to avoid the need for DEM co-registration.

8. Both DSM and DTM reference data are considered separately. It turns out that in most of our test areas, those not dominated by urban or forested land cover, the DTM and DSM vary little. Some of the minor differences arise from the different characteristics of the source data for the reference DEM: for example, in Spain, these have different resolutions of 2 m and 5 m respectively and were created at different times.

9. The intercomparison was pixel-by-pixel at every point in the sample tiles, providing approximately 130,000 points per tile. This is a significantly greater number of points compared to studies using geodetic benchmarks, airport runways, or even measurements from laser altimeters such as GEDI and ICESat-2. In addition to providing often orders of magnitude more points for the intercomparisons, the pixel-by-pixel option provides data in

both steep and forested parts of the test area, where point-based reference data on benchmarks or airport runways are often not available (Gesch et al., 2014).

10. A large number of diverse test areas – in comparison, for example, Uuemaa et al. (2020) reported using 608 million 1 m pixels (after interpolation of the 1" test DEMs) which corresponds to only six DEMIX tiles. Purinton and Bookhagen (2021) on the other hand used the equivalent of 400 contiguous DEMIX tiles, but from a single mountainous test area.

11. The DEMIX intercomparison contest covered a wide range of land types (**Figure 9**) which, combined with the large number of test areas, greatly increases the validity of the presented results. In contrast, some studies such as Gesch (2018) and Hawker et al. (2022), focused their analysis primarily on coastal or floodplain terrain which probably accounts for our finding that FABDEM often ranked lower than the CopDEM from which it was derived, particularly in mountainous terrain. At the other elevation extreme in mountain areas, Purinton and Bookhagen (2021) preferred CopDEM while Florinsky et al. (2019) and Trevisani et al. (2023) found that relative performances between COP and ALOS change according to local morphology, with ALOS slightly better in steep slopes. None of those works provide a confidence level for their conclusions.

Previous work did not quantify the ranking as developed within the DEMIX wine contest method. Our approach quantifies the intercomparisons with additional metrics beyond RMSE or MAE and also considers elevation, slope, and roughness. Furthermore, the wine contest was applied to a representative land type sample covering a total area of 23,600 km^2 across three different continents and islands adjacent to Africa. The closest approach to a wine contest applied to DEM characterisation is Table 4 in Dobre et al. (2021), where they created a "confusion matrix" that ranked the DEMs from best to worst on four criteria, but did not compute statistical significance for the results.

3.9. Vertical Datum Transformation

Vertical datum transformation for the candidate DEMs was straightforward. Grids from NGA, obtained from the PROJ database via OSGEO4W, allow shifting those DEMs using EGM96 to the WGS84 ellipsoid and then the EGM2008 geoid. This changes only the vertical coordinates.

Vertical datum transformation for the source DEMs used two procedures:

1. For those DEMs using an equivalent for the WGS84 horizontal datum, we used grids from PROJ database via OSGEO4W or in the case of Italy from the local mapping agency, to shift the WGS84 ellipsoid and then the EGM2008 geoid.

2. For the US DEMs on NAD83, we used GDAL (which used grids from PROJ database via OSGEO4W) to shift both the horizontal and vertical coordinates. This was done on the source data before aggregation to the 1 second spacing.

3.10. Disentangling Vertical and Horizontal DEM Errors

DEM errors can result from horizontal mis-location of the feature, or from assigning an incorrect vertical elevation. This can most clearly be assigned for clearly defined points like mountain peaks, where the DEM should record the correct values with single pixel precision. If both were "correct" and giving the best result possible with their sampling areas and pixel representation, ALOS and CopDEM would have the peak located with an offset of ½ pixel both the east-west and north-south directions, and a different elevation because their respective sampling areas shared only a ¼ overlap. Identification of peaks, and perhaps ridges, and their offset from the reference DEM, represents a possible future criterion for assessing DEM which might be particularly relevant for assessing satellite registration and terrain corrections.

Away from clearly defined features like peaks, separating the horizontal and vertical errors proves very difficult. For the current DEMIX work, we assume the elevations are the best estimate at the nominal location and treat the error as vertical.

4. The DEMIX Wine Contest

Section contributors: Carlos H. Grohmann, Peter Guth, Conrad Bielski, Carlos López-Vázquez, Laurence Hawker, Dean Gesch, Sebastiano Trevisani, Virginia Herrera-Cruz, Serge Riazanoff, Hannes I. Reuter, Peter Strobl

4.1. Global DEMs Comparison

The DEMIX wine contest is a novel and flexible tool to provide users with a practical approach to intercompare a set of candidate global DEMs based on predefined criteria and a statistically sound ranking approach. The framework provides the wider geospatial community tailored recommendations regarding available DEM products that are not limited to one domain, geographic area, or landscape type. The method is flexible and customizable in relation to the specific needs and requirements of users in their particular application.

The wine contest approach was applied for the comparison of the six global DEMs that are available at a spatial resolution of 1" (ALOS AW3D30, ASTER GDEM, CopDEM, FABDEM, NASADEM, SRTM) using a variety of criteria on 236 widely distributed test tiles from 24 test areas. The method is based on four overall steps (**Figure 10**):

Step 1: Obtain high quality reference elevation data (data with significantly higher accuracy elevation measures and much smaller sampling spacing than the DEMs to be tested);

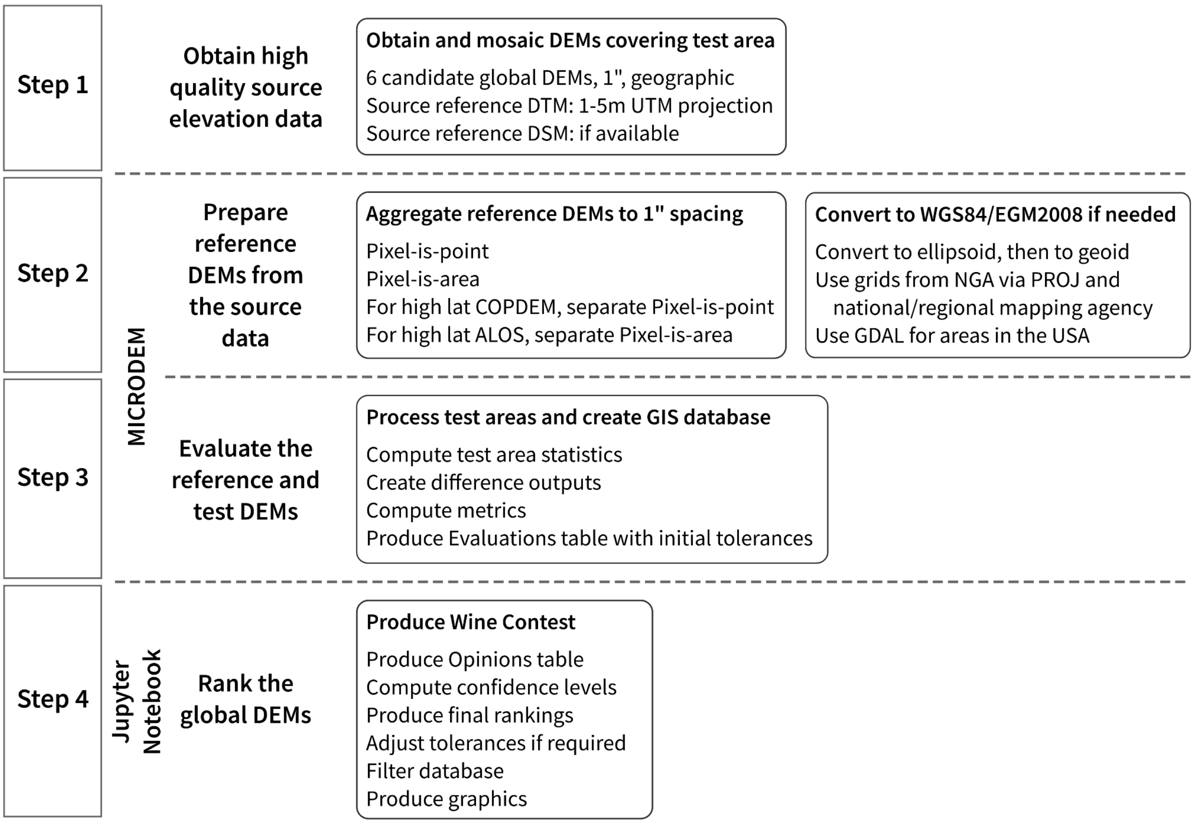
Step 2: Prepare reference DEMs from the reference data. If required, adjust the vertical datum of reference DEMs;

Step 3: Evaluate the reference and global DEM data, for every test area and for every criterion to produce the evaluations table (**Figure 12**);

Step 4: Rank the global DEMs according to the wine contest rules to produce the opinions table. Produce final rankings based on chosen statistical confidence levels.

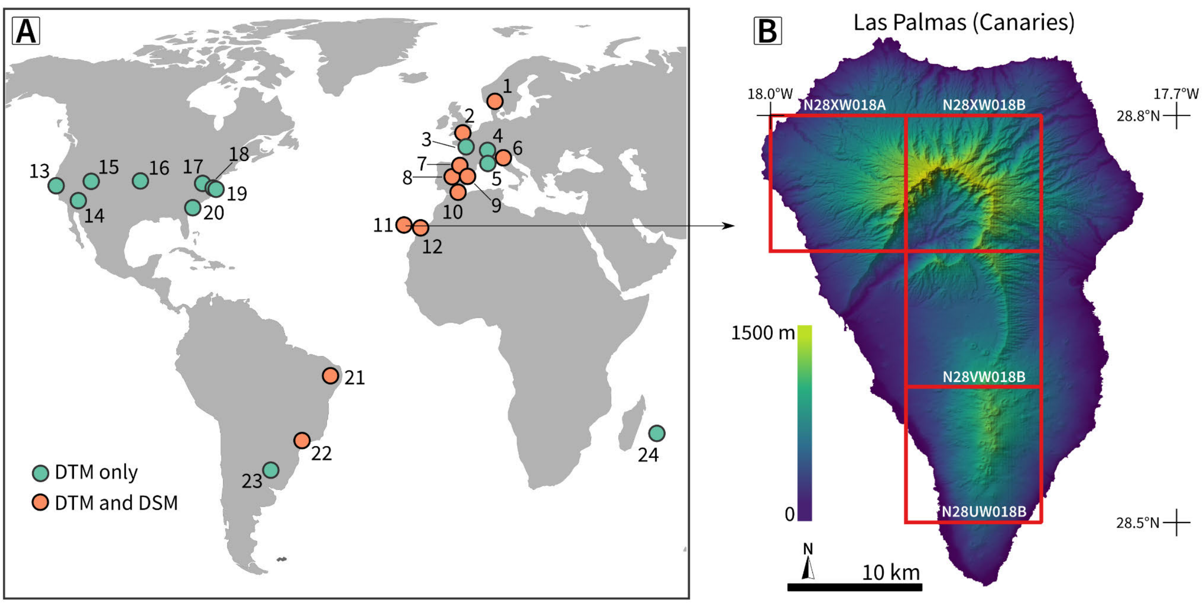
For each test area where reference DEMs were available, one or more DEMIX sampling tiles were extracted. The reference data for the intercomparison of global DEMs has significantly higher accuracy elevation values and usually much smaller sampling spacing than the DEMs to be tested. These reference DEMs were produced by a variety of mapping agencies from around the world with sampling intervals between 1-5 m which is more than five times finer spatial resolution than the global DEMs considered. While DTMs are available for all 24 test areas and 236 tiles, DSMs are only available in 11 of those areas (134 tiles) because some mapping agencies do not provide them. Of the global DEMs under consideration, only FABDEM claims to be a DTM. The rest of the candidate global DEMs are closer to DSMs but most likely fall somewhere between a true DTM and DSM (Guth et al. 2021c). Due to this ambiguity, all the global DEMs considered were compared against both, reference DTM and DSM when available.

Figure 10. The general DEMIX wine contest steps as executed for the comparison of global DEMs.



Source: Bielski et al. (2024)

Figure 11. A) Location of the 24 test areas. B) Distribution of DEMIX tiles over Las Palmas Island. The names of the test areas shown on the map A: 01 - Norway, 02 - Oxford, 03 - Caen, 04 - Valonne, 05 - Vanoise, 06 - Trentino, 74 - Pyrenees, 08 - Madrid, 09 - Ebro Delta, 10 - Almeria, 11 - Las Palmas, 12 – Canary East, 13 - Redwoods, 14 - State Line, 15 - Canyon Range, 16 - Republican River, 17 - Shenandoah, 18 - Blackwater, 19 - Chincoteague, 20 - Charleston, 21 - Pernambuco, 22 - São Paulo, 23 - Uruguay, 24 - La Reunion.



Source: Bielski et al. (2024)

For a comparison of global DEMs, the reference DEMs must cover a representative sample of the different landforms observed on Earth. The quality of the global DEMs being compared are not expected to depend on national or continental borders because the same acquisition technique and processing was used to create each of the DEMs under consideration. Therefore, it was necessary to be sure that a representative sample was available for the wine contest and covered landforms available on all continents. As a first approximation this was achieved by calculating statistics and classifying landforms found within sample tiles. The results show that a wide range of surface characteristics are covered by the set of tiles used in the wine contest. Our tile selection includes a large range of different landscapes and provides a much better diversity of surface characteristics than previous efforts comparing global DEMs (see Section 4.6 and **Figure 13**).

4.2. Preparation of the Reference DEMs

The preparation of the reference DEMs was performed in MICRODEM version 2023.6.20 (Guth, 2009, Guth, 2023i). The following key global DEM characteristics were managed to produce the reference DEMs:

- **Match Grid Type** - the global DEMs under consideration have different grid representations which affect their spatial location: pixel-is-point and pixel-is-area (Guth et al. 2021c). Only one quarter of the area is shared by pixels defined by the two representations, and although they have the same nominal coordinates in uneven terrain, they are unlikely to have the same elevation.
- **Spatial Sampling** - a 1'' DEM cannot be compared directly with a reference DEM having a spatial resolution of 1-5m because of the significant differences in spatial scale. The global DEMs have a spatial sampling of approximately 30m (1'') at the equator and their sensing techniques can be assumed to deliver an average elevation in each given cell. Therefore, the reference DEMs must be aggregated in order to match the spatial sampling of the global DEM.
- **Aggregation** - global DEMs are provided in geographic coordinates whereas reference DEMs typically use local projections such as UTM to avoid spatial distortions. The projected coordinates of each pixel in the source reference DEMs were converted to WGS84 latitude and longitude, and mapped to the corresponding pixel in the test DEM. The value for the reference 1'' DEM is the mean of the points in the high resolution source DEM, ranging from about 36 points for a 5m DEM to about 900 for a 1m DEM.
- **Latitudinal Effects** - the pixel spacing changes with latitude for CopDEM and ALOS but remains constant for the others.
- **Vertical Datum** - We moved the reference DEMs to the EGM2008 vertical datum using bilinear interpolation and the highest resolution available transformation grids from the PROJ library (https://proj.org/) or using GDAL to provide combined vertical and horizontal datum shifts for the United States DEMs. In a few cases we used a local transformation geoid grid, not included in the PROJ library and which does not have an EPSG code.

4.3. Criteria Used for DEM Comparison

For this report, only pixel-by-pixel comparison was applied for the assessment of global DEMs, using a reference DEM with higher accuracy elevation values. The quantitative assessments carried out on the global DEMs were based on pixel-by-pixel differences of three geomorphometric parameters: elevation, slope, and a roughness index. The following parameters were computed for the 6 candidate DEMs for all DEMIX tiles with respect to the reference DTM and reference DSM where available: elevation differences (ELVD); slope differences (SLPD) using the (Evans 1980) algorithm and roughness differences (RUFD) from the standard deviation of slope in a 5x5 window (Grohmann et al. 2011). It should be noted that both slope as well as the roughness indices considered in reality provide information on a neighbourhood (respectively 3x3 and 5x5 pixels). Accordingly, when computing differences in slope and roughness we are comparing information related to these neighbourhoods and not just to the pixel to which these values are associated.

The differences between the reference DEM and global DEM were always computed in the same manner:

$$difference = globalDEM - referenceDEM$$

where positive values indicate that the global DEM has a higher value than the reference DEM at a specific pixel.

In order to achieve a quantitative assessment required for the wine contest based on the above parameters, it was necessary to compute derived statistics from the distribution of the differences for each of the 236 DEMIX tiles:

- Standard deviation (STD);
- Average deviation (AVD) -- the mean difference was subtracted from each difference, and the average of its absolute value computed (Press et al., 2007);
- Root mean square error (RMSE);
- Mean absolute error (MAE);
- Linear error 90 (LE90) -- the 90th percentile of the distribution of the absolute differences.

In addition to computing statistics for every pixel in the tile, we computed separate pixel-by-pixel statistics for different slope and land cover classes. We used three land cover classes (forest, urban, and barren) from the Copernicus Global Land Cover Layers - Collection 2 (Buchhorn et al., 2020), and four slope categories cliff (slope > 50%), steep (slope > 25% and < 50%), gentle (slope < 25% and > 12.5%), and flat (slope < 12.5%).

4.4. Statistical Methods

Ranking a collection of DEMs from a given set of candidates leads to a mathematically similar problem to comparing wines. We use this analogy in order to develop a novel, extensible and user-tuneable comparison of global DEMs. For every combination of DEMIX tile and criterion there will be one opinion, formally similar to the one produced by a judge (tile and criterion) analyzing wines (DEMs). We are aware that for wines most criteria are subjective and require considerations over many sensory aspects. However, objective opinions can be addressed as well and are formally similar with the chosen. We cannot discern from the numerical result if an opinion based upon wine acidity was decided with the tongue or with readings over a pH instrument. Both might be valid inputs for the rest of the process. We wish to underline that the method can also be applied to comparisons of other geospatial products and is in no way limited to ranking of global DEMs and only requires that the products be ranked according to defined criteria.

The proposed DEMIX wine contest provides a statistical procedure to use the results from quantitative and/or qualitative criteria to rank DEMs and provide statistical confidence to the final ranking. The adopted statistical procedure is formally known as the randomized complete block design (RCBD – Addelman 1969). We can consider `wine contest' as a synonym to RCBD and apply it in the context of an intercomparison of global 1'' DEMs to identify the most appropriate DEM for a given set of criteria and sample sites, while providing statistical confidence to the outcomes.

Based on this design (RCBD) two tables will be built. The first one will be denoted here as the evaluation table (**Figure 12** , left side). It has k columns (one per global DEM) and N rows (one for each criterion applied to each site) holding the numerical results from evaluating the criterion for a site and for a DEM. For example, one row might hold the elevation RMSE values for a particular site. A ranking for each row of the evaluations table is produced and stored as the corresponding row of the opinions table (**Figure 12**, right side). Each row of the opinion table holds a ranking, which (in a case without ties), is a permutation of the integer values from 1 to k. The opinion table can be interpreted as the starting point of the RCBD, so no requirements are posed over how it was produced. Thus, the evaluations can be either qualitative (i.e., subjective based on the opinion of an expert) or quantitative based on some objective method that can produce a numerical result amenable to be sorted (like the RMSE). If a criterion cannot produce a ranking it cannot be part of the wine contest. The opinion table is built with the rule of “lower is better”. In the RCBD context, the following associations can be made with respect to the wine contest:

- The wines are presented to the judge(s) in an arbitrary order and therefore the test is Randomized. This is important only for subjective criteria.
- Each judge must produce a valid opinion for every wine based on their individual taste, so the test is Complete. There are no voids in the opinions table.

There is no reason to expect the judges (block) to produce identical opinions for the same set of wines.

Figure 12. A realistic wine contest example applied to the intercomparison of DEMs - the evaluations table (left side) records the assessment outcomes for each of the candidate DEMs. The opinions table (right side) translates the evaluations table to ranked opinions.

	Evaluations							Opinions						
Criterion	FABDEM	COPDEM	ALOS	NASADEM	SRTM	ASTER	tolerance	FABDEM	COPDEM	ALOS	NASADEM	SRTM	ASTER	
ELVD_RMSE	2.46	2.44	4.42	5.36	5.78	9.00	0.5	1.5	1.5	3	4.5	4.5	6	
ELVD_LE90	2.38	2.29	6.04	8.36	8.67	14.20	0.5	1.5	1.5	3	4.5	4.5	6	
SLPD_RMSE	3.52	3.47	3.79	9.22	9.04	12.48	0.5	2	2	2	4.5	4.5	6	
SLPD_MAE	1.28	1.19	2.20	5.83	5.70	8.26	0.5	1.5	1.5	3	4.5	4.5	6	
RUFD_AVD	0.51	0.52	0.63	2.28	2.22	2.61	0.5	2	2	2	4.5	4.5	6	
RUFD_RMSE	1.24	1.25	1.03	3.15	3.09	3.70	0.5	2.5	2.5	1	4.5	4.5	6	
							Ri	11.0	11.0	14.0	27.0	27.0	36.0	
							Rj^2	121.0	121.0	196.0	729.0	729.0	1296.0	sum=3192.0

Source: Carlos López-Vázquez

The null hypothesis is that there is no difference among the DEMs, and a consensus based on the opinions cannot be achieved. However, if the null hypothesis is rejected, then the contest ranking is not based on chance (given a chosen confidence level) and some conclusions can be obtained.

Ties, individual results which are considered equally good (or equally bad), can be considered by applying the mid-rank procedure (Amerise and Tarsitano 2015) and arise after considering tolerances. When applicable, tolerances should account for measurement uncertainty to identify minor differences in the evaluations table that will have an impact in the opinions table but are not really different. For example, results in the evaluations table are recorded to two decimal places, but many of these DEMs only record elevation to the nearest meter and thus differences of centimeters or even decimeters would not imply a significant difference among DEMs. Tolerances can also reflect chosen reference data comparison algorithms, and thematic requirements when justified to support the creation of the opinions table.

With the opinions table in-hand, the overall opinion ranking can be computed where R_j is the sum of the ranked value for a given DEM. Lower is better, since the rank order is from 1=best to *k*=worst. Sorting the R_j values might produce the requested wine contest ranking result. However, there is yet no statistical confidence associated with these rankings, so the next step is to apply the appropriate statistical tests.

A first step is to analyse if the differences between Rj values are just due to chance. The null hypothesis is that the DEMs are indeed equivalent. Its analysis can be accomplished using either parametric or non-parametric tests.

One of the strengths of the wine contest is that both quantitative and qualitative evaluations can be integrated into the contest. Therefore, a non-parametric test should be used which can operate over both types of rank results. The chosen non-parametric test for the wine contest is the Friedman Test (Friedman 1937), because among the non-parametric choices at hand, it is the best alternative for minimizing the risk of paradoxical results (Fey and Clarke 2012).

The general Friedman's statistic χ^2_r, is presented in **Equation 1**, valid with or without ties (N=number of opinions/assessments, k=number of DEMs).

Equation 1. The general Friedman statistic.

$$\chi_r^2 = \frac{N(k-1)\left[\sum_{j=1}^{k} \frac{R_j^2}{N} - C_F\right]}{\sum_{i=1}^{N}\sum_{j=1}^{k} r_{ij}^2 - C_F}; C_F = \frac{Nk(k+1)^2}{4}$$

To apply the Friedman Test, the entries in the opinions table (**Figure 12**) are denoted as the elements r_{ij}, and the final row presents the column sums by R_j. The remaining values to compute the Friedman statistic (χ^2_r) can be extracted from the same table: with CF=441, sum of r_{ij}^2=537.0, and sum of R_j^2=3192.0, then χ^2_r =28.871.

For a given k, N and confidence level alpha, the Friedman's statistic value χ_r^2 is compared to the critical value; if χ_r^2 is larger, the null hypothesis is rejected, and the conclusion is that we cannot accept that the DEMs are equivalent. Since we are willing to accept ties, the standard critical values tables from the Friedman Test are not suitable. From the table for k=6 provided by (López-Vázquez et al. 2021, López-Vázquez and Hochsztain 2020), and at the 95% confidence level, the row N=6 offers a critical value to compare of χ^2_{crit}=10.489. Compared to the χ^2_r result of 28.871 which is larger, one should reject the null hypothesis that the opinions table entries are purely at random, implying that the DEMs are not equivalent.

The null hypothesis has been rejected based on the Friedman Test, so there are statistically significant differences among the set of DEMs under consideration. The lower the values of R_j, the better, but it is still necessary to assess whether the difference between the pairs of ranked candidate DEMs are statistically significant or otherwise conclude that the pair is tied. The process is denoted as post-hoc analysis and there exist different options to carry out such an analysis. In this case and following (Pereira et al. 2014), we propose to use the test by (Dunn 1961) applying the Bonferroni correction (Equation 2).

A pair of DEMs is considered significantly different if

Equation 2. Applying the Bonferroni correction.

$$|R_i - R_j| \geqslant z_{1-\alpha/k/(k-1)}\sqrt{\frac{Nk(k+1)}{6}}$$

where *z* stands for the inverse of the cumulative normal distribution, and α is the prescribed confidence level.

4.5. Building the GIS Database

MICRODEM was used to produce a wine contest GIS database (Guth, 2023b), combining the tile characteristics, an evaluations table, and an opinions table with an initial set of tolerances. The opinions table ranks the six DEMs from 1 to 6 (low score is best), with the possibility of ties. We assign a tolerance for each criterion; evaluations within the tolerance will be treated as a tie. Many of the candidate global DEMs are rounded to integer type, therefore it does not make sense to treat differences on the order of centimetres or even decimetres as significant for any elevation-related metric.

The current DEMIX database includes 55,699 opinions which are rows found within the database. The following fields are included to help reading the database:

- Tile name, area, and centroid coordinates;
- Tile characteristics required for per tile filters;
- Whether the reference DEM is a DTM or a DSM;
- The land type, with ALL for every pixel, or one of the seven categories defined above if there are at least 100 pixels of that type within the tile, and the percentage of the tile in the land type;
- Criterion (quantitative assessment metric) name for the 15 metrics applied in this wine contest (plus six other signed metrics for analysis purposes not applicable to the wine contest) defined above;
- Evaluations table using all six candidate DEMs;
- Tolerance used to populate the opinions table and identify ties;

- Opinions table for each computed metric including the identification of potential ties due to the pre-defined tolerances.

Any combination of opinions can be used in the final wine contest ranking. The database can be filtered to show only opinions for a single land type, like `steep' or `barren', or a user may only be interested in the tiles where the forest cover type covers at least 50%-75% of the tile, or where the relief is greater than 1000m. This would filter the evaluations table to only include tiles that fulfill the desired requirement.

4.6. DEMIX Wine Contest Results

Figure 13 summarizes the results of wine contests applied to subsets of the database, integrating almost forty thousand opinions in the DEMIX database. The top rows show the overall ranking using all tiles and criteria based on the DTM and DSM reference DEMs. For this set of evaluations (ALL land type), the DTM winner is FABDEM and the DSM is CopDEM, both at the 95% confidence level.

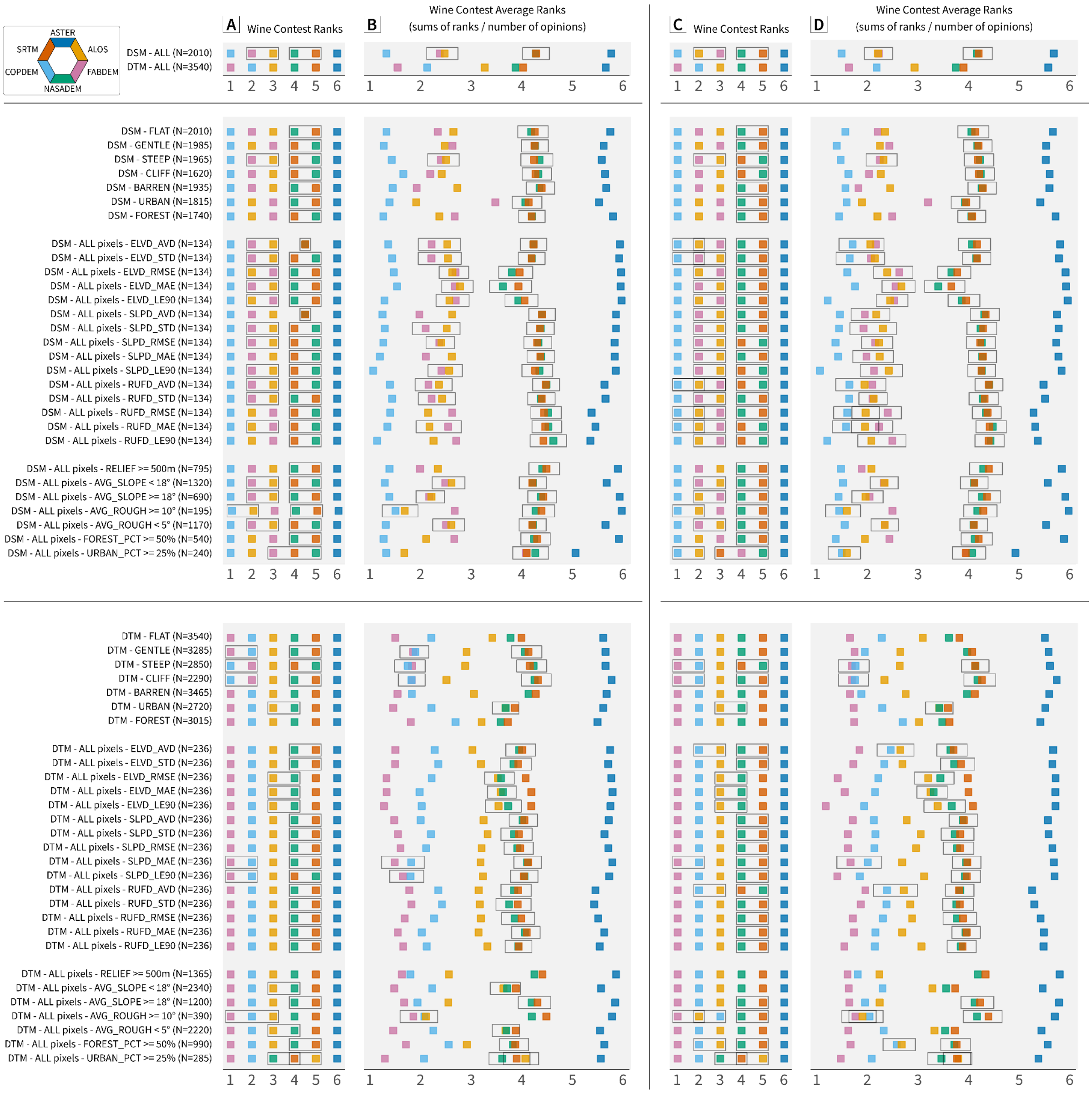
The DEMIX wine contest also associates statistical significance to the ranking and this is presented in **Figure 13** by drawing a box around those global DEMs whose rank cannot be differentiated from a random result, i.e. the pairwise rank outcome does not pass the significance test.

The subsequent rows have been grouped by DTM or DSM, land type filter (FLAT, GENTLE, FOREST, etc), and criterion (ELVD_RMSE, SLPD_LE90, RUFD_MAE, etc). By filtering the database, not only the order can change but also the number of ties and between which candidate DEMs the ties have occurred. This illustrates the power of the wine contest procedure in the context of the intercomparison of global DEMs because depending on the user requirements, the 'best' DEM will emerge based on the set of chosen criteria, far from a situation of one-option-fits-all.

To try and visualize the final ranking outcomes with a bit of context, **Figure 13** , columns B/D presents the same outcomes with the goal of showing how many times a DEM ranks higher/lower over the number of opinions. The output is the sum of the ranks for any particular DEM for each row in the opinions table divided by the number of opinions. The more times a DEM is ranked higher in the opinions table, the higher the ranking. In this case, lower is better.

To demonstrate the effects that changing tolerances can have on the final rankings, **Figure 13** B/D presents the outcomes based on higher tolerances: 0.5 → 1.0 and 0.2 → 0.4. The increasing of the tolerances has had the effect of producing more ties between the different global DEMs, but the overall results are robust. This fact is an important result for the intercomparison of global DEMs.

Figure 13. The final rankings of the DEMIX wine contest, based on the intercomparison of six 1'' global DEMs. The different outcomes are based on the chosen land types, the 15 contest criteria and filters for tiles with specific characteristics. Tolerances for A/B: ELVD=0.5, SLPD=0.5, RUFD=0.2. Tolerances for C/D: ELVD=1.0, SLPD=1.0, RUFD=0.4; the increase in tolerances led to a greater number of ties in C/D compared to the rankings shown in A/B.

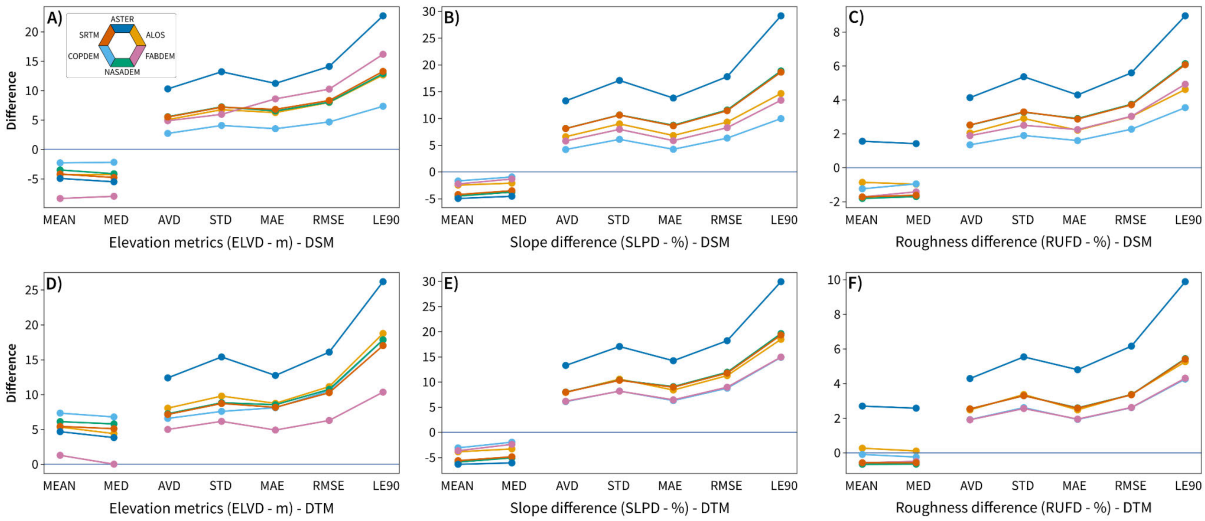


Source: Bielski et al. (2024)

Figure 14 displays seven criteria based on different metrics applied to elevation (**Figure 14** A/D), slope (**Figure 14** B/E) and roughness (**Figure 14** C/F) differences between the reference and test DEMs for tile N43PW002B in Spain. The first two results displayed in each case (left side of each graph) are the mean and the median differences. These signed results cannot be ranked and therefore are not applicable to the wine contest directly; however, they provide valuable summary statistics for the tile under consideration. Specifically, they show how candidate DEMs compare to the reference DEM in terms of the direction of the differences: positive or negative.

For the signed results, one observes some negative bias for results based on the elevations (when compared to a reference DSM), slope and roughness parameters. ASTER results tend to be outside of the group of results being an outlier in terms of not corresponding with the reference DEM compared to the other global DEMs. The five unsigned metrics used as criteria as part of the wine contest show similar trends across the chosen tile and within each geomorphometric parameter. It would however be difficult to choose the best global candidate DEM based only on these graphs, especially with any statistical confidence.

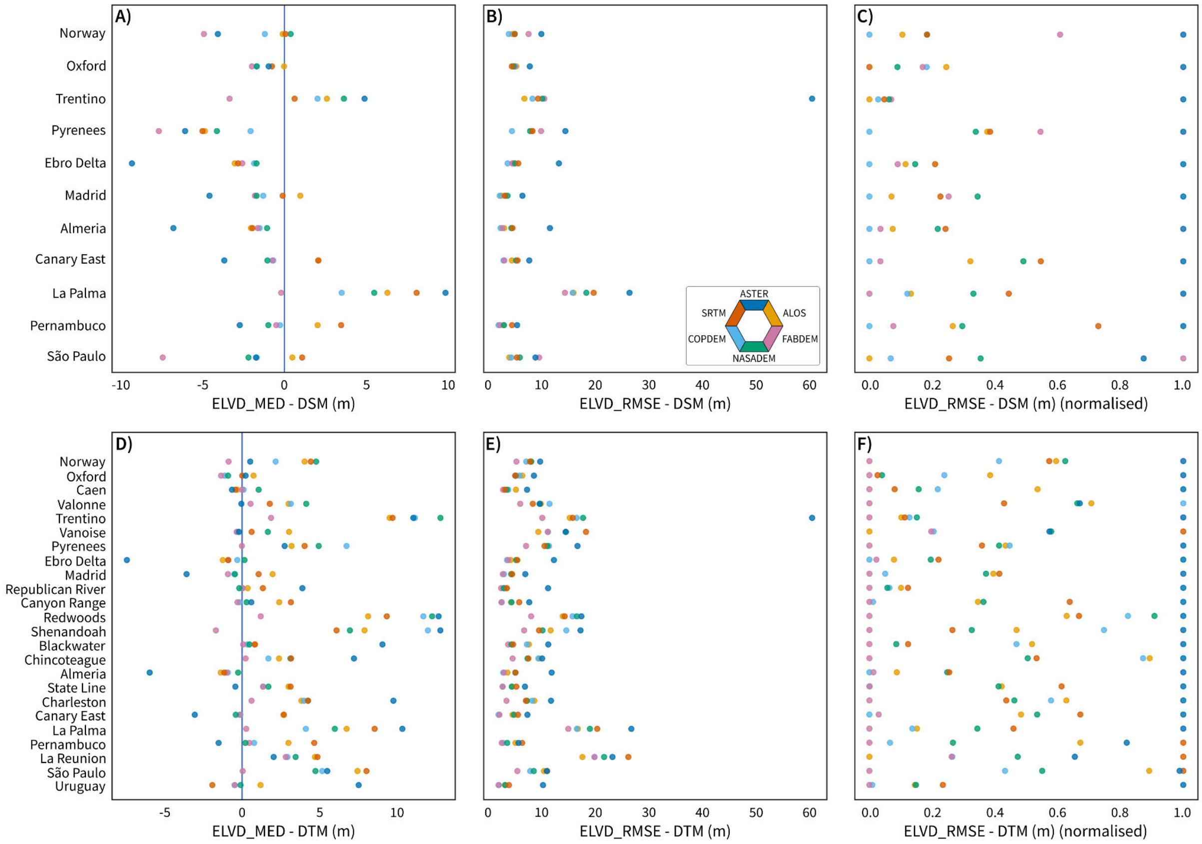
Figure 14. The distribution of seven elevation, slope and roughness metrics comparing the difference between the test and reference DSMs (top) and DTMs (bottom) for tile N43PW002B in Spain. The mean and median on the left of each graph are signed; only the five rightmost parameters are used in the wine contest. The lines connecting the points, while not statistically valid, show the effect on the parameters by their weighting of the tail of the distribution.



Source: Bielski et al. (2024)

Figure 15 A and D show the medians of the elevation difference distributions for the 24 test areas, which average the results from 1-35 tiles in each area, highlighting a general negative bias for results based on the elevations when compared to a reference DSM and a positive bias compared to the DTM.

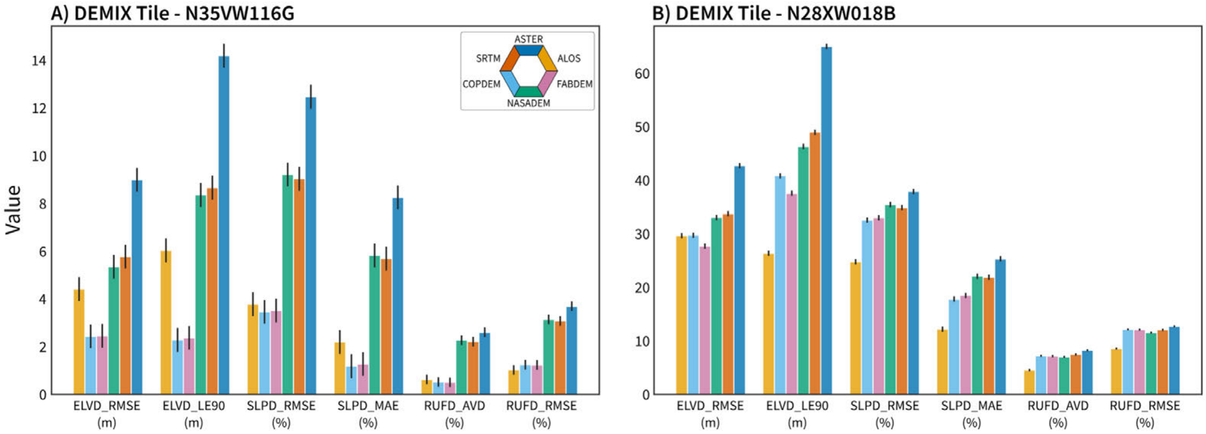
Figure 15. Distribution of two signed metrics for all areas. A/D: ELVD_MED. B/E: ELVD_RMSE. C/F: ELVD_RMSE with normalised values. Negative values for the median indicate the DEM is lower in elevation than the reference DEM, while positive values indicate the DEM has higher elevations than the reference DEM. The areas are arranged by latitude from north to south.



Source: Bielski et al. (2024)

Figure 14 and **Figure 16** provide some guidance on the selection of criteria, and the tolerances for accepting ties. In **Figure 14** , the criteria used for the wine contest on the right are separated from the signed means and medians. The five unsigned criteria are arranged from left to right in rough order by the effect of outliers; LE90 always has a much larger value compared to the others. With rare exceptions the lines connecting the values do not cross, and the rankings for the criteria are the same. The tolerances used for ties apply to the difference scale on the vertical axis; the amount of "white" space between the lines shows the tolerance that would be required to consider the DEMs tied. In most cases visual assessment of graphs like this reinforces our conclusions about the rank ordering of the DEMs: CopDEM and FABDEM, then ALOS, then NASADEM and SRTM, and finally ASTER. **Figure 16** shows the evaluation results for 6 criteria and two different tiles. Note the very different vertical scales; tile N28XW018B in the Canaries is very steep and rough. Scores for the different criteria are also very different, although this also depends on the local topography. Salient results from this graph are that most of our results are immune to any reasonable choice of tolerances or which criteria are used, and that our ranking of the DEMs is robust. Tile N28XW018B is anomalous in showing ALOS as the best DEM for 5 of the 6 criteria.

Figure 16. Bar plots of the evaluation results for 6 criteria with two tiles compared to a reference DTM. The small black line atop each bar shows the tolerance we used for that criterion; and show the evaluation another candidate DEM would need to be tied.



Source: Bielski et al. (2024)

4.7. Mining the DEMIX Wine Contest Database

Section contributor: Peter Guth

4.7.1. Exploring Diversity Among the DEMIX Contest Tiles

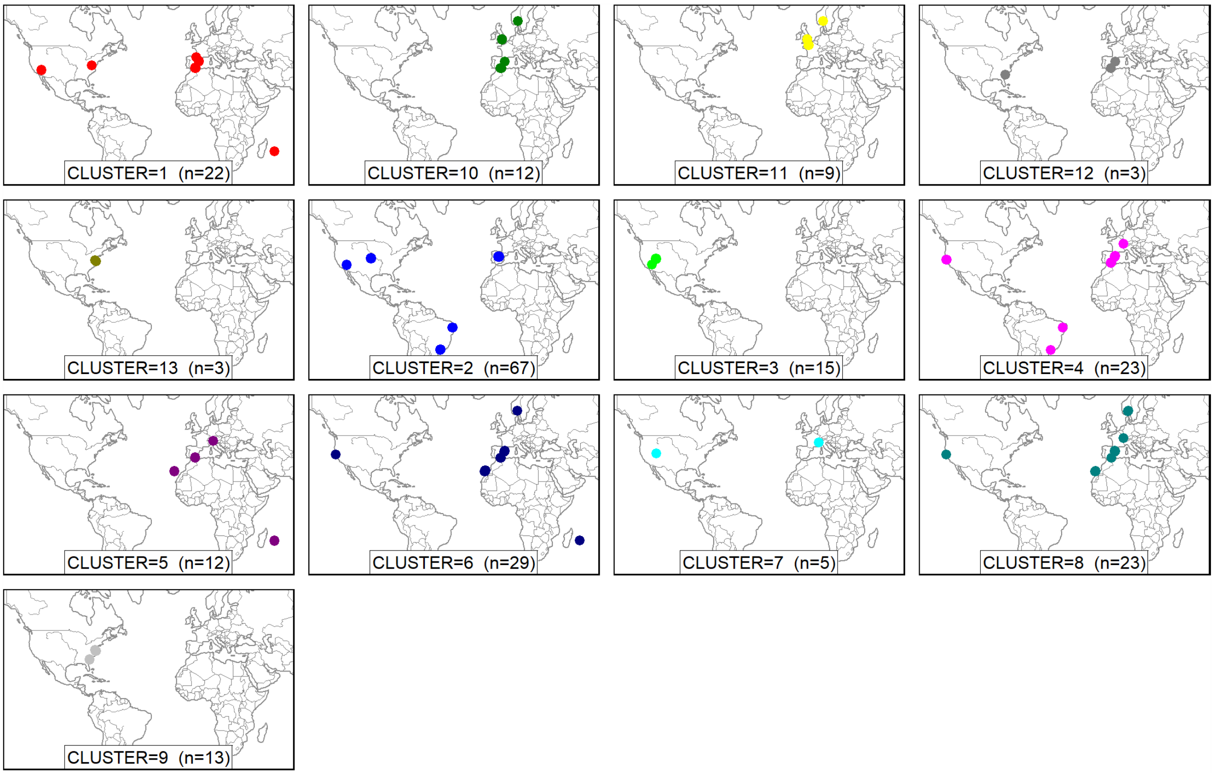
The DEMIX database (Guth, 2023b) contains 7 characteristics for each of the 236 tiles: average elevation, average slope, average roughness, relief, and the percent of the tile classified as urban, forest, and barren. **Figure 17** and **Figure 18** how the DEM rankings can be sorted by the single characteristics such as average slope.

A more complex analysis can look at all 7 characteristics together, using K-means clustering. The algorithm seeks to find 15 clusters but might return fewer and in this case could only differentiate 13 clusters. **Figure 17** shows the clusters, along with the number of tiles in each cluster. At this map scale multiple tiles from the same test area cannot be differentiated. **Figure 18** shows scatter plots of the characteristics, showing which characteristics differentiate the clusters. **Table 2** shows the mean value for all 7 characteristics in each of the 13 clusters from the K-means clustering. As an unsupervised method, K-means clustering only groups similar tiles without any geomorphic interpretation. Looking at both figures and the table allows interpreting them. Clusters 9, 12, and 13 contain flat coastal tiles, with cluster 13 having generally lower relief and cluster 9 having a large fraction of the tile forested.

K-means clustering algorithms have a number of tuneable parameters for the algorithm, the possibility for transforming the data before running the analysis, and the choice of variables to include in the clustering. We accepted the default settings in the MICRODEM implementation, which uses code from Fred Edberg written 20 years ago (no valid current links/references could be found), did not transform the data, and used all 7 characteristics.

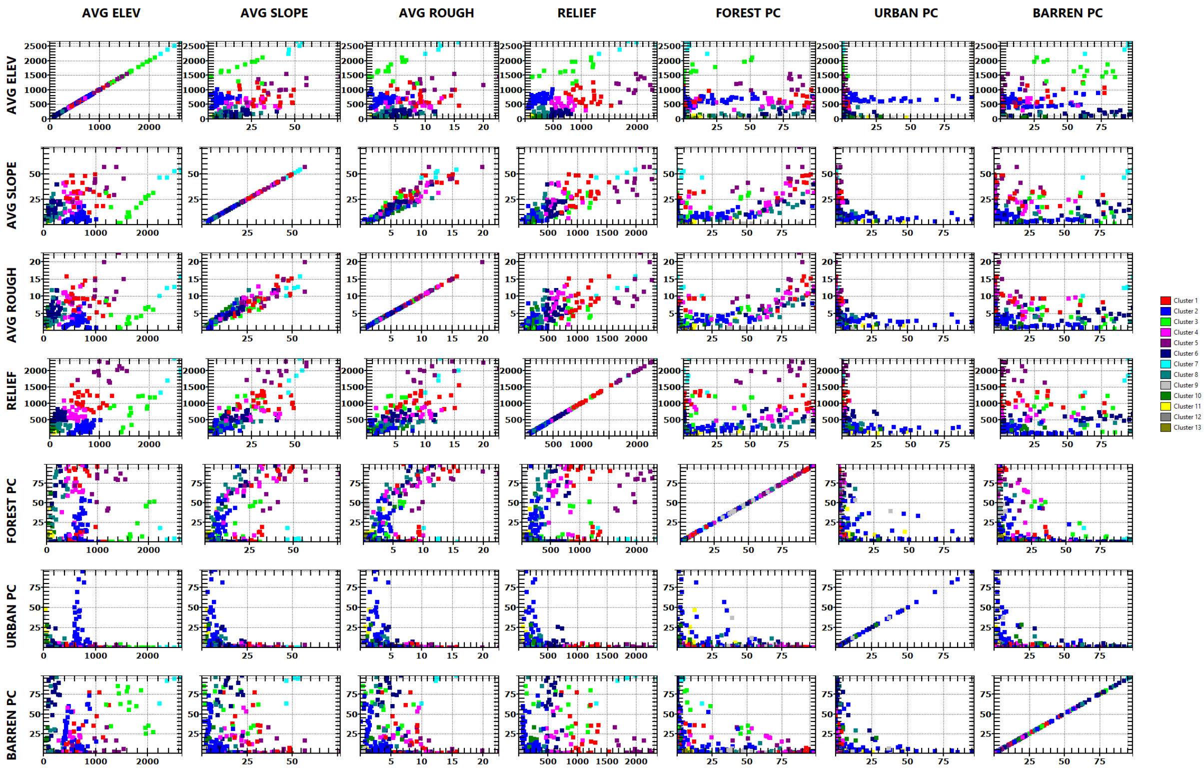
Users who wanted to test whether the current database represents the landscape diversity in their region of interest could compute the 7 characteristics for their tiles, add the values to the database, and then rerun the clustering. If the tiles did not belong to a cluster, or the cluster was small, additional tiles with similar characteristics could be acquired to expand the database. Users can add additional characteristics to the DEMIX database (Guth, 2023b) by computing them for each of the test tiles. Users can remove characteristics from the clustering. The Meybeck et al. (2001) typology for mountains has 4 categories based on elevation, and the arbitrary boundaries might not be appropriate for tiles near the category boundaries that contain two categories that are probably not really that different.

Figure 17. Tile clusters from the 236 tiles in the DEMIX database using all 7 characteristics (Note – Cluster 9 is presented with grey dots on the US East Coast).



Source: Peter L. Guth

Figure 18. Scatterplots from the 236 tiles in the DEMIX database using all 7 characteristics, with the same colours as **Figure 17**.



Source: Peter L. Guth

Table 2. Mean value for the 7 tile characteristics for each of the 9 clusters.

CLUSTER	N	AVG_ELEV	AVG_SLOPE	AVG_ROUGH	RELIEF	FOREST_PC	URBAN_PC	BARREN_PC
1	22	759	32.52	9.17	1073.75	43.42	0.69	21.57
2	67	618.08	6.43	2.41	199.95	14.43	15.71	17.81
3	15	1671.54	16.21	3.97	756.03	16.52	0.36	57.33
4	23	525.61	23.6	7.5	666.61	51.95	1.33	15.02
5	12	1145.18	43.32	12.23	1953.14	70.92	0.9	8.98
6	29	196.55	17.65	5.6	593.61	26.67	3.49	42.62
7	5	2481.68	50.31	12.6	1839.82	5.73	0.29	88.81
8	23	207.49	12.62	4.65	341.88	40.84	3.28	17.64
9	13	2.77	0.74	0.45	16.48	49.06	4.68	2.37
10	12	68.77	6.38	2.83	211.19	14.86	7.18	10.22
11	9	62.55	3.36	1.3	112.79	13.3	15.01	3.28
12	3	1.61	0.52	0.35	22.21	0.17	3.1	1.98
13	3	-0.5	0.3	0.23	8.71	6.97	1.34	2.57

Source: Peter L. Guth

Section 3.6 shows 4 classifications of the test areas with the percentage of the tile in each category; we have similar figures for the 236 tiles. The variety of categories in each area shows the difficulty of assessing the dominant classification in each tile, and the limited number of locations where they would be useful. As part of creating the figures, we also created databases with the percentage of each tile belonging to each category. **Table 3** summarizes the number of categories in each classification scheme, as well as the number of clusters found by the K-means algorithm. Two of the classifications only differentiate three clusters, and those might be less useful.

Filtering the GIS database allows selection of tiles that match desired characteristics, and test whether the results of the wine contest differ in that particular landscape. We are confident that our current selection of tiles faithfully captures a representative sample of the world's landscapes, and that our conclusions that CopDEM and FABDEM are the best choices, will be generally true almost everywhere.

Table 3. Tile characteristics for the five tile classification databases.

Classification Database	**Classification Categories**	**Clusters**
DEMIX	7	13
ESACCI-LC-L4-LCCS	38	7
Geomorphons	10	3
Iwahashi & Pike	16	10
Meybeck & others	15	6

Source: Peter L. Guth

5. Using MICRODEM for DEMIX

Section contributor: Peter L. Guth

5.1. Introduction

DEMIX wanted open-source routines and software to perform the Wine Contest. As such it should be easy to translate to other environments like C, C++, Matlab, Python, R, or Rust. MICRODEM supplied a local machine platform for development of the methodology, creation of the database, initial graphical display of the results, and further exploration of the database.

5.2. MICRODEM Software

MICRODEM is a GUI program written in Delphi for the Windows 64 operating system. While Delphi can target some other operating systems and there has been some experimentation with both Android and Mac platforms for MICRODEM, porting the DEMIX code to other platforms would require significant effort.

MICRODEM has been under development for several decades (Guth et al. 1987) and more recently has focused on geomorphometry (Guth 2009). Because one of its first primary data sets was DTED, MICRODEM has always considered native usage of geographic coordinates for computation and map display.

MICRODEM is open source. The Delphi source code, approximately 400,000 lines, is on github (Guth, 2023h). The most recent compiled executable and the help file in CHM format are at microdem.org (Guth, 2023i).

While Delphi is a commercial product, MICRODEM can be compiled with its community edition (Embarcadero, 2023) which is free for limited non-commercial usage (consult license for legal details). As a teaching language, Delphi reads a lot like pseudo code which makes it easy to translate into other languages.

5.3. DEMIX Processing Step Overview in MICRODEM

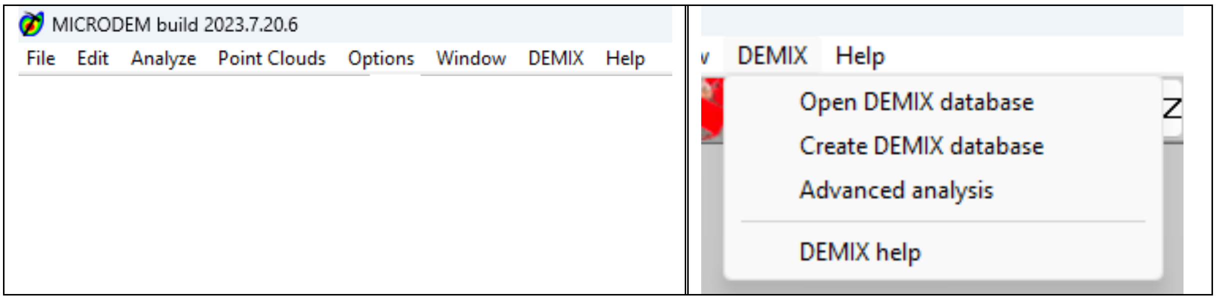
The complete processing chain involves the following steps:

- MICRODEM Setup, including GDAL
- Collect landcover data for the test areas
- Collect Reference data and down sample to 1" spacing by aggregation
- Collect and index test/candidate global DEMs
- Evaluate DEMIX wine contest criteria
- Create DEMIX database in DBF format (transpose and merge criteria CSV files)
- Manipulate database for graphics and statistics
- Export database in CVS format

5.4. Accessing the DEMIX Options in MICRODEM

The main program menu (**Figure 19** – left side) has a submenu (**Figure 19** – right side) that opens all the required options.

Figure 19. Finding the DEMIX options in the MICRODEM Menu: left side – all menu items, right side – DEMIX options.



Source: Peter L. Guth

The four major options:

- Open DEMIX database opens the database, which must have already been created, and allows a large number of options from a control form.
- Create DEMIX database: runs through the required steps to manage the reference data and create the database.
- Advanced analysis: experimental options not yet accessible from the Open DEMIX database.
- DEMIX help: opens the MICRODEM help file at the main DEMIX help page. This has many hyperlinks, should be the primary source for running the wine contest.

5.5. Using MICRODEM for DEMIX

Manipulation of the DEMIX database continues as active research, and the program will continue to evolve. The source code is available (Guth, 2023h), so you can resolve questions and modify it.

Adding more test areas is straightforward and just requires following the steps in the processing chain.

Adding additional criteria will require modifying the source code.

Adding additional candidate DEMs will require modifying the source code.

6. The DEMIX Wine Contest Jupyter Notebook

Section contributor: Carlos Henrique Grohmann

6.1. Introduction

Jupyter Notebooks have become the most widely used system for shareable and reproducible research (Shen 2014, Rule et al. 2019, Kluyver et al. 2016). It supports a variety of programming languages, such as Julia, R, JavaScript, and C, allowing the user to create literate programming (Knuth 1984) documents combining code, text, and results with visualizations and other rich media (Pimentel et al. 2021).

Within the DEMIX Working Group, a first Jupyter Notebook was developed for the analysis and visualization of the DEMIX Wine Contest. This article briefly describes its characteristics and usage.

In summary, the contest is designed to rank a set of Global DEMs based on a number of objective criteria (although subjective criteria could be used as well). Each criterion (such as RMSE) is used to rank the Global DEMs against a reference DEM. Then, a set of criteria is defined and the Friedman Test (Friedman 1937) is used to determine if the DEMs can or cannot be considered statistically different. If there are statistically significant differences among the set of DEMs, they are compared pairwise using the test by Dunn with Bonferroni correction (Dunn 1961) (a step called post-hoc analysis) and a final ranking is produced.

The notebook, as well as auxiliary files, is available in GitHub (https://github.com/CarlosGrohmann/DEMIX_wine_contest).

6.2. Characteristics

The DEMIX Jupyter Notebook was developed based on Python version 3.10.x and the Pandas, qgrid, Numpy, Matplotlib, and Seaborn libraries (Python 2021, qgrid 2020, Harris et al. 2020, McKinney 2011, Hunter 2007, Waskom 2021).

As the idea behind the notebook was to provide a simple interface to analyse the DEMIX wine contest data, the statistical and plotting functions were implemented in the demix_wine_functions.py file. The user can alter some plotting options, such as colours or symbols. One must note that the plotting functions were tailored to produce the plots and figures presented in (Bielski et al. 2024) and running the notebook with a different number of candidate DEMs will require adaptation of the existing code.

The notebook can be run on a user's local machine or in the cloud using the Google Collaboratory (Colab) platform. While running locally requires setting up a working python/jupyter environment, it allows for more flexibility in terms of files' location. Running it in Colab requires installing a specific version of qgrid and downloading external files from GitHub and Zenodo. All the instructions necessary for the user to run the notebook are included as comments within code cells or as rich text.

6.3. DEMIX Notebook Usage

As input the Jupyter notebook will read a .CSV file containing, at least, the criteria for ranking and their values for each test DEM (the evaluation table). Other information can be present in the database and used for filtering, such as the reference surface type (DSM/DTM) or landcover types.

In any case, this auxiliary information must be the initial set of columns of the database, followed by a column named "CRITERION" holding the acronyms for each criterion (as "ELVD_RMSE", "SLPD_STD" etc). The last columns must have the values for each criterion and test DEM. **Table 4** shows an example of the structure of the CSV file.

Table 4. Example structure of the CSV file used as input for the Wine Contest by the Jupyter Notebook.

(auxiliary columns)					*(mandatory columns)*				
AREA	**DEMIX_TILE**	**REF_TYPE**	**LANDCOVER**	**...**	**CRITERION**	**DEM1**	**DEM2**	**DEM3**	**DEM4**
Area_1	XXX	DSM	FOREST	...	ELVD_RMSE	0.0	0.0	0.0	0.0
Area_1	XXX	DTM	FOREST	...	ELVD_RMSE	0.0	0.0	0.0	0.0
Area_2	YYY	DTM	BARREN	...	SLPD_STD	0.0	0.0	0.0	0.0
...									

Source: Carlos Grohman

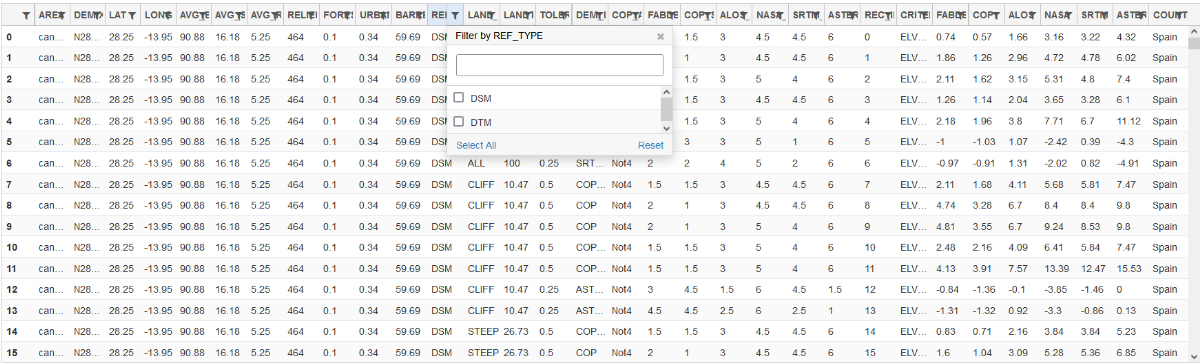
The database can contain signed (mean, median) and unsigned values (RMSE, LE90, MAE etc.) of the differences between the test DEMs and reference DEMs, but only the unsigned values are read for the Wine Contest. The signed values are used to produce plots that help the user to explore and understand the set of data being analysed.

Before exploring the database, the user can define the values for tolerances, which will impact the occurrence of ties between the DEMs in the rankings.

A Pandas dataframe will be created from the database, and the qgrid library will provide an interactive spreadsheet widget (**Figure 20**). This widget allows the user to query the database by various properties, such as selecting between DTM and DSM, filtering by DEMIX tile, area or difference metric.

It is imperative to select at least between DTM and DSM, as the analysis will not provide meaningful results if these two surface references are mixed.

Figure 20. Database rendered as a spreadsheet by qgrid. The selection between DTM and DSM is shown.



	AREA	DEM	LAT	LONG	AVGE	AVGS	AVGR	RELIE	FORE	URBA	BARR	REF	LAND	LAND	TOLE	DEM	COPY	FABD	COPY	ALOS	NASA	SRTM	ASTER	REC	CRITE	FABD	COPY	ALOS	NASA	SRTM	ASTER	COUNT
0	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69								1.5	3	4.5	4.5	6	0	ELV...	0.74	0.57	1.66	3.16	3.22	4.32	Spain
1	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69								1	3	4.5	4.5	6	1	ELV...	1.86	1.26	2.96	4.72	4.78	6.02	Spain
2	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69								1.5	3	5	4	6	2	ELV...	2.11	1.62	3.15	5.31	4.8	7.4	Spain
3	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69								1.5	3	4.5	4.5	6	3	ELV...	1.26	1.14	2.04	3.65	3.28	6.1	Spain
4	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69								1.5	3	5	4	6	4	ELV...	2.18	1.96	3.8	7.71	6.7	11.12	Spain
5	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69								3	3	5	1	6	5	ELV...	-1	-1.03	1.07	-2.42	0.39	-4.3	Spain
6	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	ALL	100	0.25	SRT...	Not4	2	2	4	5	2	6	6	ELV...	-0.97	-0.91	1.31	-2.02	0.82	-4.91	Spain
7	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.5	COP...	Not4	1.5	1.5	3	4.5	4.5	6	7	ELV...	2.11	1.68	4.11	5.68	5.81	7.47	Spain
8	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.5	COP	Not4	2	1	3	4.5	4.5	6	8	ELV...	4.74	3.28	6.7	8.4	8.4	9.8	Spain
9	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.5	COP	Not4	2	1	3	5	4	6	9	ELV...	4.81	3.55	6.7	9.24	8.53	9.8	Spain
10	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.5	COP...	Not4	1.5	1.5	3	5	4	6	10	ELV...	2.48	2.16	4.09	6.41	5.84	7.47	Spain
11	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.5	COP...	Not4	1.5	1.5	3	5	4	6	11	ELV...	4.13	3.91	7.57	13.39	12.47	15.53	Spain
12	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.25	AST...	Not4	3	4.5	1.5	6	4.5	1.5	12	ELV...	-0.84	-1.36	-0.1	-3.85	-1.46	0	Spain
13	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	CLIFF	10.47	0.25	AST...	Not4	4.5	4.5	2.5	6	2.5	1	13	ELV...	-1.31	-1.32	0.92	-3.3	-0.86	0.13	Spain
14	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	STEEP	26.73	0.5	COP...	Not4	1.5	1.5	3	4.5	4.5	6	14	ELV...	0.83	0.71	2.16	3.84	3.84	5.23	Spain
15	can...	N28...	28.25	-13.95	90.88	16.18	5.25	464	0.1	0.34	59.69	DSM	STEEP	26.73	0.5	COP	Not4	2	1	3	4.5	4.5	6	15	ELV...	1.6	1.04	3.09	5.28	5.36	6.85	Spain

Source: Carlos Grohman

A new dataframe is created based on the selection in the spreadsheet widget. This new dataframe will then be passed to the function responsible for the Friedman Test. This function will return information (**Figure 21**) about the tolerances, which filters were applied to the database and the initial result, stating if the DEMs have statistically significant differences among them, and if the user can proceed to the post-hoc analysis (**Figure 22**).

Figure 21. Example of results of the Friedman Test. Filters include a selection of DTM as surface type, CLIFF and STEEP as land types.

```
Ranking with user-defined tolerances (might take a while...)

Filter settings for column REF_TYPE:['DSM']
Filter settings for column LAND_TYPE:['CLIFF', 'STEEP']

Results of the DEMIX Wine Contest

For k=6, CL=0.05, and N=3585, the critical value to compare is chi_crit=11.038
And since chi_r (12041.957) is greater than chi_crit (11.038)...
Yay!! We can reject the null hypothesis and go to the Post-Hoc analysis!!
```

Source: Peter L. Guth

The next step is the post-hoc analysis, and the function will return a table (**Figure 22**), where each row corresponds to one DEM; the columns are:

- Rank – the final ranking, where lower is better;
- Sum of ranks – numerical value of the sum of all ranks for the DEM;
- Sum of ranks divided by number of opinions – this value might be used for comparing ranks made with different sets of criteria;
- Ties with – with which DEM there is a 'tie', that is, these DEMs don't have a statistically significant difference.

Figure 22. Example of the post-hoc analysis results, using the same filters as in **Figure 13**. Here SRTM and NASADEM are tied in 4th place.

	Rank	Sum of ranks	Sum of ranks divided by number of opinions	Ties with
COP	1.0	6081.5	1.696	
FABDEM	2.0	8728.0	2.435	
ALOS	3.0	9275.5	2.587	
SRTM	4.0	15419.5	4.301	NASA
NASA	5.0	15713.0	4.383	SRTM
ASTER	6.0	20067.5	5.598	

Source: Carlos Grohman

Included in the notebook are a plot of correlation matrices, a customized scatterplot (**Figure 14**) of 60 wine contest ranks (including database filters on surface type, land type, land cover, geomorphometric indices and two sets of tolerances), scatterplots of signed metrics (means, medians), unsigned metrics (RMSE, standard deviation), and of selected criteria per selected tiles (chosen as representative of the metrics' behavior). All figures can be saved in common formats (e.g., .png, .svg). Note that the plots are not produced in 'publication-ready' formatting and might require "finishing" edits in an illustration software.

7. Expanding the DEMIX Database

Section contributors: Peter L. Guth, Carlos Grohman, Sebastiano Trevisani, and Conrad Bielski

7.1. Introduction

This section presents the results from three studies presented at the 2023 DEMIX session of Geomorphometry 2023 (Guth et al., 2023e, 2023f, and 2023g). They extend the methodology used by Bielski et al. (2024) and show the incredible range of information present in the DEMIX database (Guth, 2023b) which the original paper could only hint at.

7.2. Subjective Hillshade Criterion

This section is based on the presentation by Guth and others (2023e), with the addition of a wine contest conducted by expert-based visual analysis given in Iasi, Romania.

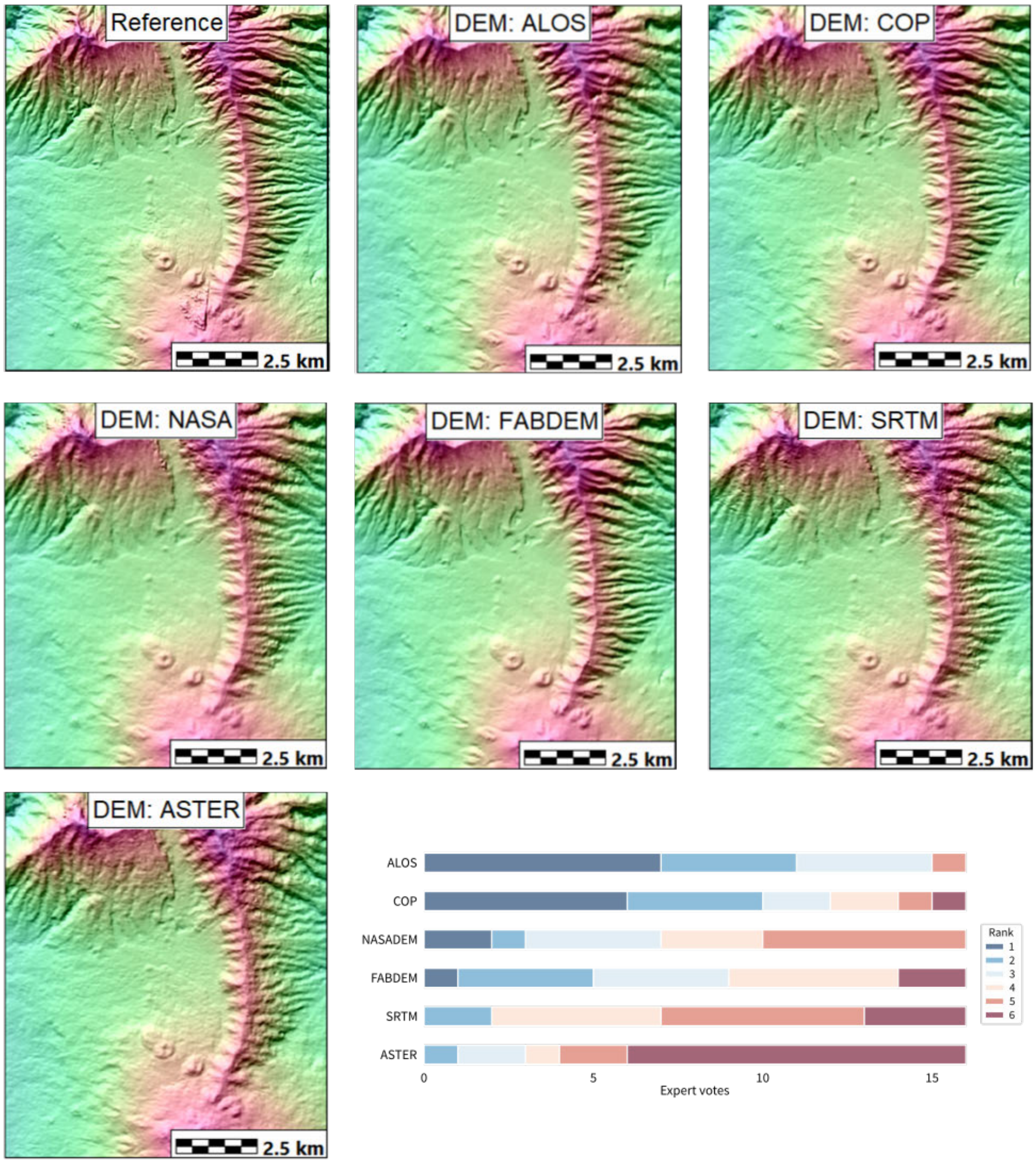
7.2.1. Designing a Hillshade Wine Contest

Digital elevation models (DEMs) represent a fundamental building block for work in science, engineering, social science, government, and the military. DEMs at 1arc second provide the best resolution freely available globally. The DEMIX group is working to compare and rank 6 of those DEMs, and created a database to support their work. The DEMIX wine contest provides a framework for ranking DEMs and providing statistical significance for the results. An oenological wine contest frequently involves subjective assessments from experts, and the DEMIX group noted the ability to use subjective assessments for a DEM wine contest, but did not include any subjective criteria in their initial results. We will use a subjective, visual criterion, show the challenges in applying it to a large number of test areas, and demonstrate that our application of the subjective criterion validates the DEMIX group findings that CopDEM, ALOS, and FABDEM are demonstrably much better than SRTM, NASADEM, and ASTER.

Nothing in the wine contest precludes subjective criteria tests; for demonstration purposes, during spring 2022 we experimented with showing 16 “experts” hillshades of the DTM from DEMIX tile N28VW018B covering part of La Palma in the Canary Islands (**Figure 23**). The DTM was created by aggregating a source DTM from the national mapping agency, using the 2 m DTM to create a 1 second DTM to match the global DEMs. Using a Google form (Google, 2023), we asked the “experts” to rank the subjective visual quality of the maps. In addition to the images, they had an animation cycling though the hillshades which highlights differences. They were not allowed to have ties in their rankings.

During spring 2023 we repeated the contest with a larger number of “experts”, several different test areas (two in the western US, and one in the Italian Alps), and improved methodology. Our initial assumption was that students who constituted the bulk of our “experts” did not know anything about the 6 DEMs, and the original test included the DEM names (as shown in **Figure 23**). The revised test removed the animation and the DEM names, and presented the DEMs in a different random order for each test area. We also ran the contest during Geomorphometry 2023 in Iasi (Romania), both to demonstrate the method and to collect additional data.

Figure 23. Hillshades of DEMIX tileTile N28VW018B, and the distribution of expert opinions for each of the 6 ranks. Low score in best. Later versions of the contest removed the DEM names to avoid biasing the judges.



Source: Peter L. Guth

7.2.2. Geoscience Student Results

The Google Form (Google, 2023) provides the test administrator a figure online (**Figure 24**) showing a quick visualization of the results, as well as individual results from each judge which we do not need. The Form program downloads the results in a CSV file for import into a spreadsheet. **Figure 25** shows the sample hillshades for the three additional DEMS used for the second iteration of the contest. We rearranged the results to get the alternative graphic (**Figure 23**) which we feel more closely shows the results. We also ran statistics (**Figure 26**), using the wine contest Jupyter Notebook (Grohmann 2023b, Chapter 6 on page 50). **Table 5** summarizes the scoring for each iteration of the contest, and **Figure 27** shows the overall evaluation of the overall results.

Figure 24. Google Forms automatic display of survey results.

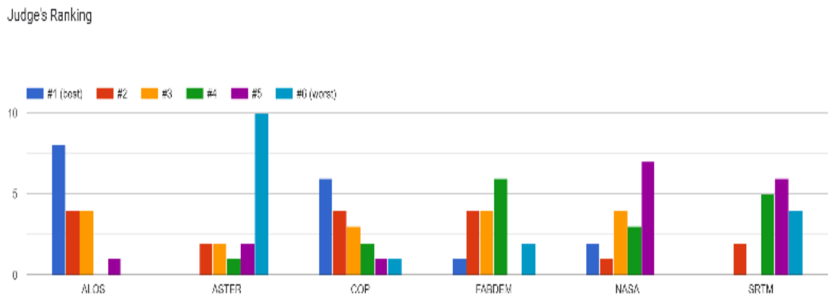


Figure 25. Hillshades for the three additional DEMs used for the second iteration of the contest.

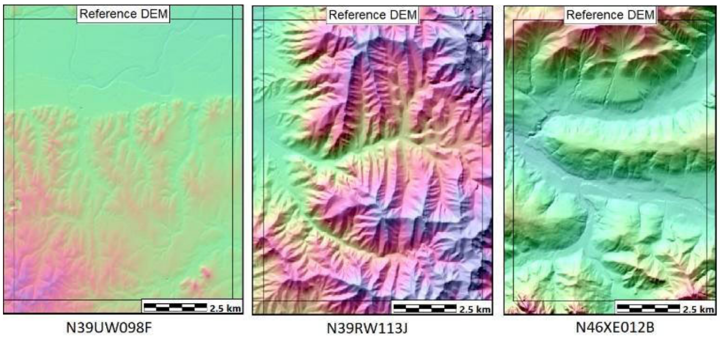


Figure 26. Wine contest ranking and statistical significance matrix. “Ties with” means the DEMs are not statistically different in this test.

```
Ranking without tolerance

No filters applied

Results of the DEMIX Wine Contest

For k=6, CL=0.05, and N=188, the critical value to compare is chi_crit=11.038
And since chi_r (430.675) is greater than chi_crit (11.038)...
Yay!! We can reject the null hipothesis and go to the Post-Hoc analysis!!
```

	Rank	Sum of ranks	Sum of ranks divided by number of opinions	Ties with
FABDEM	1.0	422.5	2.247	ALOS, COP
COP	2.0	457.5	2.434	ALOS, FABDEM
ALOS	3.0	465.0	2.473	COP, FABDEM
NASA	4.0	815.5	4.338	SRTM
SRTM	5.0	818.5	4.354	NASA
ASTER	6.0	969.0	5.154	

Source Fig. 24-26: Peter L. Guth

7.2.3. Contest Results from Geomorphometry Experts

Seven judges, all PhD scientists working in geomorphometry, rendered 16 opinions using DEMs from 4 different study areas (**Table 5**). Results from the Friedman statistics using the Jupyter notebook rank the DEMs in three groups: CopDEM, FABDEM, and ALOS are tied for best; SRTM and NASADEM tie a significant difference back, and ASTER is clearly worst, which only three opinions ranking it 5th instead of last.

Table 5. Iasi wine contest opinions.

Area	**ALOS**	**ASTER**	**CopDEM**	**FABDEM**	**NASA**	**SRTM**
Bolzano	3	6	1	4	5	2
Bolzano	2	5	6	1	3	4
Bolzano	3	6	2	1	5	4
Bolzano	2	6	1	3	4	5
Bolzano	3	5	2	1	3	4
Bolzano	5	6	2	1	4	3
Republican River	4	6	1	2	5	3
Republican River	3	6	2	1	5	4
Republican River	3	6	1	2	5	4
La Palma	1	6	2	3	5	4
La Palma	3	6	2	1	5	4
La Palma	2	6	1	3	5	4
La Palma	1	6	2	3	4	5
Canyon Range	3	6	1	1	4	5
Canyon Range	1	6	2	3	4	5
Canyon Range	2	5	1	3	6	4
Average score	2.56	5.81	1.81	2.06	4.50	4.00

Source: Peter L. Guth

Figure 27. Statistical evaluation of the Iasi wine contest.

	Rank	Sum of ranks	Sum of ranks divided by number of opinions	Ties with
COP	1.0	29.5	1.844	ALOS, FABDEM
FABDEM	2.0	33.5	2.094	ALOS, COP
ALOS	3.0	41.5	2.594	COP, FABDEM
SRTM	4.0	65.0	4.062	NASA
NASA	5.0	72.5	4.531	SRTM
ASTER	6.0	94.0	5.875	

Source: Peter L. Guth

7.2.4. Hillshade Criterion Validates Pixel Differences Distribution Results

The results show a clear preference for CopDEM, FABDEM and ALOS; the results are quantitatively confirmed when using the wine contest statistics. The top three DEMs, and the bottom three significantly lower in the opinion of the judges, are the same as those from the DEMIX results (see Section 4 above, Bielski et al., 2024) which relied on over 20,000 quantitative opinions for 15 criteria using 133 100 km² tiles from 19 areas spread over three continents.

The hillshade map can show elevation with colour, but for the visual perception of hillshade maps the fine-scale morphology based on derivatives of elevation (slope and surface roughness) dominate. In the quantitative DEMIX comparison fine-scale morphology are represented by specific criteria for slope and roughness. For many users these derived parameters are more important characteristics of the DEM, but as emphasized by the DEMIX group, users must select the comparison criteria that most closely match their requirements.

For these areas, the difference between FABDEM and CopDEM are minimal and would be very hard to differentiate in a hillshade map, verified by the similarity in their contest scores. While NASADEM improved on SRTM for the elevation differences compared to the reference DEM, it generally has very little effect on the slope and roughness differences. Since slope and roughness determine the hillshade, the judges did not clearly differentiate NASADEM and SRTM.

Some reviewers had suggested that only highly trained experts could evaluate and rank hillshade maps. Our results show the groupings from such experts produce the same results as our students in geoscience courses and suggest that the results are robust and reliable.

Other potential subjective assessments for DEMs include topographic profiles (Guth, 2010; Grohman, 2018) or elevation-slope plots (Guth, 2006; Guth, 2010; Grohman and Steiner, 2008). DEM quality varies with land cover, landforms, and the slope of the terrain, so the test areas should cover a wide range of conditions.

The design and implementation of an expert-based approach to criteria evaluation is not a trivial task. The approach requires a considerable effort to collect this data and does not easily scale to multiple test regions. The demands on the judges to evaluate multiple DEMs mean that we could never reach the hundreds of test areas, and over a dozen criteria, which are possible with automated quantitative criteria. In our first iteration the DEMs were always in the same order; for the second iteration, we used multiple test areas which always had the DEMs in the same order but varied in not showing the same DEM first or last in every test area. This was the best we could do using Google forms. It would require custom programming to make an ideal survey, and an effort to collect multiple experts willing to judge a number of tiles. Custom programming would also allow judges to give ties.

Despite the challenges, the test shows the power of the wine contest to evaluate DEMs, and that subjective criteria can be used. While the statistical validity of qualitative criteria may have caveats due to relatively small sample sizes, it provides another metric that users can evaluate in deciding which DEM they prefer to use, which in the end comes down to a value judgment. The mean of the differences to terrain parameters cannot be used in the wine contest but means also provide information about where the candidate DEMs are low or high, too steep or too flat, and too rough or too smooth. Users should choose among CopDEM, ALOS, or FABDEM, all of which are very similar to the reference DTM.

7.3. Half Second Grid for Direct Map Comparisons and Differences

This section is based on the presentation by Guth et al. (2023f).

7.3.1. Rationale

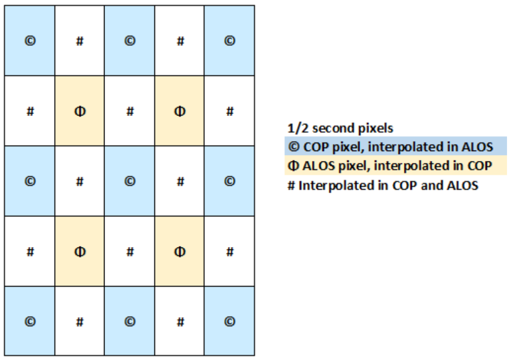
DEMs represent a fundamental building block for work in science, engineering, social science, government, and the military. DEMs at 1 arc second (~30 m) provide the best resolution freely available globally. The DEMIX group is working to compare and rank those DEMs. We use their database to show the geomorphometric and geospatial patterns of the differences. While we performed this analysis for 6 DEMs, space restricts us to concentrate only on CopDEM and ALOS which are demonstrably much better than SRTM, NASADEM, and ASTER. We also limit consideration of FABDEM, which attempts to approximate a DTM but only improves on CopDEM for some landscapes, so that we can look at the simpler case of only two DEMs.

7.3.2. Methods

The wine contest (see Section 4 above and Bielski et al. 2024) uses geomorphometric criteria to rank DEMs. The criteria must allow numerical ranking, which means the evaluations must be unsigned, and the method requires at least three DEMs for the statistical significance to be valid. While the criteria must be unsigned, the signed parameters like the mean and median differences provide important auxiliary information.

We use ½ arc sec grids (**Figure 28**), so that CopDEM and ALOS are equally affected by interpolation. We create the reference DTM and DSM where available by aggregation from high resolution source data. For ALOS and CopDEM the points in the original DEM are in their positions within the ½ sec grid, and they are surrounded by points from bilinear interpolation. This allows direct difference maps between CopDEM and ALOS, and shows the geometric pattern of the differences, without introducing variable interpolation effects between the DEMs. We consider elevation, slope, and roughness defined as the standard deviation of slope in a 5x5 window (Grohman et al., 2011). We make pixel by pixel comparisons for approximately 520,000 pixels in each DEMIX tile which covers approximately 100 km^2. The difference grids allow us to map the differences. We must set tolerances for what we consider significant differences before creating some grids, because the grids classify the map area into categories using the tolerances (**Figure 29**).

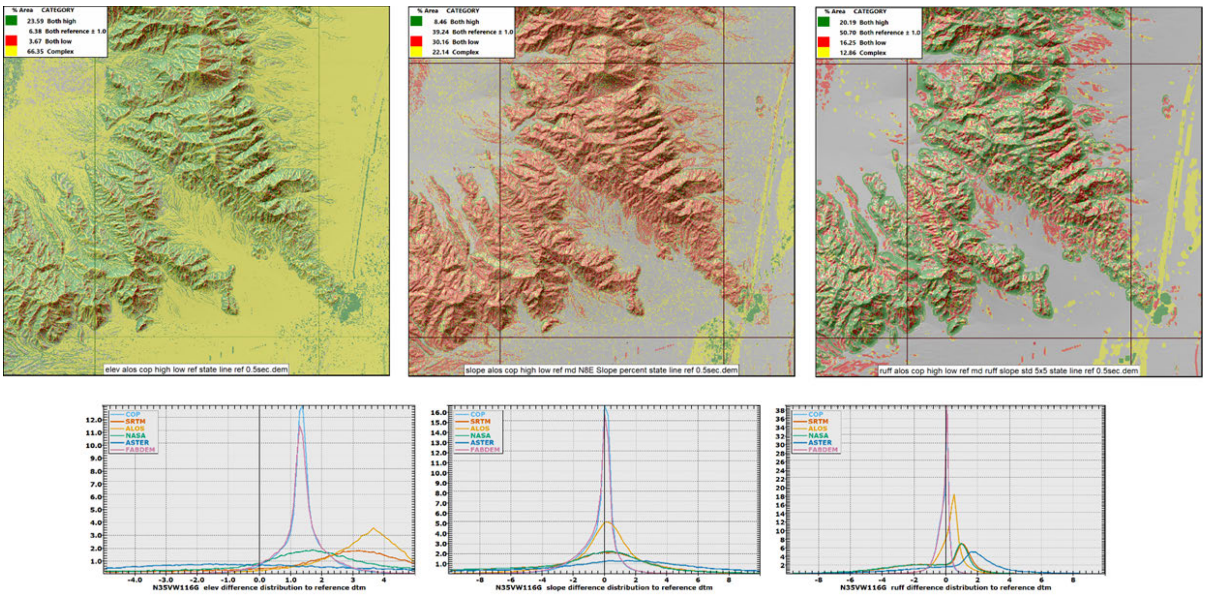
Figure 28. ½ second resampling for arc second DEMs with rectangular pixels. Interpolation reserved ¼ of the points in the grid where the nodes on the 1" and ½" grids coincide, and interpolates new elevations for ¾ of the points in the ½" grid.



Source: Peter L. Guth

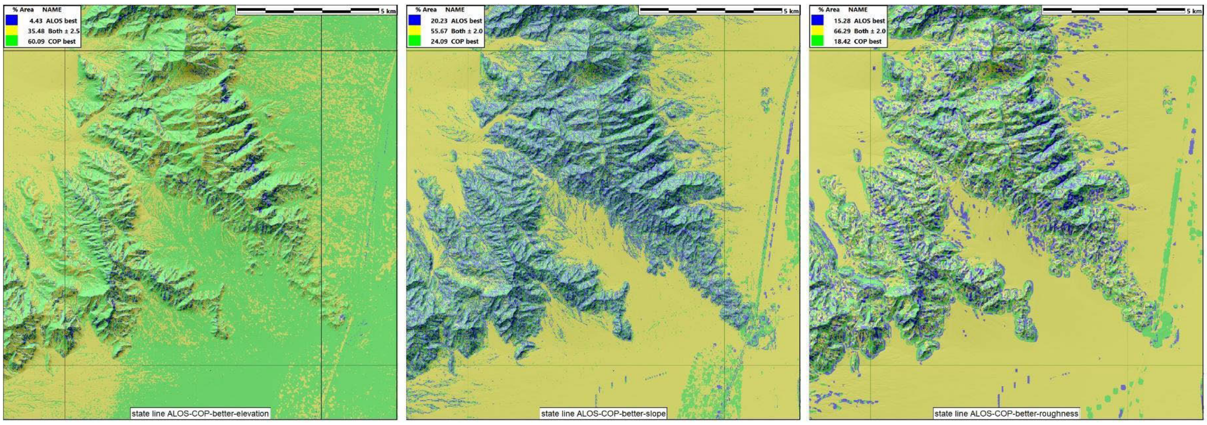
We will show two representations of the results (**Figure 29** and **Figure 30**), and the summary statistics for 20 test areas with over 200 100 km² tiles (**Figure 31** and **Table 6**). **Figure 31** considers the pixel a tie if both CopDEM and ALOS are within the tolerance of the reference DTM; otherwise, the DEM closer in absolute value to the reference DTM is the best.

Figure 29. Comparison of CopDEM and ALOS to the reference half second DTM for DEMIX tile N35VW116G in the centre of the map. Difference maps above and difference histograms below. Instate 15 on the right side of the map shows up difference appearance in the three DEMS.



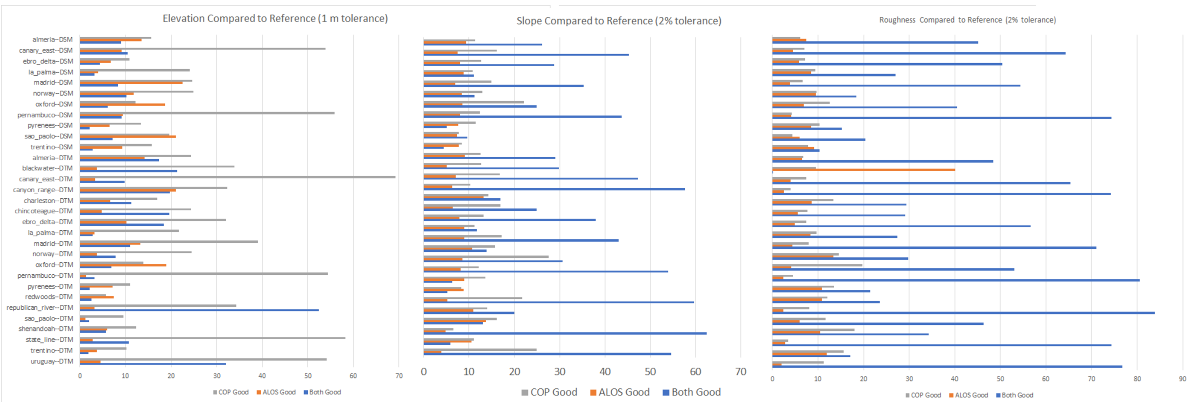
Source: Peter L. Guth

Figure 30. Comparison of CopDEM and ALOS to the reference half second DTM for DEMIX tile N35VW116G. Difference maps above and difference histograms below.



Source: Peter L. Guth

Figure 31. CopDEM and ALOS comparisons to the reference half second DTM and DSM where available for 20 test areas.

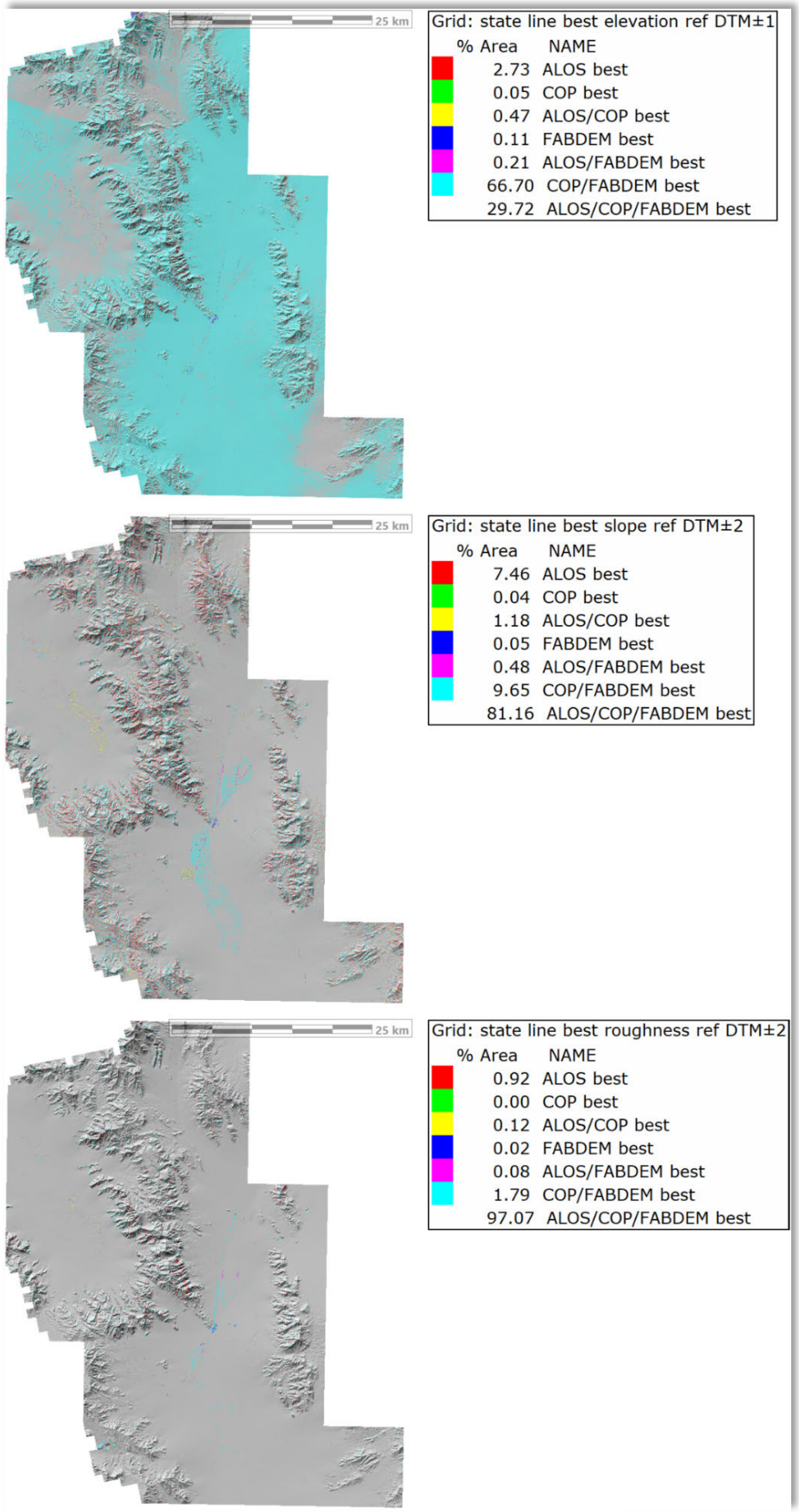


Source: Peter L. Guth

7.3.3. Mapping Locations where CopDEM, FABDEM and ALOS are Best

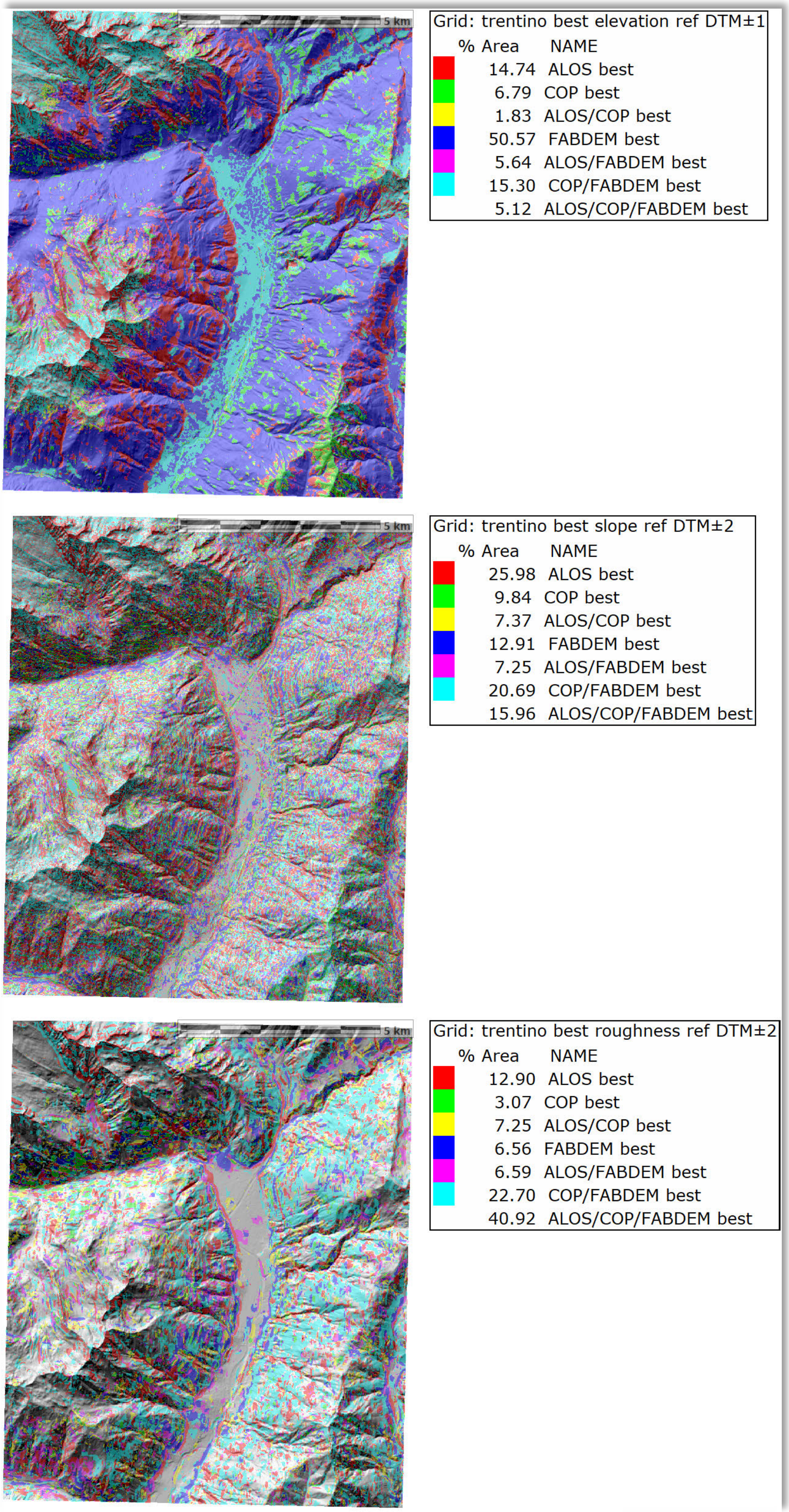
The next series of maps, using ½ second grids, show where CopDEM, FABDEM, and ALOS tie for the best results compared to the reference DTM for three areas, with maps for elevation, slope, and roughness. For each area and parameter there is a map showing where each DEM is best or tied, and one showing nine categories.

Figure 32. State Line (USA) area maps showing the best DEM with the elevation (top) slope (middle) and roughness (bottom) in terms of matching the reference DEM. The associated map legend shows the tolerances applied.



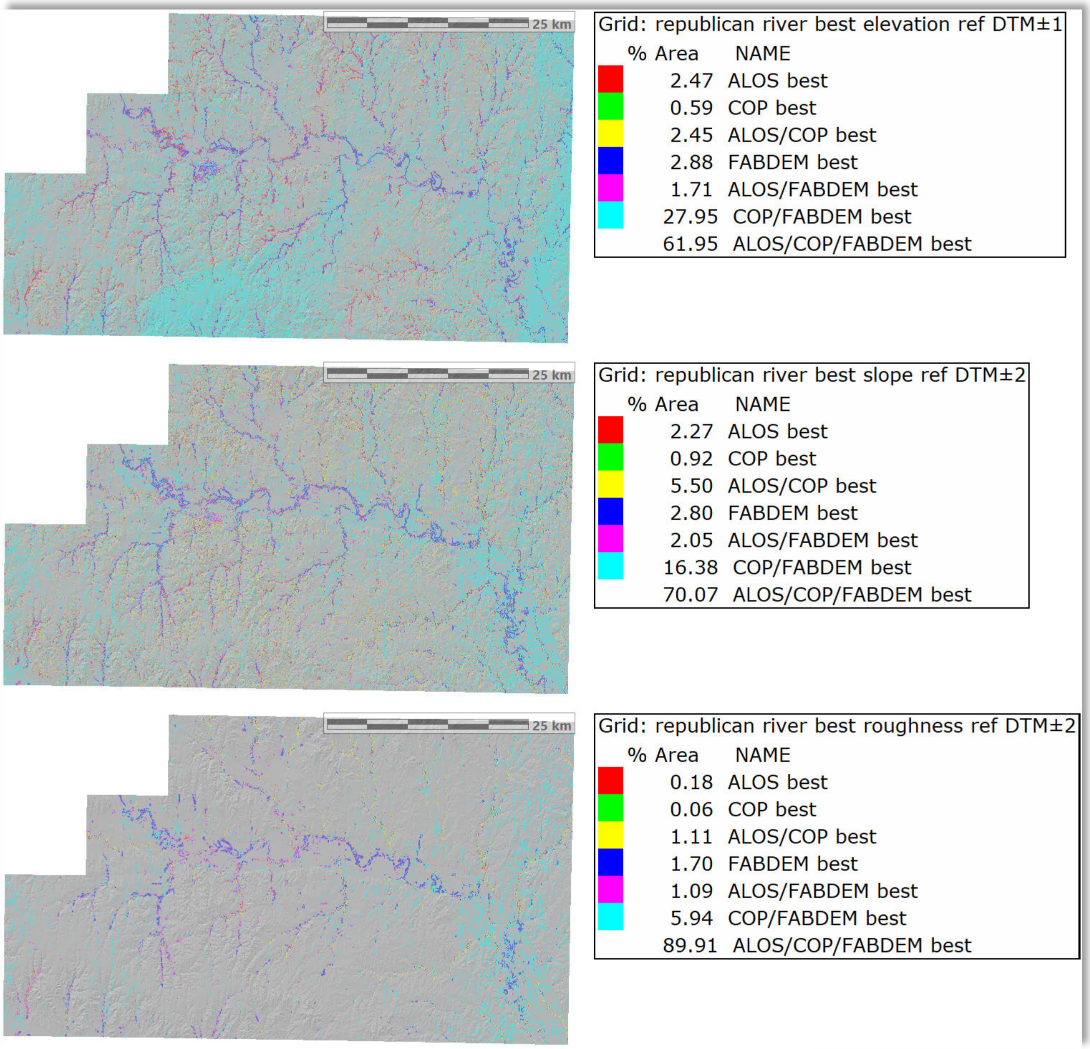
Source: Peter L. Guth

Figure 33. Trentino (IT) area maps showing the best DEM with the elevation (top) slope (middle) and roughness (bottom) in terms of matching the reference DEM. The associated map legend shows the tolerances applied.



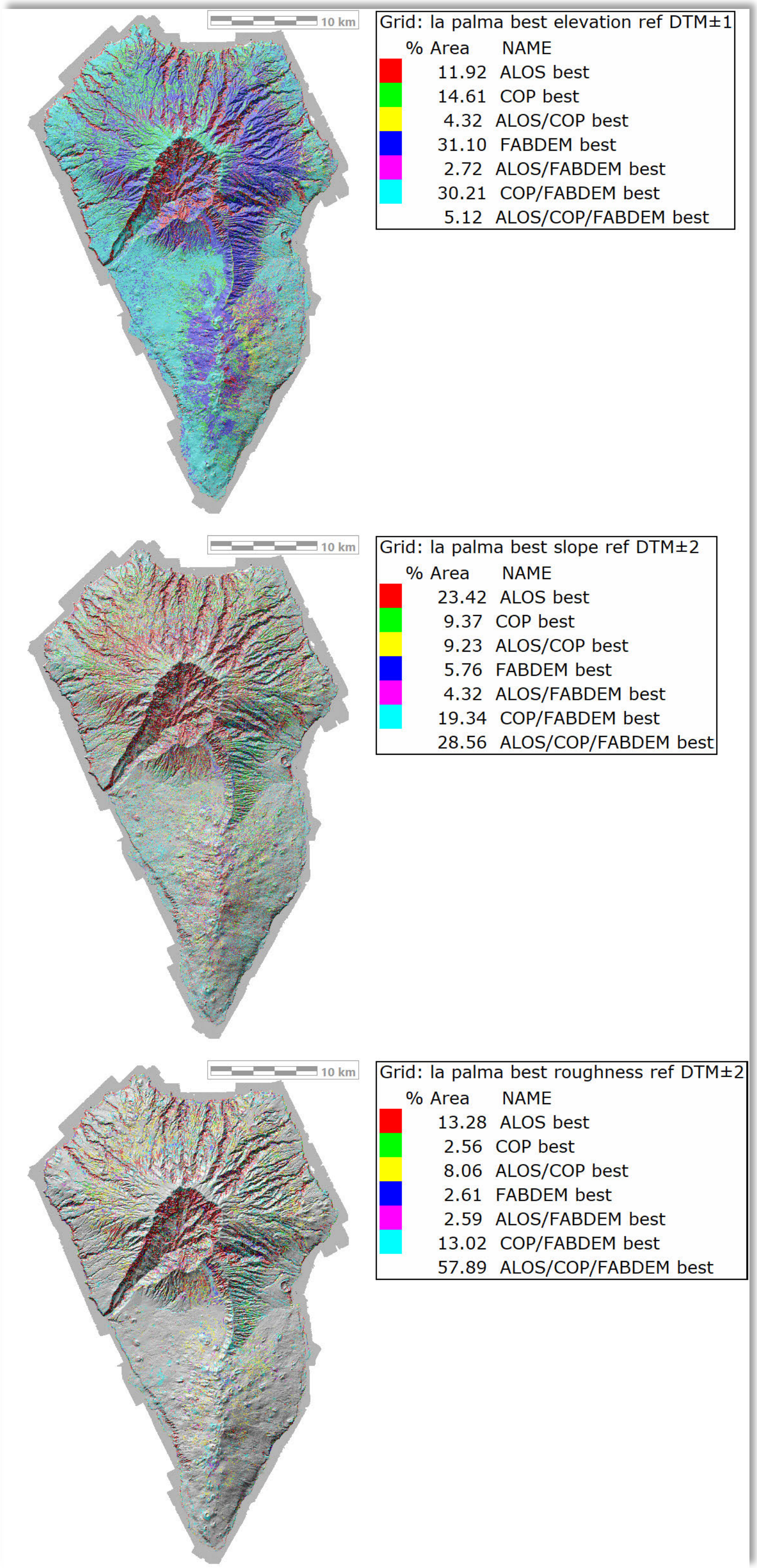
Source: Peter L. Guth

Figure 34. Republican River (USA) area maps showing the best DEM with the elevation (top) slope (middle) and roughness (bottom) in terms of matching the reference DEM. The associated map legend shows the tolerances applied.



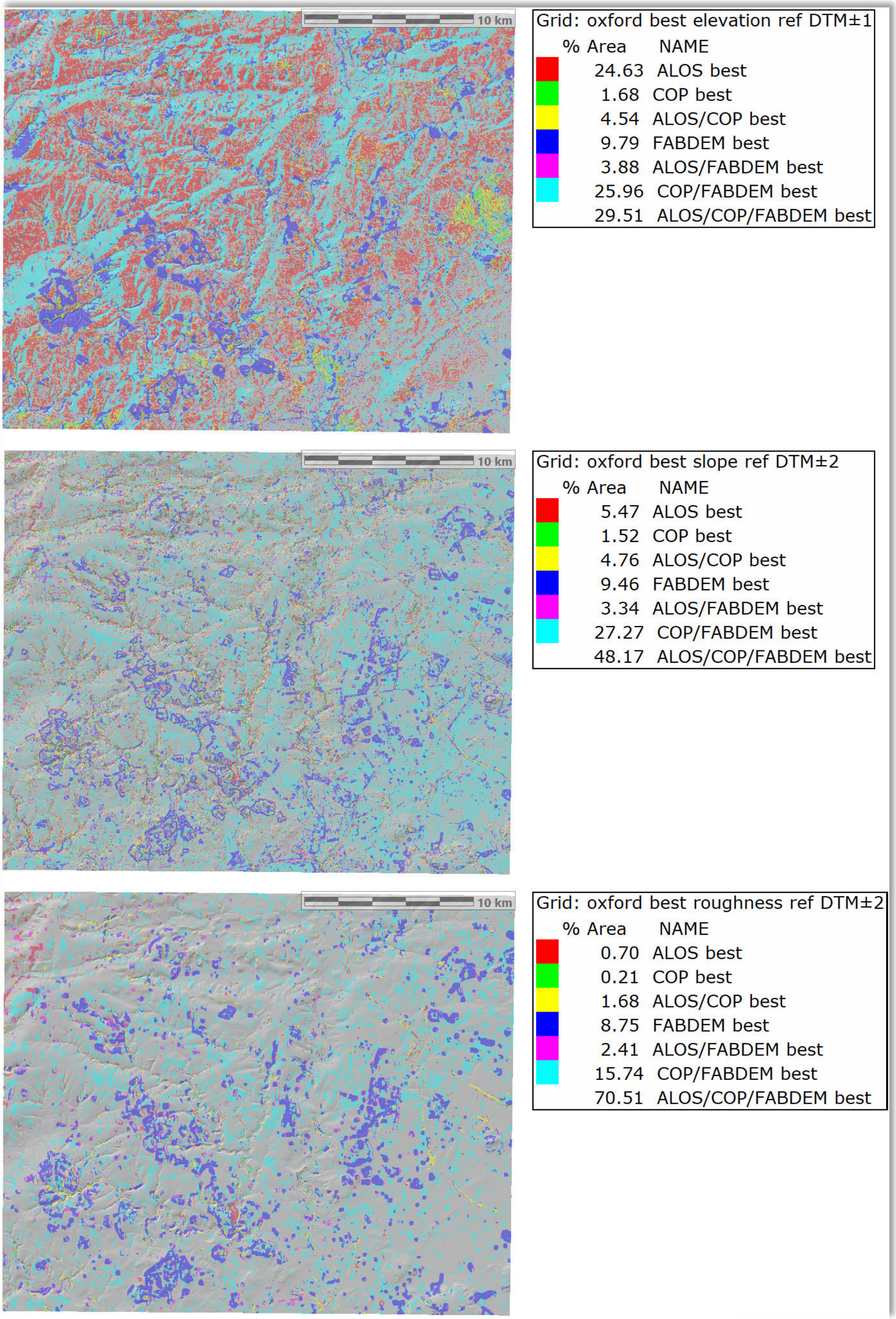
Source: Peter L. Guth

Figure 35. La Palma (ES) area maps showing the best DEM with the elevation (top) slope (middle) and roughness (bottom) in terms of matching the reference DEM. The associated map legend shows the tolerances applied.



Source: Peter L. Guth

Figure 36. Oxford (UK) area maps showing the best DEM with the elevation (top) slope (middle) and roughness (bottom) in terms of matching the reference DEM. The associated map legend shows the tolerances applied.



Source: Peter L. Guth

7.3.4. Results

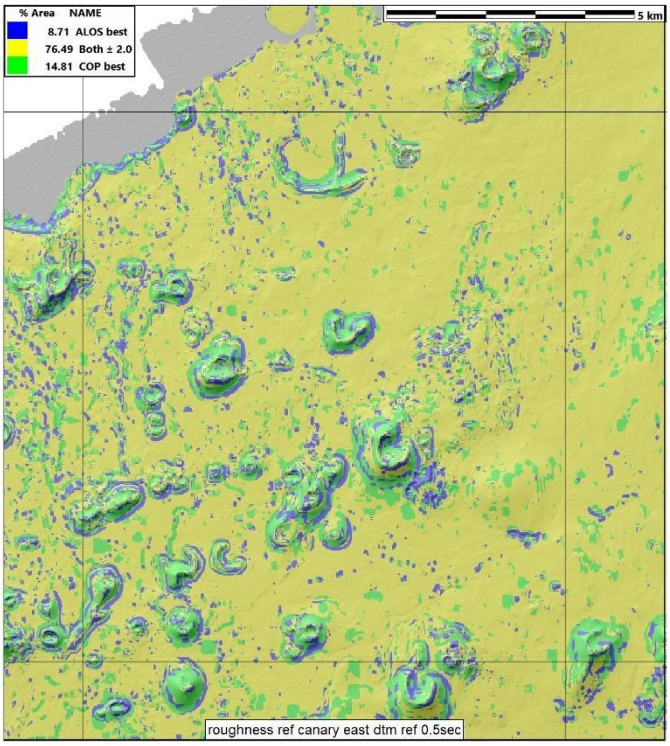
The above presented results (**Figure 32** to **Figure 36**) shows the largest differences from the reference DEM occur in steeper terrain. The computation has nine categories (each test DEM can be high, equal within tolerance, or low). The complex category lumps into 6 categories.

The elevation bias in this tile is not representative of all tiles; the lack of bias for the slope and roughness are representative. The very small standard deviations are characteristic of CopDEM and FABDEM, as is the greater dispersion for ALOS, and the large dispersion for the other DEMs with a poorly defined mode.

Despite the negative overall bias for slope difference, the distribution skews toward less steep slopes. This results from comparing the ½” reference DEM with an interpolated test DEM which has lower average slopes.

Figure 37 shows roughness differences for a tile in the Canary Islands. The volcanic cones show the effect of trying to compare a pixel-is-area DEM and a pixel-is-point. The half pixel offset means that peaks, ridge crests, and valley bottoms in ALOS and CopDEM occur at different locations, and each is “better” compared to the reference DEM about half the time and the map pattern closely mimics the topography.

Figure 37. Roughness differences between CopDEM and ALOS from a reference DTM for DEMIX tiles N29MW014D.



Source: Peter L. Guth

Table 6. Best DEM by area.

	DTM		DSM	
	ALOS	CopDEM	ALOS	CopDEM
Elevation	2	18	2	8
Slope	1	19	0	10
Roughness	1	19	3	7

Source: Peter L. Guth

7.3.5. Discussion and Conclusion

The wine contest criteria used by DEMIX use 5 metrics for the unsigned difference distribution. For each parameter (elevation slope, roughness), the 5 metrics are highly correlated and do not provide greatly different results. The 5 metrics progressively get larger as they factor in increasingly extreme values in the tails of the distribution, but the global DEMs generally follow the same ranking. Beyond the evaluation numbers and rankings, the criteria do not provide a simple, visual and intuitive way to assess the results.

Our three new criteria for elevation, slope, and roughness take the percentage of points in the DEM where each DEM is closest to the reference DEM and ranks the DEMs. As designed these criteria have less influence from the tails of the difference distribution. With only two DEMs in our test, we can create maps showing the spatial patterns, and relate them to the characteristics of the region and see how slope, roughness, and aspect affect where ALOS and CopDEM diverge from the reference DEM. The maps allow a subjective assessment of DEM quality that adds to the quantitative results; as one example, we have noted that ALOS frequently has small anomalies tacking the satellite's orbit path where different images where merged, something we did not observe with CopDEM.

We assume that the reference DEM is the closest to a true value of what the elevation should be for a 1 arc second pixel. Aggregating the high resolution (reference) DEM to 1 arc second necessarily loses detail, and the largest differences to the global DEMs occur in steep terrain. In these pixels there is a large variation in elevation, and picking a single value to represent the pixel presents a challenge. We must consider the possibility that the choice for the elevations in the reference DEM has much more uncertainty in steep areas than in flat areas.

These comparisons (**Table 6**) show that CopDEM is clearly better than ALOS, but the differences are generally small. CopDEM is better in the most heavily forested areas, reinforcing the suggestion that the radar sensor has greater penetration in the canopy compared to an optical sensor as suggested in evaluating the positions of the global DEMs in lidar point clouds (Guth and Geoffroy, 2021).

Using ½ second grids allows direct comparison of ALOS and CopDEM but does affect the slope differences and clearly shows the ½ pixel difference between the two DEMs.

7.4. Mining the DEMIX Database

This section is based on the presentation by Guth et al. (2023g).

7.4.1. Rationale

DEMs provide a critical base layer for almost all earth science studies. The DEMIX group is working to compare and rank 6 global DEMs: CopDEM, FABDEM, ALOS, NASADEM, SRTM, AND ASTER (Bielski et al., 2024; Grohmann et al., 2023a). We use their database (Guth, 2023b) and reference DEMs (Guth et al., 2023d) to demonstrate preliminary results of the rich potential to mine those datasets. Two other papers at this conference have examined the ability to use qualitative evaluation of hillshade maps to rank DEMs (Guth et al., 2023e), and to look at the difference maps to study the spatial patterns where ALOS and CopDEM differ (Guth et al., 2023f).

7.4.2. Results

The wine contest (Bielski et al., 2024) cannot use signed results because their choice of statistics requires that the values be ranked from low score (best) to high score. They commented that signed values like the mean and median provide additional context, which we will explore with **Figure 38** which shows the mean, median, mode, and standard deviations of the difference distributions of elevation, slope, and roughness for all 236 tiles in the database. This summarizes the individual difference histograms for each tile. It allows the following generalizations:

- The difference distributions for CopDEM and FABDEM, and to a lesser extent ALOS, have a very sharp peak close to 0, indicating very little bias compared to the reference DTM. The other DEMs have much smaller and flatter peaks; for many it is hard to find a mode.
- NASADEM has little bias from the reference DTM only for elevation differences, where it is much better than SRTM.
- NASADEM does not improve on slope and roughness compared to SRTM.
- ASTER is clearly the worst performer.

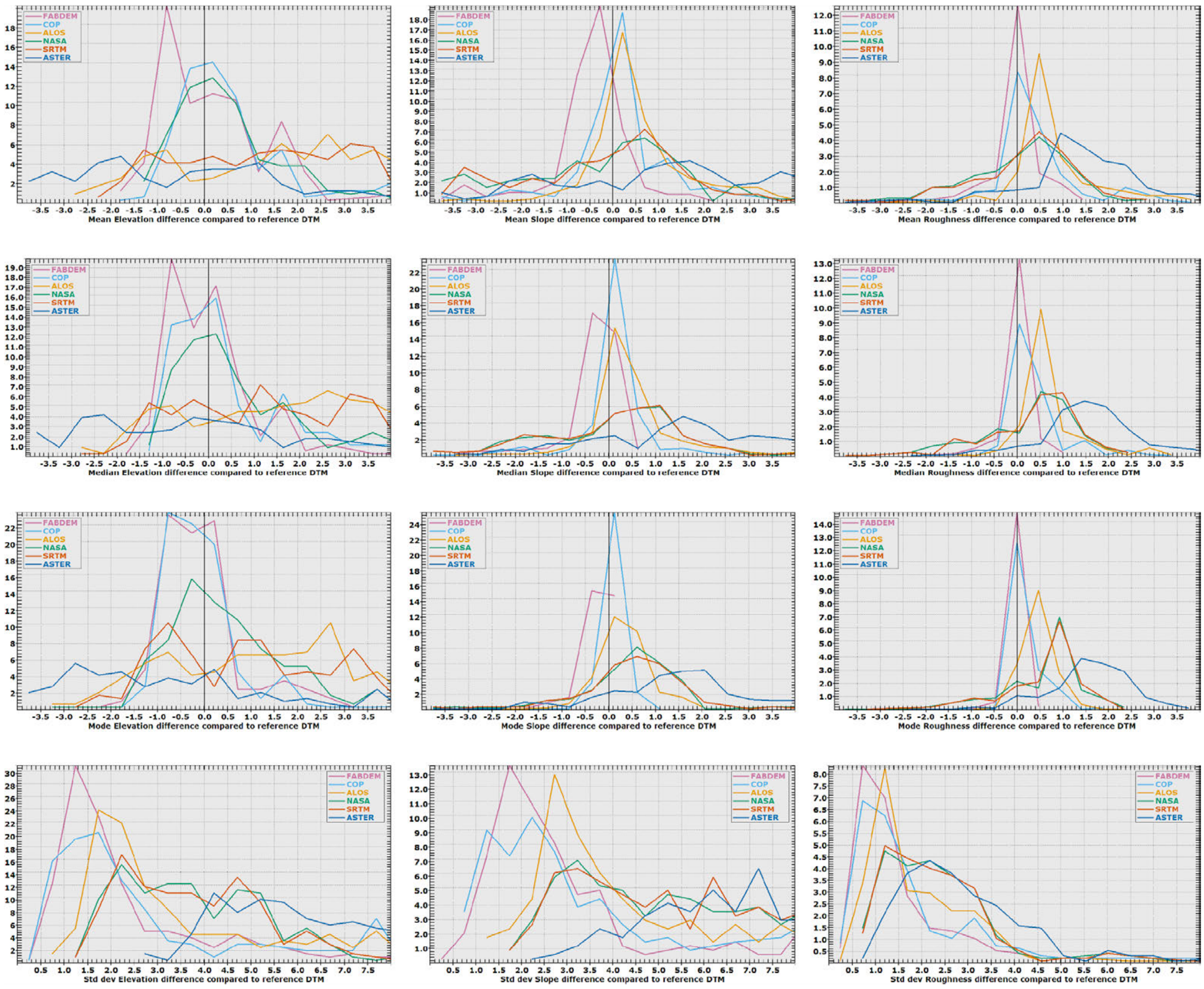
Overall, **Figure 38** supports the DEM ranking from the wine contest (Bielski et al., 2024) but provides insight into how the DEMs differ.

Figure 39 shows where each of 6 DEMs is tied for best, with the tiles sorted by average slope and percent forest, using three criteria of the 15 used in the wine contest (Bielski et al., 2024). There are 236 opinions, one per tile for each criterion with the DTM as reference and 134 opinions for the DSM as reference. There can be 1 to 6 DEMs tied, and the figure suggests where each is best:

- Fewer tiles have a DSM, which must be factored into looking at the figure.
- CopDEM and FABDEM are almost always at the top, FABDEM more often when compared to a DTM and CopDEM when compared to a DSM.
- Based on a limited number of very steep and very rough tiles, ALOS may outperform the other DEMs in that type of topography.
- NASADEM and SRTM perform best in low slope tiles, and in unforested tiles and some tiles with about 70% forests.

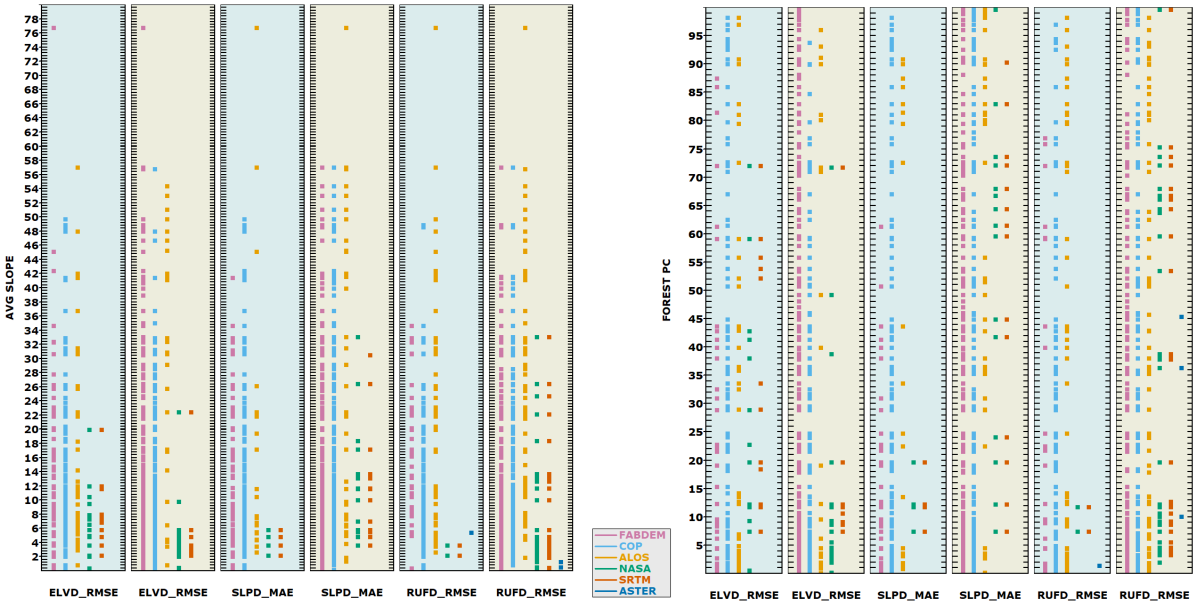
- ASTER is tied for first for only 6 tiles, two of which are very flat coastal DEMs in which none of the 1 arc second DEMs performs very well.

Figure 38. Summary statistics (mean, median, mode, and standard deviation) for the difference distribution compared to the reference DTM for 6 global DEMs and 236 DEMIX tiles. Graphs of the difference distributions are available online (Guth, 2023c).



Source: Peter L. Guth

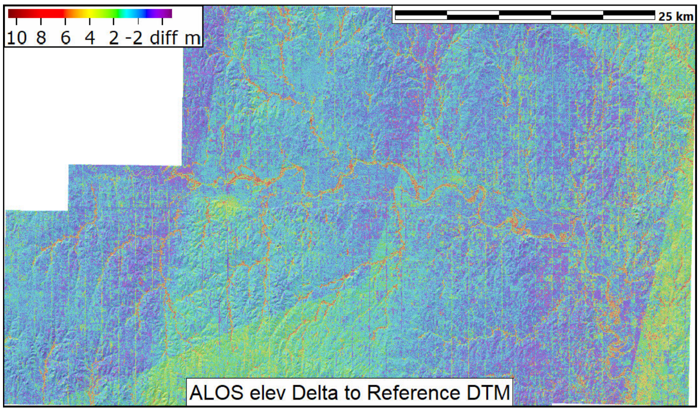
Figure 39. Tile by tile results where each DEM is ranked best within tolerance, for three parameters and both DSM (blue background) and DTM (brown background). On the left side the tiles are sorted by average slope of the tile in percent, and on the right by how much of the tile is forested.



Source: Peter L. Guth

Inspection of the difference maps reveals aspects of the DEMs that do not always show up in the statistics. For example, **Figure 40** shows linear anomalies with the ALOS DEM showing where images were merged; we have seen the same patterns elsewhere with ALOS and notably have not seen similar artefacts with CopDEM. The scale of these anomalies makes them hard to see on individual DEMIX tiles and is one reason we prefer to work with larger areas with many DEMIX tiles. The other, and perhaps bigger reason, is that the geographic tile boundaries do not line up well with the UTM boundaries of the source DEMs used to aggregate the reference DEMs, and it is much easier to get data for an area and then extract the individual tiles on the fly.

Figure 40. Elevation differences, ALOS minus reference DTM for the Republican River test area in Kansas.



Source: Peter L. Guth

Figure 41 and **Table 7** show the results from the DEMIX database, with the power of the filters possible. They show the results combining all 15 criteria, sorted by the average slope in the tile, and with a separate plot for each of the 8 land types. For the ALL land type, there are 3540 opinions for the DTM (236 tiles * 15 criteria), and 2010 for the DSM (134 * 15). The number of opinions drops for the other land types, because some test tiles do not have that land type. The coloured symbols show where each DEM is tied for the best performance. If there are multiple DEMs with the same average slope, they will plot on top of each other. If the DEM ties for only one criterion, there is no way to distinguish it for another DEM which is best for all 15 criteria. The table provides additional insight. It shows the percentage of the opinions where each DEM was tied for best.

The graphs show that CopDEM, FABDEM, and ALOS perform similarly, but the table shows how many more opinions favour CopDEM and FABDEM.

This analysis shows the rare cases where ASTER performs well, which is overall no more than about 2% of the opinions. This is only for very flat tiles (ASTER appears only at the bottom of the graph), and primarily for only a few land types.

Figure 41. Combined results for the seven land types all 15 DEMIX criteria, showing the results sorted by average slope for the test tile.

Source: Peter L. Guth

Table 7. DEM winning percentages for 15 criteria.

REF_TYPE	LAND_TYPE	OPINIONS	FABDEM	CopDEM	ALOS	NASA	SRTM	ASTER
DSM	ALL	2010	43.18	91.19	31	4.18	4.08	0.3
DSM	FLAT	2010	48.16	95.27	28.96	6.27	5.27	0.2
DSM	GENTLE	1985	34.86	91.99	25.44	3.83	4.03	2.07
DSM	STEEP	1965	40	83.31	24.07	3.41	3.46	2.09
DSM	CLIFF	1620	45.62	72.28	32.53	3.52	3.64	1.91
DSM	URBAN	1815	13.44	88.98	64.74	6.56	5.07	0.72
DSM	FOREST	1740	21.78	87.93	30.4	4.94	5.11	0.06
DSM	BARREN	1935	63.77	92.25	28.17	4.39	3.98	0.1
DTM	ALL	3540	86.3	68.81	21.3	10.96	9.32	1.36
DTM	FLAT	3540	91.1	67.66	21.5	13.14	10	1.64
DTM	GENTLE	3285	69.16	80.24	24.78	11.32	10.47	1.22
DTM	STEEP	2850	67.96	80.88	18.6	8.42	8.28	0.98
DTM	CLIFF	2290	62.88	68.86	33.14	3.45	3.36	0.17
DTM	URBAN	2720	87.43	57.39	17.72	11.51	9.34	0.99
DTM	FOREST	3015	70.98	54.33	31.41	20.7	18.64	1.82
DTM	BARREN	3465	87.36	78.15	21.99	6.78	6	1.1

Source: Peter L. Guth

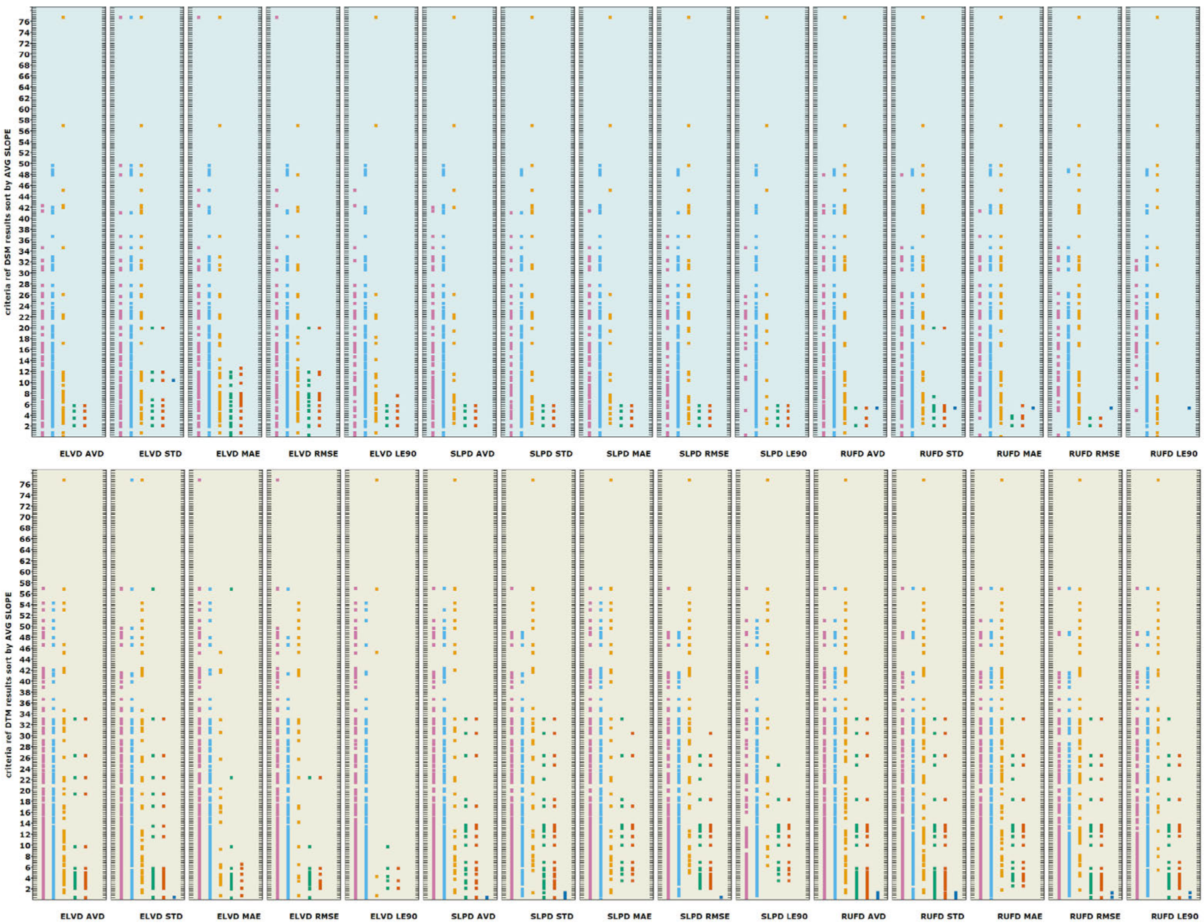
7.4.3. Which Criteria Work Best

Figure 42 shows the results by tile for each of the 15 criteria, using the ALL and type separately for the reference DSM and DTM. For the DSM, it suggests that for the two steepest tiles ALOS is best, but for the DTM only the steepest tile. This sampling, one or two tiles out of over 200 for the DTM, suggests that the generalization might not be generally, and the results for the tiles with slopes of 30-50% vary greatly depending on which criteria are considered.

7.4.4. Conclusions

We foresee several ways that the DEMIX wine contest and its associated database will be extended (1) adding more tiles, particularly in mountainous areas, (2) adding new criteria, and (3) looking beyond the numerical rankings to understand what causes the differences, and when each DEM performs best. We do not expect that the relative rankings of the DEMs will change much, but that understanding where ALOS performs best could lead to improved DEMs in areas where CopDEM underperforms; understanding the difference between optical and radar sensors might lead to better composite DEMs. All of these DEMs are composites, using additional data to fill voids, and FABDEM removed trees and buildings from CopDEM very well in creating the best (and only) DTM in the group.

Figure 42. DEMs which tie for best on each of the 15 criteria, sorted by average slope.



Source: Peter L. Guth

8. Copernicus DEM Wins

Section contributors: Conrad Bielski, Carlos H. Grohmann, Peter Guth, Carlos López-Vázquez, Laurence Hawker, Dean Gesch, Sebastiano Trevisani, Virginia Herrera-Cruz, Virginia Herrera-Cruz, Serge Riazanoff, Hannes I. Reuter, Peter Strobl

8.1. Introduction

The DEMIX SG2 developed, implemented and applied the wine contest in order to produce an intercomparison of the six global digital elevation models which included the Copernicus DEM (CopDEM) product (Bielski et al., 2024). Fifteen criteria related to elevation, slope and roughness were used for this intercomparison and demonstrated the superiority of the CopDEM over the other global DEMs included in the contest. From an overall final ranking of the global 1" DEM intercomparison, CopDEM and FABDEM are clearly the frontrunners based on the chosen criteria and test areas: CopDEM when compared to a DSM and FABDEM when compared to a DTM. The best rankings depend on the choice of parameters and what the user wants to do with the DEM. ALOS was generally the third best but sometimes moving into second place.

The following section builds on the conclusions presented in the DEMIX article (Bielski et al. 2024) to provide more support for retiring other DEMs including the EU-DEM product[1] due to the strong dependence on the SRTM and ASTER global products which scored worse in the DEMIX wine contest. An important conclusion and recommendation of the DEMIX SG2 is that SRTM be retired and ASTER be avoided.

8.2. DEMIX Wine Contest Suggests Using Copernicus DEM

An important achievement of the DEMIX methodology was to provide DEM users with a flexible procedure that can be applied anywhere for the intercomparison of DEMs. The procedure permits a final DEM ranking considering user-defined and user-chosen criteria that are supported by user chosen statistical significance level. The statistical significance associated to the ranking highlights those situations where the DEM rankings cannot be differentiated from a random result. Users should be aware that the selection of different criteria may lead to different ranking results with respect to what was published in Bielski et al. (2024). The considered criteria were applied because they cover relevant characteristics of DEMs and produce compatible results to the ones obtained considering the visual perception of DEMs by experts (see Section 7 above). In particular, in the comparison performed the criteria are based on statistical metrics calculated on three main groups of variables: elevation, slope, and roughness (computed as standard deviation of slope), differences against a reference DEM. Considering the high number of tiles and the wide range of landscapes included in the analysis, the results of the wine contests provide general and useful information concerning the relative quality of the selected global DEMs when compared to fiducial local data.

The result of the DEMIX intercomparison based on the wine contest showed that most of the time FABDEM and CopDEM are the winners, respectively when considering reference DTMs and DSMs. ALOS is ranked only slightly lower, and in some rare settings, e.g., high roughness, steep terrains, is

[1] https://www.eea.europa.eu/en/datahub/datahubitem-view/d08852bc-7b5f-4835-a776-08362e2fbf4b?activeAccordion=735550

tied with CopDEM or even better. However, criteria concerning the presence of artefacts have not been considered in the ranking and neither the global completeness in coverage. This is particularly relevant for the ALOS product, which as outlined in various works and also in the Section 7 above, can significantly be affected by artefacts and data voids.

The FABDEM product outperforms or is tied with CopDEM when the reference is a DTM, especially for the criteria related to elevation differences. However, in steep terrains FABDEM occasionally appears inferior to CopDEM, even when the reference is a DTM. This can be expected in some situations given that FABDEM is derived by post-processing CopDEM by means of machine learning approaches. Accordingly, the predicted elevation depends both on the predictive approach adopted as well as on the characteristics such as the resolution, uncertainty, etc. of the input features considered. The input features considered which have the greatest effect when applying such novel techniques include the layers from which the elevations of building objects and trees for example are derived and whose quality is not homogenous around the globe (Hawker et al. 2022). Another important aspect for users to consider is related to the licensing constraints of FABDEM because it has been published using a Creative Commons "CC BY-NC-SA 4.0" licence that restricts the use to non-commercial purposes only. Consequently, CopDEM can be considered a safe and reliable choice for any user when the analysis requires both a DSM as well as a DTM because it does not have a restrictive usage license.

In some settings, specifically in lowland coastal areas, FABDEM appears to be the better choice when a DTM is required based on the wine contest outcomes. However, it is highly recommended that users take care to check the input features used for the supervised prediction and inspecting the elevation differences with CopDEM. A clear indication of potential issues can be identified when FABDEM elevation data is higher than CopDEM in such lowland areas because it would be produced by the prediction model and not a real measurement.

8.3. Consistency of DEMIX Wine Contest Results with Other Studies

In general, the DEMIX wine context results, including the final suggestion to adopt CopDEM, are compatible with many other DEM studies that focus on specific areas around the world and consider other approaches and DEM characteristics to perform the DEM quality evaluation. These studies were conducted both by researchers involved in the DEMIX group as well as by other scientists and practitioners outside of the DEMIX group. The interesting aspect is that despite the many different approaches and perspectives, including the lack of a final ranking with a confidence level, all these different studies converge with the main findings of the DEMIX intercomparison which further supports the robustness of the DEMIX wine contest results.

Guth and Geoffroy (2021) evaluated five global DEMs (FABDEM had not yet been released) considering as reference LiDAR point cloud elevations and ICESat-2. Terrestrial lidar data typically has 1500-10,000 returns within each pixel of the 1” DEM from vegetation, ground slope, and man-made features. In open flat terrain, all 5 DEMs approximate the ground surface. In other areas, the Copernicus DEM elevations cluster tightly near the centre of the elevations within the point cloud, while the other DEMs are spread throughout and even outside the lidar point cloud. Guth and Geoffroy (2021) showed the superiority of the Copernicus DEM is evident in tests against both lidar and ICESat-2 data, in eight (high relief) test areas, in different vegetation types, and on both gentle and steep slopes.

The ESA Copernicus DEMs Quality Assessment Summary (Gross et al. 2021) presents an extensive evaluation of the characteristics of CopDEM, including technical details on the sensing methodology. Details can be found in Section 11.

The assessment of global DEMs for mapping low elevation coastal zones by Gesch (2023) is an interesting domain-specific study dealing with a key issue linked to sea level rise and related hazard mapping. Global DEMs have been used extensively for coastal hazard assessments, but their effectiveness in accurately mapping and assessing risk and vulnerability in the coastal zone is directly related to data quality, namely their vertical uncertainty. This study evaluated global DEMs for high-confidence mapping of the low elevation coastal zone (≤ 10 m above sea level) through comparison with very high-accuracy reference elevation benchmarks, resulting in vertical accuracy statistics. The results show that among the 30-m global DEMs, CopDEM and FABDEM (which is derived from CopDEM) are clearly the best performers in terms of absolute vertical accuracy producing the lowest vertical RMSE and smallest mean error, or vertical bias in the studies performed. Both CopDEM and FABDEM allow for extremely high-confidence (> 95%) delineation of the low elevation coastal zone and are also suitable for mapping of a 5-m shoreline zone at a high confidence level (> 68%).

Another interesting domain-specific study was the evaluation of CopDEM for Landsat Collection-2 Processing by Franks and Rengarajan (2023). This study evaluated the global suitability of CopDEM for the orthorectification of Landsat imagery and compares its performances with the USGS Collection-2 DEM, conventionally adopted for the processing. The study concludes, even if the results are not homogeneous in the different regions of the globe, that in general CopDEM can be preferred for the Landsat orthorectification and conclude that using CopDEM will improve the results over the current Collection-2 DEM. One of the factors leading to the spatial heterogeneity of the results is due to the fact that the quality of CopDEM and Collection-2 DEM change according to the data and approaches adopted in the gap filling process. From this perspective, it should be underlined that in high relief and steep areas CopDEM is characterised by a high proportion of voids filled data. Apart from these specific areas, the CopDEM quality at the global level is quite uniform.

Some of the DEMIX wine contest findings relating to CopDEM versus ALOS in steep terrain are coherent with the results of the study of Trevisani et al. (2023), concerning a significant alpine area (6500 km^2) and focusing on the capability to reproduce fine scale terrain morphology. A LiDAR based DTM was used as reference (coarsened via pixel aggregation at 25 m) and the fine-scale morphology is characterised considering the residual DTM and a short-range omnidirectional roughness index. The study highlights that for most of the study area CopDEM is superior to ALOS, with the exception of very steep and rough terrains. It also evidences the widespread presence of artefacts in ALOS and the limitations in doing DEM comparisons on the sole basis of metrics of elevation differences obtained on a pixel-by-pixel basis. It outlines how steep slopes and peaks in mountain environments are characterised by widespread gaps in both DEMs original source data, requiring the extensive integration with other products (e.g., SRTM). Tendentially, CopDEM appears smoother than ALOS in regard to the omnidirectional roughness index considered. The study also highlights the relevance, for the quality evaluation of DEMs, of geomorphometric parameters such as roughness indices and residual elevation, capable of describing specific characteristics of the spatial variability of topographic surfaces.

The necessity and the advantages of moving beyond a pixel-by-pixel evaluation of elevation differences is also explored by the work of Purinton and Bookhagen (2021), where the DEM comparison between ALOS, ASTER, NASADEM, TANDEM-X and CopDEM) was performed taking into account fine-scale morphology considering the residual DEM and an analysis of inter-pixel consistency metrics using spectral approaches such as Fourier analysis on high-pass filtered hillshade. The inter-pixel consistency metrics are essentially surface roughness indices. These approaches also take into account the presence of specific DEM artefacts. A limitation of the study is that the comparison is performed without considering a reference DEM which implies some assumptions in the analysis, e.g. CopDEM free from longer wavelength's spectral peaks and a short-range smoothness of topography. The study area is located in the Central Andes of western Argentina covering an area without or with little vegetation and a relatively smooth topography. In the area considered, CopDEM and TANDEM-X resulted to be the most suitable DEMs from the perspective of their smoothness and absence of long wavelength artefacts.

In the work of Karlson et al. (2021), the evaluation was carried out considering, in addition to vertical accuracies with respect to a Lidar based reference DTM, the performances in deriving the topographic wetness index (TWI) and a morphological classification based on the geomorphons approach. The candidate DEMs considered were Arctic DEM, ASTER, ALOS and CopDEM covering a watershed of size 18,000 km^2 in northern Sweden that was characterised by different types of land cover. The study reports that CopDEM performed the best both in terms of vertical accuracy as well as in regard to TWI and landform classification.

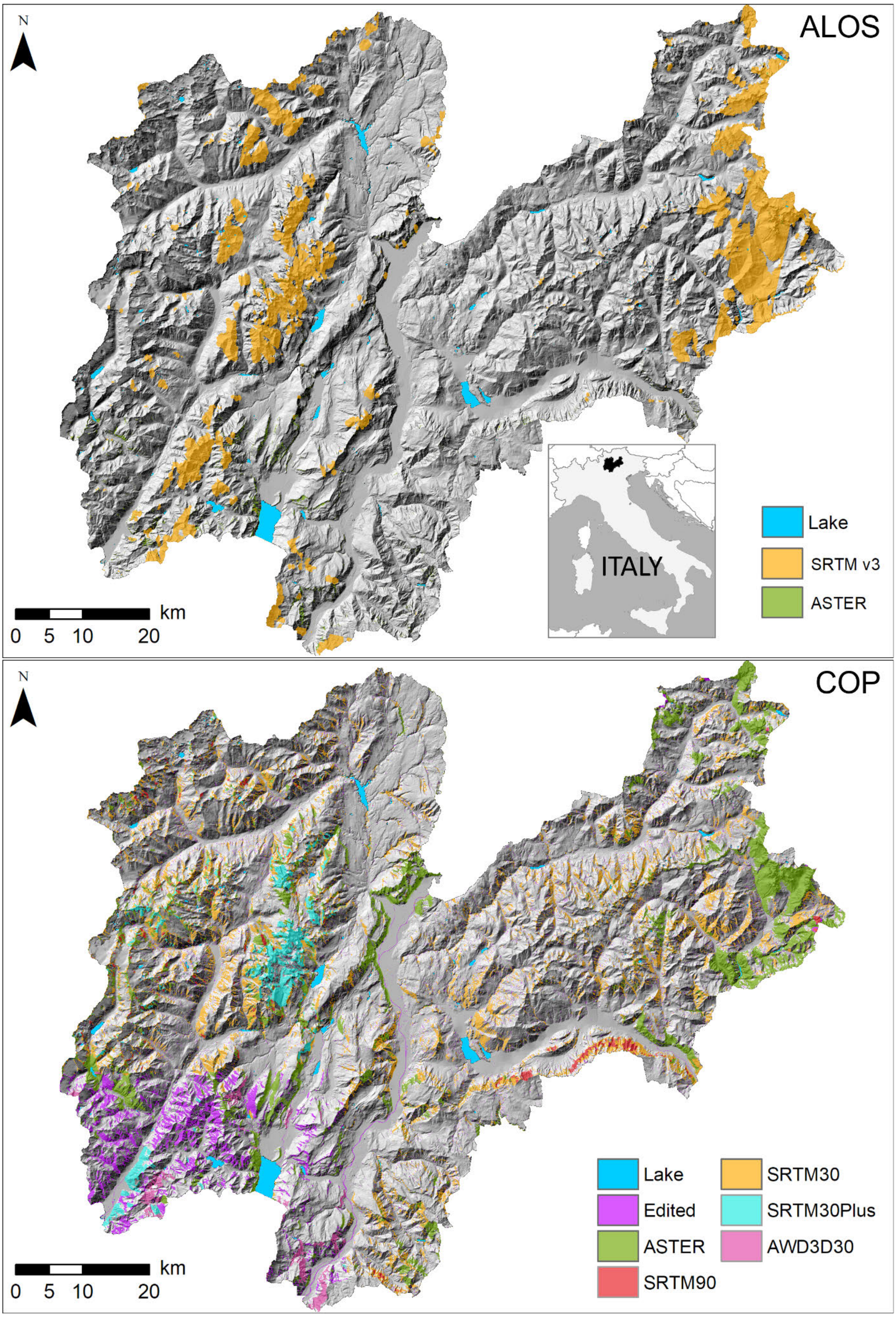
Other examples of DEM comparisons involving CopDEM can be found among the reference list presented in López-Vázquez and Ariza-López (2023), which considers an extensive review of the literature concerning the comparison of DEM products.

8.4. Considerations on the Wine Contest Results

The DEMIX intercomparison of global DEMs (Bielski et al. 2024) was performed taking into account that, to be representative, the reference DEMs must cover a wide range of landforms observed on Earth. Moreover, another premise of the approach was that the quality of the global DEMs being compared are not expected to differ across national or continental borders because the same acquisition technique and processing was used to create each of the DEMs under consideration. These requirements are to a high degree met in that study however, it should be taken into account that in specific geomorphological settings, e.g., sand dune deserts, have not been considered because of their dynamic temporal characteristics. In geomorphological dynamic settings the different acquisition times of the global DEMs and of the reference DEM should be considered especially if adopting pixel-by-pixel based metrics. Moreover, in some specific settings the quality of global DEMs can change across borders due to the data integration approaches deployed by the DEM producers. For example, CopDEM considers not only other global DEMs but also integrates national/regional DEMs that were acquired using different types of measurements. These are highlighted in the product description for Norway and Spain, (see the product release notes: https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model).

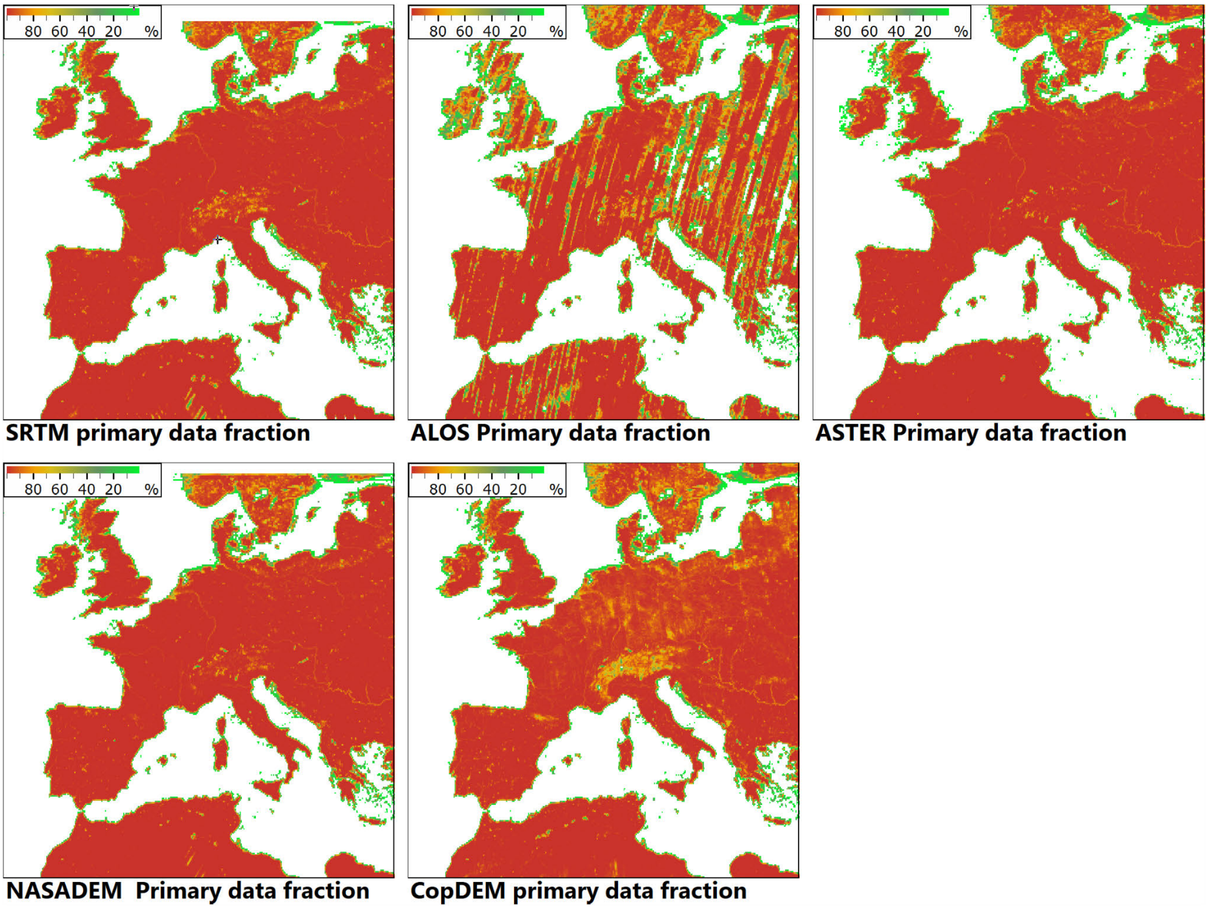
The extent of data gaps across different global DEMs and how these are filled by the producers also deserve some attention because they can come from a variety of sources. In mountain regions, global DEMs such as CopDEM and ALOS are characterised by extensive areas of high elevation with gaps filled with other DEM products such as SRTM, Aster, etc. which are older and sometimes acquired using different technologies. Accordingly, the potential impact of the integration of DEM products with different characteristics such as measurement accuracy, time of acquisition, the presence of artefacts, and others should be taken into account. **Figure 43** presents some of the gap filling introduced into CopDEM for a region in North-eastern Italy.

Figure 43. Example of extensive presence of gaps filled by other DEMs product for ALOS (V3.2) and CopDEM (Trentino Province, Italy).



Source: Sebastiano Trevisani

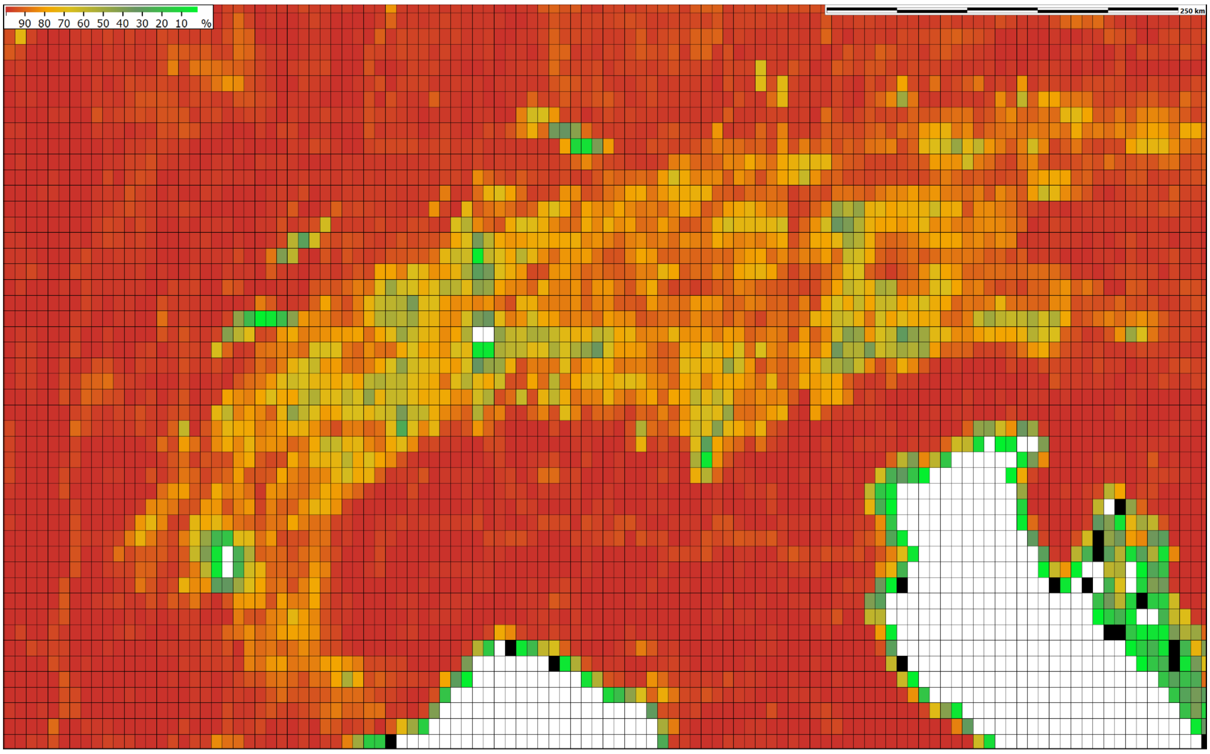
Figure 44. Showing the primary data fraction for five different global DEMs centred on Europe.



Source: Peter L. Guth

Figure 44 demonstrates the differences between the available primary data acquired to produce each of the different global DEMs that were compared. The variety observed between the products is related to the instruments used for the data acquisition. For the CopDEM example, one can see that data filling is mainly related to relief because it is concentrated across the Alps (**Figure 45**). Note that one of the original requirements for the DEM was to make sure that the data mostly came from the same instrument. However, it is evident that is not the case. However, the presented figure does not guarantee that information is available for all of the global DEMs but as is presented in the report, the Copernicus DEM product delivers high quality metadata which helps users better understand the origins of each elevation pixel. Furthermore, based on the DEMIX wine contest results, it is shown that global DEM products that are based on more primary data do not necessarily make them better quality, e.g. ASTER DEM.

Figure 45. A CopDEM example showing data filling across the Alps.



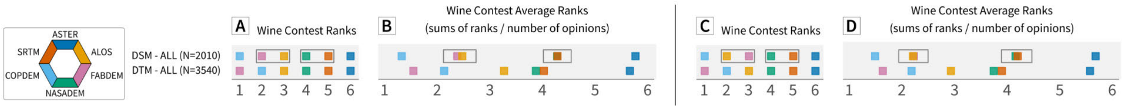
Source: Peter L. Guth

Global DEM related products are in continuous evolution and newer versions may improve their quality in specific locations and/or overall level. The presented suggestions provided in this sub-section are valid only for the versions of global DEMs considered in the DEMIX wine contest intercomparison study. When newer versions and/or DEM products will be available, the DEMIX intercomparison should be rerun in order to update the wine contest ranking based on the new data.

8.5. Sensitivity Analysis based on Changing Tolerances

The DEMIX wine contest considered many criteria and an important characteristic of a criterion is an understanding of the tolerances associated with the measurements. There are no defined tolerances related to the criterion applied in the wine contest and therefore it was possible to use the wine contest method to test different tolerances for certain criterion. The tolerances considered for the different metrics were educated guesses based on the DEM product specification and user requirements. However, by running the wine contest against different tolerances, it is possible to analyse the effect on the ranking results of some variations. This was done by performing two computations: one with initial tolerance values, and the second with those tolerance values doubled. By doubling the tolerance values, it is more likely that ties between the DEMs could occur in the opinion tables but the effect cannot be easily anticipated. The outcome of these two computations are described and presented in full in the reference (Bielski et al. 2024) from where **Figure 46** is taken in part.

Figure 46. Sensitivity analysis of the final rankings of the DEMIX wine contest based on the intercomparison of six 1" global DEMs. The different outcomes are based on 15 contest criteria. Tolerances for {A, B} are: ELVD=0.5, SLPD=0.5; RUFD=0.2. Tolerances for {C, D} are doubled. Rankings in A and C show the same winners (CopDEM and FABDEM) while only slight movements in secondary places are detected. An expanded version is available in Bielski et al. (2024).



Source: Bielsi et al, 2024

Applying the two different wine contests based on different tolerances, one can observe that the effect of significantly increasing the tolerances is rather small. The computations reported here are just a summary of what is presented in Bielski et al. (2024), where detailed results discriminated by land use class, relief, parameter, etc. are available. In a ranking problem such as this which is not continuous but discrete by nature, it is unlikely that small modifications of the parameters will lead to noticeable modifications in the ranking output. Our results support this providing some reassurance on the final ranking and the sensitivity related to the choice of tolerances which may not be known.

8.6. Use of CopDEM in ESA and Copernicus Services

The demonstrated quality of CopDEM considering the results of the DEMIX intercomparison and the additional findings of the cited studies, it is highly recommended that the CopDEM product be adopted as the basis for supporting the creation of other geospatial information layers especially those at the European scale. It is not surprising that key Copernicus Services and ESA processing workflows have already adopted CopDEM. We highlight the following products and services which are both pan-European and global: the European Ground Motion Service, the Copernicus Emergency Management Services, and the Copernicus Sentinel-2 pre-processing services.

- The European Ground Motion Service (EGMS) which is part of the Copernicus Land Monitoring Service monitors ground deformation across the EU member states based on the application of Synthetic Aperture Radar (SAR) interferometric techniques. The technique is applied using the Copernicus Sentinel-1 data and details on the product can be found here: https://land.copernicus.eu/pan-european/european-ground-motion-service . The CopDEM is integrated in the processing workflow to produce the outputs of the EGMS (Crosetto et al 2020; and ESA/Copernicus report). Due to the high quality of CopDEM in urban areas, this product was adopted for SAR processing and is particularly beneficial in the context of monitoring of the built environment and its interaction with the geo-environmental factors.

- The global monitoring workhorse Copernicus Sentinel-2 is pre-processed to level 1 and above by the ground segment under the responsibility of the European Space Agency (ESA). This optical Multi-Spectral Instrument (MSI) acquires imagery at a nominal spatial resolution of 10m since March 2021. To deliver high quality products to Sentinel-2 users, the Copernicus DEM GLO-90m product was adopted for the orthorectification process to achieve Level 1C. See https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi/definitions .

- The Copernicus Emergency Management Service (CEMS) is regularly activated around the world to help with the mapping and provision of situational awareness for both natural and anthropogenic disasters and emergencies. The emergency mapping producers depend on the CopDEM in order to deliver accurate maps for many different types of disasters including the mapping of flood disasters (https://emergency.copernicus.eu/).

The DEMIX SG2 outcomes fully support the integration and updating of Copernicus products and services based on the CopDEM in order to deliver the most accurate pan-European and global products to users and practitioners.

8.7. A Vibrant Future for Global DEMs

This extensive report has hints in many places indicating the variety of aspects and possibilities that can characterise the rankings of DEM products. While digging into the DEMIX wine contest results database it was also possible to get a glance of the fact that alternative criteria and DEM characteristics could be taken into account in performing a DEM comparison. The investigative potential of the DEMIX wine contest approach is timely considering the speed at which new DEM products can be produced today based on novel techniques which can increase day after day, both in terms of file sizes as well as in terms of their inherent complexities. This is related to both the deployment of new sensors and survey missions as well as the increasing capabilities in data analysis and modelling via geo-computational approaches including machine learning and artificial intelligence (AI). The FABDEM is a good example of this because it takes advantage of various input features, including CopDEM elevations to apply a supervised learning approach based on the Random Forest algorithm to estimate the bare earth elevations (Hawker et al. 2022). The derivation of FABDEM also includes other post-processing steps to remove artefacts such as smoothing filters. Other examples are represented by the integration of different DEM products to derive an ensemble improved DEM such as the Ensemble Digital Terrain Model (EDTM) of the world (Version 1.1) (Ho et al. 2023).

The possibility to create new DEM products integrating different informative layers and adopting post-processing algorithms also increases their complexity. For example, when applying machine learning approaches it is possible to include a variety of input features, such as elevations and geomorphometric derivatives from different DEM products, e.g., ensemble models. From this perspective, spatial heterogeneity in the quality and resolution of the input features will be reflected in the outputs, possibly with spatial characteristics that are not detectable by simple metrics of elevation errors. Furthermore, the metadata linked to such products may not indicate all the resources used turning the DEM product into a black box.

Based on the DEMIX intercomparison performed, the outcome highlighted the necessity to consider a wide set of reference data that is represented by much finer resolution and more accurate elevation data which can potentially represent both DSMs and DTMs. DEM products that cover wider regions from continental to global will continue to be produced with increasing accuracy, resolution and frequency of updating. Consequently, such DEM products should be paired with access to available high quality, finer resolution, much more accurate, and multi-temporal reference elevation data that can be used as reference. It is highly likely that an initiative to make available an open repository of reference elevation data with prescribed standards regarding the accuracy, the geoid transformation, etc. would pave the ground for more transparent quality standards in global DEM production and quality assessments.

8.8. Abandoning the EU-DEM for CopDEM

The literature comparing EU-DEM to at least some global DEMs is somewhat scarce.

In the single case found involving also CopDEM, (Kramm and Hoffmeister 2022) compared eight DEMs in two regions in Germany. In one, EU-DEM was the best, while in other it was behind SRTM (both 30 and 90 m) and NASADEM. CopDEM was worse in both sites. The comparison was in terms of RMSE. In terms of NMAD the order is reversed, and EU-DEM was below CopDEM for both sites.

Tavares da Costa, Mazzoli and Bagli (2019) compared seven DEMs (some of them global) in terms of its performance in hydrological applications in Italy. The reference was a DEM based upon LiDAR. They used RMSE, MAE and Pearson correlation coefficient as quantitative metrics, and other more related with the application. EU-DEM was worse than SRTM (both 30 and 90 m) but better than ASTER and ALOS in terms of RMSE. In terms of MAE, it was behind both SRTM and ALOS but better than ASTER.

Mouratidis and Ampatzidis (2019) compared EU-DEM to SRTM and ASTER (its predecessors) using 70.000 GNSS points as reference taken over a 7000 km2 site in Greece. They concluded that EU-DEM improved both, either using RMSE, STDE, NMAD or the maximum absolute error as the criteria. Jozsa et al. (2014) used the same datasets in a test site in Hungary. They concluded that EU-DEM was worse than SRTM but better than ASTER in terms of RMSE.

Aside from SRTM and ASTER, Florinsky et al. (2019) analysed also ALOS and EU-DEM in the Trentino region, Italy. EU-DEM and ASTER were tied in the best position according to the RMSE, but lagged behind in terms of the MAE, STDE and maximum absolute error. It is the best (but tied with ASTER) in terms of number of outliers.

Hnila and Elicker (2021) tested EU-DEM in a 1.5 km2 site in Armenia, using just 17 GNSS readings as reference. In such a limited situation, EU-DEM was the worst when compared to SRTM 30, NASADEM and ASTER either in terms of RMSE or the maximum absolute error.

Becker et al. (2017) used three metrics in a small catchment in Andalusia, Spain. He analysed Pearson correlation coefficient for both the elevation and the slope. For the elevation, all of the DEM considered behaved similarly, while for the slope he concluded that EU-DEM was below SRTM (both 30 and 90 m) but above than ASTER. The same order was concluded in terms of elevation RMSE.

A very complete work was presented by Zingaro et al. (2021) using elevation, slope and aspect as parameters of interest. Three sites and the whole Italy were considered using point elevation values as a reference. For the sites also a reference DEM was available. EU-DEM was the worst in terms of elevation RMSE and STDE for the whole of Italy. It was below SRTM30 and ALOS in terms of the STDE of slope and aspect for the three sites.

In summary, the literature shows that EU-DEM was in most cases below SRTM (both at 30 and 90 m) while slightly superior to ASTER. In the evaluation table (**Table 8**) we collected the numbers obtained in all cases and for all the metrics. Unlike the one created within DEMIX, this evaluation table is incomplete thus precluding running a wine contest given our current restrictions.

Table 8. Evaluation table for the cases found where EU-DEM was part of the comparison. Only those global DEMs considered in this report were included. This table is an example of the RICB design, when the I stand for incomplete.

			SRTM-30	SRTM-90	NASADEM	ASTER	ALOS	CopDEM	EU-DEM
Tavares et al. (2019)	EL_RMSE	IT	5.118	4.823		9.393	7.942		7.403
Mouratidis and Ampatzidis (2019)	EL_RMSE	GR		4.8		10.2			4.3
Florinsky et al. (2019)	EL_RMSE	IT	10.5			16	7.7		17.6
Jozsa et al. (2014)	EL_RMSE	HU		7.3		10.1			7.8
Kramm (2019)	EL_RMSE	DE	3.207	3.207	3.414	6.103	4.552	4.138	3
Kramm (2019)	EL_RMSE	DE	6.461	7.077	6.872	8.41	7.384	8.307	7.282
Hnila and Elicker (2021)	EL_RMSE	AM	2.77		2.36	5.82	2.64		4.54
Zingaro et al. (2021)	EL_RMSE	IT	6.5			8.9	6.5		10.1
Becker et al. (2017)	EL_RMSE	SP	4.12	4.74		8.61			7.39
Mouratidis and Ampatzidis (2019)	EL_STD	GR		4.3		7.6			3.6
Florinsky et al. (2019)	EL_STD	IT	9.2			14.9	7.5		17.6
Zingaro et al. (2021)	EL_STD	IT	6.4			8.9	6.1		10.1
Mouratidis and Ampatzidis (2019)	EL_NMAD	GR		3		6.8			2.6
Kramm (2019)	EL_NMAD	DE	1.759	1.655	1.862	4.862	2.276	1.138	1.448
Kramm (2019)	EL_NMAD	DE	6.256	6.564	6.256	8	6.153	5.231	6.564
Mouratidis and Ampatzidis (2019)	EL_Range	GR		118		88.2			86.4
Florinsky et al. (2019)	EL_Range	IT	57.3			133.9	57.2		147.6
Hnila and Elicker (2021)	EL_Range	AM	18		19	29	36		30
Tavares et al. (2019)	EL_MAE	IT	3.433	3.214		6.612	3.657		4.782
Florinsky et al. (2019)	EL_MAE	IT	7.1			9.5	5		10.4
Becker et al. (2017)	EL_R2	SP	0.99	0.99		0.99			0.99
Becker et al. (2017)	SL_R2	SP	0.76	0.65		0.64			0.64
Zingaro et al. (2021)	SL_Stdev	IT	5.7			7	4.8		6.3
Zingaro et al. (2021)	SL_Stdev	IT	5.7			7	4.8		6.3
Zingaro et al. (2021)	SL_Stdev	IT	4.4			5.3	4.2		5.5

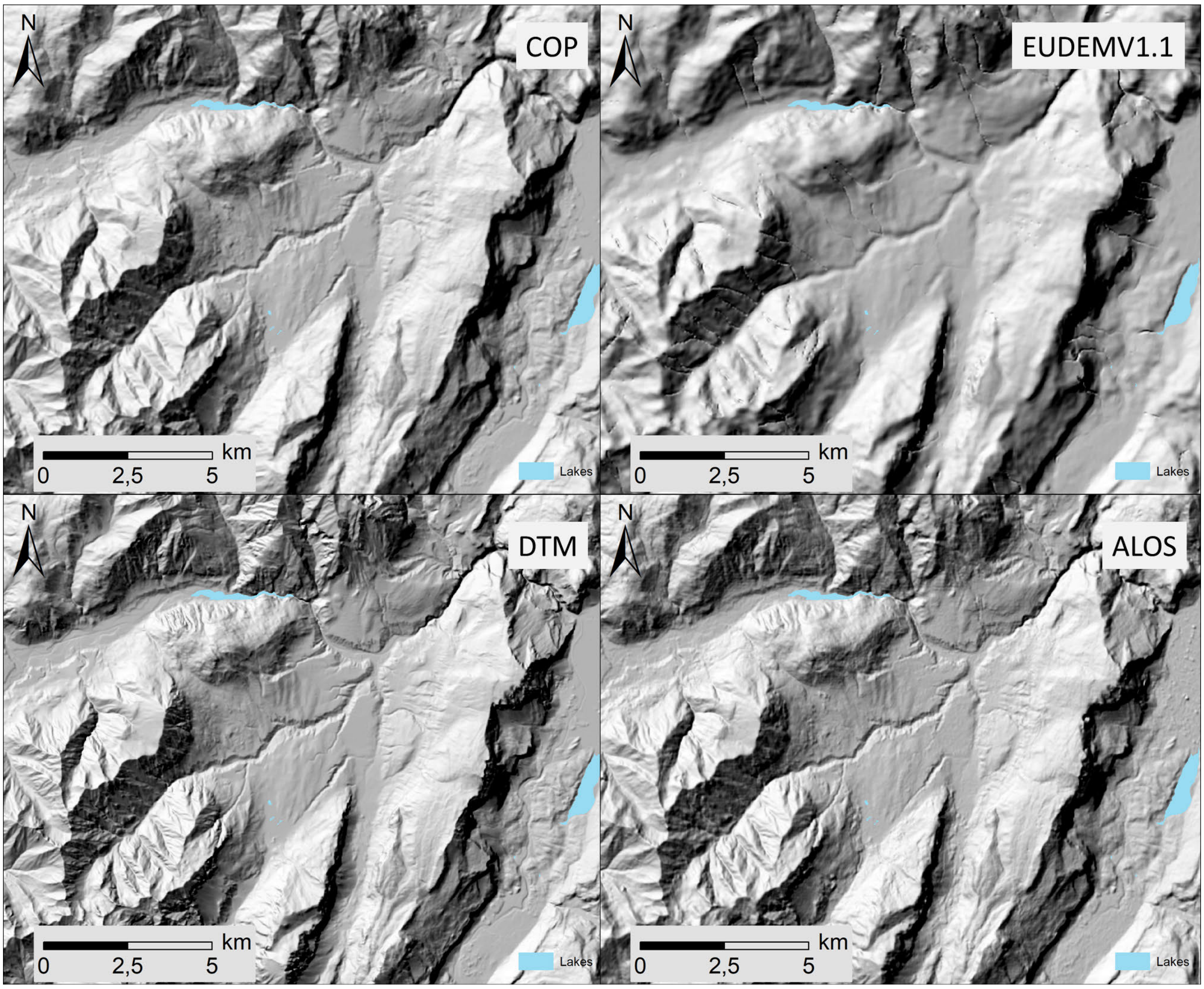
Zingaro et al. (2021)	AS_Stdev	IT	99.5			107.2	94.6		102.2
Zingaro et al. (2021)	AS_Stdev	IT	99.5			107.2	94.6		102.2
Zingaro (2021)	AS_Stdev	IT	91.4			100.5	86.1		101.6
Florinsky (2019)	#OUTL	IT	1.85			0.9	2.7		0.9

Source: own analysis

As outlined by different authors (Florinsky et al. 2019, Zingaro et al. 2021), and the VisioTerra report[2], the EU-DEM it is characterised by several artefacts, especially in the version subjected to the hydrological correction (**Figure 47**). In regard to the possibility to conduct a Wine Contest only for Europe which could include the EU-DEM product, one would expect that it would rank quite low because the EU-DEM was derived from two global DEM products which were part of the Wine Contest and ranked low overall: SRTM and ASTER. On the one hand, SRTM and ASTER DEMs resulted in the lowest ranking among the candidate Wine Contest DEMs. However, on the other hand the criteria used in the current DEMIX comparison are not specifically tailored to detect the artefacts related to the hydrological correction of the EU-DEM product. Furthermore, the EU-DEM is not a global product and therefore was not considered for this intercomparison.

2 https://visioterra.fr/telechargement/A003_VISIOTERRA_COMMUNICATION/HYP-098-VtWeb-E_CopDEM_EU-DEM_comparison.pdf

Figure 47. Hillshades for different Global DEMs and a reference Lidar-based DTM, represented in a projected system with a 25 m pixel resolution (Trentino, Italy).



Source: Sebastiano Trevisani

9. Spoilt for Choice or How to Make Informed Decisions

Section contributors: Carlos López-Vázquez, Conrad Bielski, Peter Strobl

In this section, a broader introduction into the Wine Contest is presented in order to demonstrate the wider applicability of the technique. In the context of DEMIX, it was applied to the intercomparison of global DEMs, however such a powerful technique can also be adapted to other applications where an informed choice is desired. This is especially relevant to support a data driven society where informed choices based on ranking are deemed suitable. If the quantitative and/or qualitative criteria on which the evaluations were based are open, then the Wine Contest ranking is transparent. The purpose of this section is to introduce the Wine Contest framework and all necessary terminology independent of the ultimate goal of DEM intercomparison.

9.1. A Wine Contest Framework

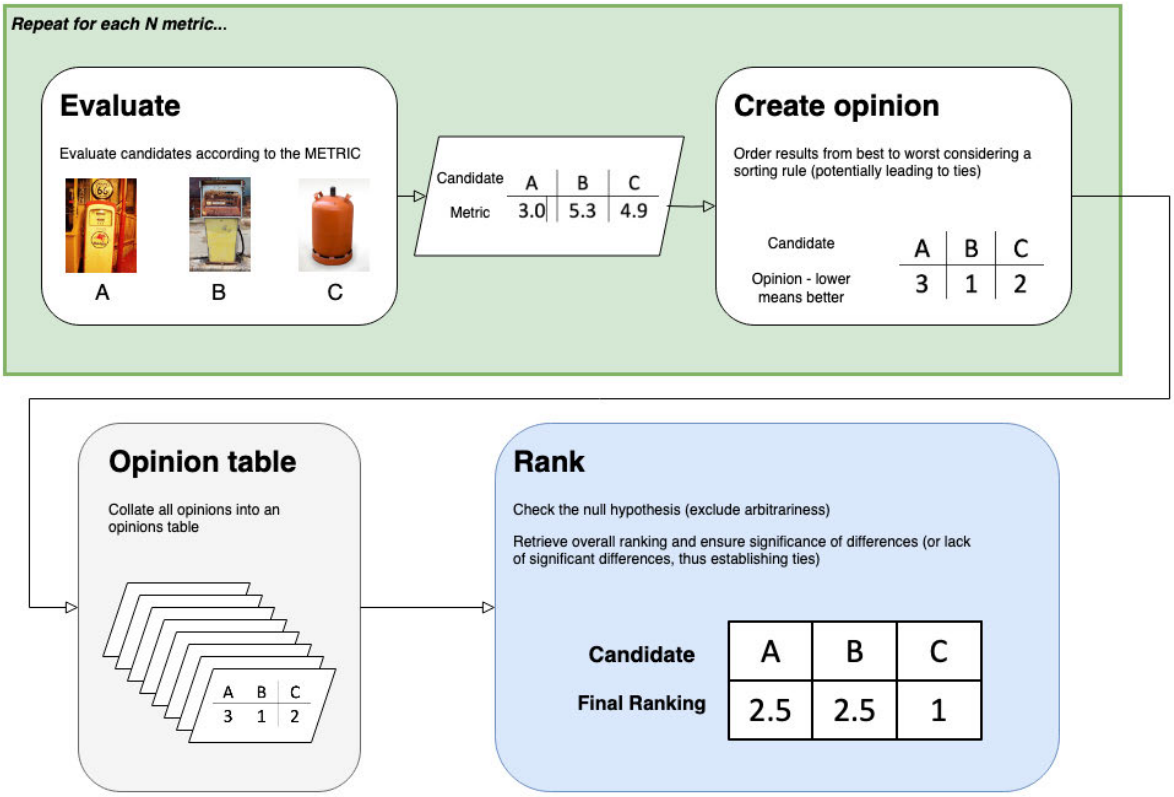
Whenever we must make a choice, the question is: which option is better or the best? From our daily lives to politics or science, we seek for a rationale to help us decide and hopefully make the best decision moving forward. However, the more options we have and the more diverse their advantages and disadvantages, the more difficult it is to come to a conclusion and trust one's judgement.

Let us begin with a simple question: if you were given a choice of wines, would you be able to tell the difference between a 'good' and a 'bad' wine? Would you go for bouquet, taste, colour, cost, or all of the above? Do you trust your own judgement, or would you rather base your choice on an expert's opinion?

The same situation occurs when one is faced with a choice of other products or services. We are interested in certain characteristics while others may not be pertinent. By being able to rank a set of choices, we can order them according to properties which we consider relevant. We do so because we are convinced that the process yields greater potential to help people decide on what they would like from a qualitative perspective and what is the most appropriate in a quantitative perspective. Think of the latest products or services where you had to make a choice. What information did you use in order to compare and choose the best for the reason at-hand?

In the course of DEMIX, this issue was tackled by combining well known and understood criteria for assessing the quality of DEMs with extensive evaluation and statistical methods in order to derive clear and substantiated answers on the question of whether and which product is better in a given examination context. This analytical framework was coined the 'wine contest' because of its similarity with the problem of evaluating wines. The 'wine contest' name is a short, intuitive, and memorable term intended to be a shorthand of 'randomised complete block design' (RCBD) which is a well-established statistical method (Siegel 1956). The whole process of an intercomparison between different choices can be envisaged as shown in the **Figure 48** and split in three different and fairly independent steps: evaluate, create opinion, and rank.

Figure 48. The main stages of the wine contest analytical framework. This figure presents the wine contest as a graphical schema with the main blocks depict creating the evaluation table, the opinion table and the final ranking.



Source: Carlos López-Vázquez

The main steps are described in the following manner:

1. **Evaluate** - A metric is applied to each candidate. For later analysis, it is critical that the metric has an informative value with respect to a specific application of the candidate. Beyond that, the metric can be anything from an ordinal label such as large, medium, small to a SI traceable physical quantity such as temperature difference in degree Celsius compared to a reference body. For the purposes of the wine contest, the only requirement is that the metric can lead to a precedence order between candidates.

The evaluation on which later an opinion is based can in principle be obtained for each wine and criterion independently and this first step neither needs to be simultaneous ('at the same time') nor complete ('for all candidates'). As happens with wines, a judge can issue an opinion only after considering simultaneously the evaluation values of all the wines. No opinion can be issued on a single wine (candidate), only an evaluation is possible.

2. **Create opinion** - It must be possible to translate differences in the metric into a sorted ordinal list, i.e. organised from 'better' to 'worse' for the given purpose. Furthermore, to complete step two and produce a valid opinion an evaluation must be available for all candidates in the set and all of them will be assigned a place on the ordinal list. A tie should be declared when the difference or dissimilarity between the candidates is not significant according to the chosen metric.

Before moving to the final step, let us consider the combination of steps 1 and 2 with respect to what wine contest judges would do by first tasting (evaluating) a wine and then creating an opinion which one is better or worse. Several judges can come up with several opinions, organised in a table with an opinion per row. For example, if a set of wines is assessed according to bouquet, colour, and other attributes (criteria), these opinions may diverge among the judges. In the current analogy, a wine contest judge is an instance that creates an opinion for every candidate wine having all of them in mind (at the same time). In statistical parlance, this is "a ranking for each block". For convenience, we will denote as "sorting" the procedure to assign an ordinal list place for each block and keep "ranking" for the process to be carried out over the full opinion table.

3. **Rank** - Each single opinion already yields a hint about which candidate is better or worse given one aspect (criterion). The ability to sort out a multitude of potentially disagreeing and maybe even opposing opinions is the core of the wine contest method or RCBD design. This final step has several important purposes from a statistical perspective:

    a) Provide statistical proof/support that the wine contest opinions are not simply random but converge;

    b) Provide an overall ranking taking into consideration all the selected opinions;

    c) Assess the existence of ties between candidates, identified as those who do not behave significantly different.

An opinion is essentially a vector of numbers and in the case of the wine contest, the number value for each wine depends (among other things) on the relative position of the wine among the candidate set. Each judge must produce a single opinion for the given set of wines. Different judges may consider or weigh differently a number of properties of the wine. While the single opinions come without a confidence level and are therefore neither stable nor reliable, the outcome of the ranking allows to provide statements such as "…candidate A is better than the others given the X% confidence level…" or "… candidate A and candidate B are better than the others, but their difference in ranking is not statistically significant at the X% confidence level…". Such statistical support is needed to understand whether differences in the ranking of candidates based on a set of chosen opinions is random or not. This is especially important if there are a large number of opinions on which to base the final ranking.

9.2. Necessary Assumptions for Applying the Wine Contest

The following are some assumptions that the user of the wine contest should take into account when considering applying the framework or using the outcome.

Assumption 1: Winner takes all! – No mixing among candidates is considered, therefore it is not possible to create a blended wine with the best performers.

Assumption 2: Every opinion receives the same weight – Essentially, there are no opinions which are better or worse than another.

Assumption 3: The judges' opinions are independent - The answer of one judge is assumed to be independent of those from others. More formally: the opinions appearing in a given block are independent of the opinions appearing in each of the other blocks (Daniel and Cross 2013]. Such a statement underlines the fact that independence between the judges does not potentially preclude correlation between the opinions because different judges can still produce the same opinions.

These assumptions are very important to consider especially when understanding the implications related to the final ranking.

9.3. The Randomised Complete Block Design (RCBD)

Given a set of candidates, selecting the best wine, choosing the best insecticide or selecting the best DEM leads to a mathematically similar problem. We continue the development of the wine contest framework with an example focusing on insecticides in order to better illustrate specific aspects of the procedure. For example, in four different agricultural fields (named Blocks) five different insecticides are tested each in small parcels. One possible metric for the evaluation of efficiency as a possible criterion takes the form of a count of insects surviving in a parcel after applying each insecticide to one of it. The blocks might differ in their orientation, slope, etc., so there is no reason to expect the exact same efficiency of the insecticides in all of them. However, it is assumed that the similarity of parcels within one field (intra-block) is larger than that across different fields (extra-block). The physical location of parcels for testing within each field is selected at random. If all insecticides are tested in every field, the procedure is named randomised complete block design Daniel and Cross (2013).

This example based on testing insecticides was chosen because it simplifies the introduction of the word 'block' within the wine contest framework which is not obvious when dealing with wines. The origin of the terminology can be traced back to the 1920's with Fisher's work at an agricultural research centre where the connection to 'blocks' is linked to agricultural parcels (Parolini 2015). In a wine contest, each judge tastes every wine in the set. Consequently, associating the wine contest analogy to the RCBD:

- A wine contest judge is equivalent to a block;
- The term complete relates to the fact that the judge tastes every wine in the set;
- Note that judges are presented the wines to taste in random order without labels and this is linked to the term randomised.

Consequently, there is no reason to expect that every judge will produce neither the same ranking, nor that for every field the insecticide performance ranking will be the same.

To apply the RCBD, one must first want to make sure that the wines are indeed distinguishably different, so that the so-called null hypothesis is that they are equivalent and we have to find a way to prove that this is wrong. There are statistical methods that are able to test whether the 'null hypotheses' of candidates' equality rejected or not. Some operate directly on a measurable quantity, such as the wine alcohol content, assuming some underlying distribution or statistical property of such values. One important class of statistical tools are named parametric methods, which are very efficient but require equal variances of the alcohol content among the wines, normal distribution, etc. In our settings, we cannot expect in general that the results arising from any metric belong to a normal distribution, or that their variances are the same. Moreover, one might want to consider not only the alcohol content, but the taste as well. Consequently, two different metrics will lead to different statistics which are not amenable to unified treatment. Fortunately, non-parametric alternatives are available which operate not on the statistics of the measurements or counts but on the precedence order resulting from measurements or counts. This precedence ordering is denoted within the wine contest as an opinion.

A well-known non-parametric test which is applicable to the wine contest context is the Friedman Test (Friedman 1937) despite the fact that there are other choices available (Iman and Davenport 1980). However, according to Fey and Clarke (2012), the Friedman method is the best alternative for minimising the risk of paradoxical results among the non-parametric tests suitable for this problem. A paradoxical result means that the transitivity property is not guaranteed (see box 1 for an example). To illustrate this, the paradox arises if you apply the procedure to {A, B, C, D} and conclude that A is better than B, while if the same procedure is applied to {A, B, C} concludes that A cannot be proved to be better than B, i.e. the conclusion changes. The inclusion of D in the set changes your conclusion about the relation A vs. B, thus leading to a paradox.

Box 2. Paradox Example from Fey and Clarke (2012)

> In the author's own words (see up to the end):
>
> "...Examples of the paradoxes that can arise through the use of rank-based procedures have been identified by several authors in the statistical literature; our review is correspondingly brief.
>
> As an example of the inconsistency that can result when treatments are ranked jointly or within blocks, consider the data on the effectiveness of hypnosis given in Table 1. The emotions of fear, happiness, depression, and calmness were requested (in random order) from each of eight subjects during hypnosis (Lehmann 2006, p. 264).
>
> Given that the data are blocked, a Friedman-type multiple comparison is appropriate (Nemenyi 1963). The results of the multiple comparison procedure are given in Table 2. When all the four treatments are compared (columns 2 and 3), none of the contrasts are statistically significant. When treatment 4 (calmness) is dropped (columns 4 and 5), however, the difference between treatment 1 (fear) and treatment 3 (depression) is statistically significant. That is, the statistical finding that the fear outcome differs from the depression outcome depends on whether or not the analyst includes the calmness treatment in his or her analysis...."
>
> In a pairwise comparison, the paradox is that A<B and B<C does not forbid that C<A. The transitivity property is not guaranteed.

The term *randomised* is significant when the criteria are subjective because the order might influence the evaluation and thus affect the resulting opinion. When the metrics are objective, it does not have any consequence.

The term *complete* implies that for each criterion, every candidate under consideration will have an entry in the corresponding row of the evaluation table. Note that many measures taken into consideration do not achieve this requirement. For example, error statistics that tend towards zero (better) are not applicable directly in this context because there is no way to differentiate whether a measure of -1 is better than a measure than +1. Such results must be processed with a sorting rule and thus stored in the opinion table. Only after the opinion table is ready the ranking procedure of the wine contest can be applied.

The chosen analogy to a real wine contest may not be too strict. For example, we are aware that judges have some rules, they qualify some prescribed aspects of the wine, assign points, and afterwards reach an opinion. In our setting, each judge is viewed as a black box that receives the set of wines and produces an opinion. The evaluation table need not to be disclosed. The opinion might be arbitrary, based upon just one aspect of the wine, based upon many, based upon either qualitative or quantitative aspects, etc. but at the end, it is just an ordered list of wines. Deliberately ignoring the mechanism to produce an opinion will let us consider subjective and quantitative opinions as well without violating the assumptions of the test.

9.4. The Friedman Test

As presented in the previous section 4.4, the Friedman Test is applied as part of the wine contest. The Friedman Test just analyses the null assumption that all the wines, insecticides or DEM are equivalent. It operates over the opinions, organised as a table of k columns (wines) and N rows (number of opinions). The test analyses the average opinions considered by columns and offers a statistic which is large if there are differences between the wines. In principle, the best wine will have the lowest average ranking provided the lower value of the opinions stands for the better wines. For example, a ranking of 1 = best whereas a ranking of 6=worst out of six wines.

The Friedman Test presented in most student textbooks and handbooks only describe the case without ties. A tie within an ordered list can arise for different reasons. Taking advantage of the wine contest analogy, two different wines may have the same alcohol content which obviously would be considered a tie. From the DEMIX perspective, if the DEM opinions arise as a mere ordering of a real valued criteria such as the RMSE, then there is little chance to observe ties. However, even for real numbers it is not realistic to consider two values as being different if their difference is below a chosen tolerance. For example, if a tolerance of 2.0 is set, then real number values of 14.04 and 15.12 imply a tie because the difference is less than the tolerance. Other more subjective criteria may also lead to ties and therefore it is reasonable to consider in more detail such cases.

In the original version of the Friedman Test (Friedman 1937), ties within a block are not allowed. The extension to consider ties was proposed for the first time by Conover (1980) with a somewhat complex formula. Other simpler but equivalent expressions appeared later in Siegel and Castellan (1988) and Corder and Foreman (2009) which apparently were derived independently despite the fact that they produce exactly the same value. The general Friedman´s statistic, valid with or without ties, is presented in **Equation 3** where N=number of opinions and k=number of candidates.

Equation 3. The general Friedman´s statistic.

$$\chi_r = \frac{N(k-1)\left[\sum_{j=1}^{k} \frac{R_j^2}{N} - C_F\right]}{\sum_{i=1}^{N}\sum_{j=1}^{k} r_{ij}^2 - C_F}; C_F = \frac{Nk(k+1)^2}{4}$$

being R_j the sum of the opinions r_{ij} over the rows i. For large k or N the statistic behaves like a χ^2 distribution of k-1 degrees of freedom. For large k and small N it behaves like a normal distribution. For low k and N values there exist tables holding the critical values to compare. The first set of such tables for the case without ties were proposed by Friedman (1937) while the most comprehensive one has been presented by López-Vázquez and Hochsztain (2019). It covers all the cases where the difference of the critical values with respect to those of the asymptotic distribution exceeds 1%. Otherwise, the user may consider safely the asymptotic estimate.

The above-mentioned tables can be used provided there will be no ties in each block. This fact is dependent upon the characteristics of the metric used to establish the ranking. If there is no expectation that a tie may occur, the reported tables could be used. Otherwise, specialised tables for the case of ties should be used instead. A small set of exact tables for the case of ties has been recently presented in López-Vázquez, Tasistro and Hochsztain (2021), while a larger unpublished table is provided by López-Vázquez and Hochsztain (2020). The latter table covers, as before, all the cases where the difference of the critical values with respect to those of the asymptotic distribution exceeds 1%.

Box 3. Choice of Statistical Test

For example, if the benefits of one insecticide are measured just based on the insect counts, all entries in the evaluation table are homogeneous and tests like ANOVA (Daniel and Cross 2013) or the one due to Quade (1979) might be applied. They operate over the evaluation table. But we might also want to consider as a figure of merit a secondary effect such as the average weight of the insects, which does not have the same units but is suitable to produce an opinion. If we plan to consider both alternatives, we have to disregard (Quade 1979) proposal and go back to Friedman´s test or another non-parametric test (Toothakere and Newman 1994), which does not expect that the original results are commensurate and instead operates over opinions. As will be argued later, that will be the case in DEM comparison so we will concentrate on Friedman´s test.

9.5. Post-hoc Analysis

A rejection of the null hypothesis based on the Friedman Test is not sufficient to identify which candidate is the winner. It only justifies that there is a difference between the candidates. Therefore, it is still necessary to create a ranking among the candidates. This is done using the so-called post-hoc analysis.

Given a wine contest, there are two possible approaches to identifying a winner which according to Fey and Clarke (2012), are not without problems:

Approach 1 - In this case, the wine contest opinions are compared pairwise: wine i vs. wine j and ignoring the rest. The authors report that there is a risk to fall into a cycle: if wine i is better than j, and j is better than k, despite intuitive there is no guarantee by the test that wine i will be better than wine k. This paradox is known as the Cycling Problem.

Approach 2 - This approach considers a ranking within blocks of all wines. The conclusions thus depend not only on wines i and j, but also on the presence of other alternatives. This might lead to a case where two candidates are different if a third one is present but show no difference if the third one is removed. This paradox is known as the Problem of Irrelevant Alternatives.

Based on the scientific literature, Approach 1 is discarded in favour of Approach 2 due to its low power (see info box 3). For Approach 2, there are a number of possible alternatives including: Bonferroni-Dunn (Dunn 1961); Nemeyi (Nemenyi 1963), Holm (Holm 1979), Hochberg (Hochberg 1988), and Hommel (Hommel 1988), among others. As the Friedman test itself most of them operate over the average rankings, a fact which has been recently questioned (Benavoli et al. 2016) promoting the use of pairwise comparisons. Instead, the authors suggest use of either the sign test or the Wilcoxon signed-rank test for pairwise comparison. Such an option can be accepted without changing the nature of the wine contest. In our implementation we choose the Bonferroni-Dunn following Pereira et al. (2014). However, for completeness, we add a brief description of the other alternatives below.

The Wilcoxon Signed Test analyses if two columns of the table of rankings are due to the same distribution. If the null hypothesis is rejected, the smallest rank sum will identify the best wine. Otherwise, no significant difference can be declared, and the two wines are deemed equivalent. The steps required to compute the test are the following:

a) Compute the paired differences between the columns;

b) Discard those which are zero, keeping n pairs out of N;

c) Take the absolute value and sort the differences with mid-rank criteria;

d) Add the positive differences weighted by its position, and do the same with the negative differences;

e) Define T as the minimum absolute value of both;

f) Compare T using tables for a given confidence value and n, and reject the null hypothesis if T is lower.

The procedure is described in detail in textbooks such as Corder and Foreman (2014).

The Sign Test is equivalent to the Friedman test when k=2, and according to Benavoli et al. (2016) it is more robust. It does not require that the observations are identically distributed. The drawback is that the test has lower power (see info box 4). The standard Sign Test ignores ties, if there are many. However, Bian (Bian et al. 2011) have developed an alternative specially designed for cases with many ties. For the case of the Wilcoxon Signed-Rank Test, Pratt (Pratt and Gibbons 1981) have also considered the case with ties and proposed the Signed-Rank Zero Procedure. In addition, if ties are allowed the post-hoc analysis test must deal with them. Note that textbooks do not consider such a situation in detail as part of the Friedman Test because the latter usually has been restricted to a no-tie situation.

Box 4. What is the Power?

When you classify a DEM as a winner you might commit two different classes of errors:

If DEM A is indeed the winner, classifying as such DEM B implies a Type I error (named alpha, for short). It is the confidence level.

If DEM A is not indeed the winner, classifying it as such implies a Type II error (named beta, for short). The power is exactly (1-beta). It depends on the value of alpha, the sample size, and properties of the test. Alpha is chosen, but beta is usually just blindly accepted. If it is not small enough the remedy is to increase the sample size. Unfortunately, without making strong assumptions the sample size is difficult to estimate so it is just ignored in many papers.

9.6. A Numerical Example

The wine contest is applied to another domain to demonstrate the flexibility and provide more examples of its application. From a numerical perspective, the wine contest first builds an evaluations table of k columns and N rows recording the different metric values for each of the candidates and an opinions table, which through a sorting rule translates the evaluations table to ranked opinions.

The following hypothetical example of these wine contest tables is presented in Figure 48. The data used was extracted in part from Szumska et al. (2019) and modified to fit our needs. The data reports some emission properties of three different fuels: gasoline, diesel and Liquid Propane Gas (LPG). All three types of fuel are of comparable cost of extraction, insurance and maintenance. The tolerances and subjective assessments were arbitrarily chosen to illustrate wine contest features.

Figure 49. A wine contest example applied to the intercomparison of fuels. The evaluations table (left side) records the assessment outcomes for each of the candidate fuels. The opinions table (right side) translates the evaluations table to ranked opinions using the mid-rank procedure for ties. In this case, tolerances have been established as identical for the three fuels.

	Evaluations				Opinions		
Assessments	Gasoline	Diesel	LPG		Gasoline	Diesel	LPG
Long term resource abundance (opinion by Expert 1)	Small	Medium	Large		3	2	1
Environmental considerations (opinion by Expert 2)	Fair	Fair	Very good		2.5	2.5	1
CO (tol=10.000)	2983.372	2922.011	2337.609		3	2	1
NOx (tol=0.700)	128.592	161.177	128.942		1.5	3	1.5
PM10	[72.801, 72.901]	[72.686, 72.786]	[7.235, 7.335]		3	2	1
PM2.5	[16.557, 16.657]	[16.525, 16.625]	[1.611, 1.711]		2.5	2.5	1
VOC (tol=20.000)	285.093	169.847	105.285		3	2	1
SOx (tol=0.10)	2.69	2.781	1.614		2	3	1
				R_i	20.5	19	8.5

Source: Carlos López-Vázquez

The evaluation table illustrates some alternatives. The first two rows hold categorical ordinal answers, which with a sorting rule easily leads to the corresponding rows of the opinions table. The next two rows hold real valued entries, where a general tolerance is set. In one case this leads to ties. Two rows hold interval values (probably with a confidence level attached) which is another equivalent way of presenting uncertain data. The length of the intervals might be interpreted as a tolerance, which might be different for each entry. The last two rows have also real valued entries. Notice that the magnitude and the units of the evaluation table are completely different, thus precluding all possible parametric methods. The entries in the opinions table (**Figure 49**, right side) are denoted as r_{ij}, and the final row presents the column sums by R_j. The remaining values to compute the Friedman statistic (x_r^2) can be extracted from the same table: with C_F=96, sum of r_{ij}^2=110.5, and sum of R_j^2=853.5, then x_r^2 =11.793.

For a given k, N and confidence level alpha, Friedman's statistical value x_r^2 is compared to the critical value; if x_r^2 is larger, the null hypothesis is rejected, and the conclusion is that we cannot accept that the fuels are equivalent. Since we are willing to accept ties, the standard critical values tables for the Friedman Test are not suitable. From the table for k=3 provided by López-Vázquez et al. (2021), and at the 95% confidence level, the row N=8 (the number of opinions) offers a critical value to compare of x_{crit}^2=5.793. The null hypothesis is that the opinions table entries are purely at random, implying that the fuels are equivalent. The x_{crit}^2 value from our case is larger than the x_r^2 value read from the table, so the conclusion is that we should reject the null hypothesis. If the x_{crit}^2 value is not larger, then the conclusion is that we cannot reject the null hypothesis on grounds of the available data.

Based on this numerical example, the null hypothesis has been rejected according to the Friedman test, so there are statistically significant differences between the set of fuels under consideration. The lower the values of R_j, the better, but it is still necessary to assess whether the difference between the pairs of ranked candidate fuels are statistically significant or otherwise conclude that the pair is tied. This process is denoted as post-hoc analysis and there exist different options to carry out such an analysis. In this case and following Pereira et al. (2014), we propose to use the test by (Dunn 1961) applying the Bonferroni correction.

A pair of candidate fuels {i,j} is considered significantly different if

Equation 4.

$$\left|R_{.i} - R_{.j}\right| > z_{1-\frac{\alpha}{k(k-1)}} \sqrt{\frac{Nk(k+1)}{6}}$$

In this example, the critical value for the post-hoc analysis at the 95% confidence level will be 4.005. To conclude whether there is a significant difference between ranked pairs the absolute difference between R_i and R_j should be greater than 4.005:

Gasoline vs. Diesel: abs(20.5-19.0) = 1.5 → not significant

Gasoline vs. LPG: abs(20.5-8.5) = 12.0 → significant

Diesel vs. LPG: abs(19.0-8.5) = 10.5 → significant

From a prospective fuel user perspective, the above wine contest results say that for the given assessment, LPG is the clear winner, while either Gasoline or Diesel could be equally recommended because there is no significant difference between them. The conclusion is valid at the 95% confidence level.

10. Additional Thoughts on the DEMIX Wine Contest

Section contributors: Carlos López-Vázquez, Conrad Bielski, Peter Strobl

As presented in the in section 9 above, the generic wine contest framework is quite extensive and based on a significant body of statistical work. This section aims at discussing further aspects of the wine contest methodology which are of specific relevance when comparing DEMs and thus to DEMIX. While some of them have been implemented and the results are presented in section 7, others were not considered due to a number of reasons including the requirement for extensive reference data, further software development, or even basic research before a use case could be presented. Since this discussion is DEMIX oriented, it will not use generic examples as in section 9.1 but focus on application to DEMs.

10.1. A Short Discussion on the Topic of Geomorphometric Parameters

Depending on DEM users' needs, some will be more interested in a ranking based upon elevation accuracy only while others may want to include DEM derived parameters. At the pixel level, the most popular DEM derived parameters are slope and aspect (Courty et al. 2019, Zingaro et al. 2021, Nadi et al. 2020), but there are many more. Such parameters are likely to be computed from elevations and not measured directly, although in some cases there might be suitable field instruments (clinometers, Abney level, etc. as described in Goudie (1990) or Young et al. (1974). To illustrate just one more case of derived parameters we used roughness. The roughness parameter used previously in DEMIX is based on the standard deviation of slope and it is one among many possible roughness indices. The choice of this index in the demonstrator experiment (Bielski et al. 2024) is related to different motivations: the algorithm is generally implemented both for geographical as well as projected coordinate systems; it is relatively well known; and it is easy to understand. However, it is a very specific roughness metric, aiming to describe roughness in the direction of the gradient. Other and more informative indices can be used, such as geostatistical based roughness indices (Trevisani et al. 2023) and popular indices such as the standard deviation of residual DEMs (TPI) or the dispersion of normal vectors (Trevisani and Guth 2025). Some other popular indices, such as topographic ruggedness indices (TRI) should instead not be used, because they provide a biased evaluation of roughness (e.g., highly correlated with slope (Trevisani and Guth 2025). For the comparative analysis of DEMs, geostatistic approaches are quite versatile because they permit to focus on specific aspects of surface roughness, including anisotropy, and could provide the basis for very specific quality metrics, including the potential identification of artefacts.

10.2. A Short Discussion on the Topic of Criteria

We consider that a criterion is equal to a metric plus a sorting rule. Without losing generality, we will require that the sorting rule produces a set of values where lower is better. Such a property is satisfied by mere sorting any mathematical norms or seminorms such as the RMSE, STD, Range, LE95, etc. But there are other metrics where the expectation goes in the opposite direction such as DEM completeness which can be presented as a percentage. DEM completeness refers to the areal coverage of a DEM dataset with respect to the user requirements. Therefore, as completeness metric increases, the percentage number increases and therefore in this case "higher is better". In cases like this, the requested behaviour can be imposed as part of the sorting by just changing its sign, or by defining some sort of voidness metric (the opposite of completeness). If the existence of a suitable sorting rule is in doubt, the metric cannot produce a ranking and thus must not be included in the contest. One example is the mean error which is signed and where "best is closest towards zero”. In most applications, a large negative error is no better or less harmful than a large positive one. Thus, despite sorting is mathematically possible the resulting precedence is not meaningful or relevant.

Metrics related to DEMs can be classified into three classes:

1. Vertical metrics - these metrics focus on the z-component data in DEMs neglecting positional (X,Y) errors;
2. Horizontal metrics - these metrics relate to the position of the x,y- components of the DEM where the vertical z-component is measured;
3. Functional metrics - these metrics compute a value based on the DEM data, thus applying some mathematical model as well as potentially extra data.

Most of the criteria found in the literature are of type vertical, (such as RMSE, LE90, MAD, etc.) either for the elevation itself or for derived values (slope, aspect, etc.) compared to reference values. Functional metrics might be closer to user needs. They are illustrated by landslide susceptibility (Brock et al. 2020, Rabby et al. 2020, Watson et al. 2019), erosion (Fijałkowska 2021, Romshoo et al. 2021, Zhidkin et al. 2021), etc., all of them involving extra data and a mathematical model. In the first case, the metric might be expressed as f(DEM-R), while the latter can be f(g(DEM)-g(R)) where R is a generic name for the reference data used. An example of function f might be RMSE, while an example for function g might be landslide susceptibility. Quantitative metrics first produce a numerical value and then, after applying the sorting rule, an opinion can be obtained. We will coin the term ‘qualitative metric’ to denote those metric types that produce an ordinal variable, and not an intermediate numerical value. Such cases are valid because the sorting rule require that only the relation “>” between objects is defined.

In general, metrics can produce numerical values (natural, integer, real), intervals, as well as ordinal categorical. The single requirement is that they can be ordered by applying some sorting rule. By convention the sorting rule should attach the lower scores to the better values and decide about the ties. For accuracy values, the natural order is an obvious choice. If there exist a tolerance, or the values are intervals, we should declare a tie when the intersection of the intervals is not empty. In addition, the sorting rule might also consider application specific tolerances. For example, if the metric is the familiar R squared ("linear correlation between data and the reference"), anything below 0.5 could be considered equally bad and thus ties might appear. It will be convenient to distinguish between an evaluation table holding the numerical metric values, and an opinions table holding the rankings. The elements of the evaluation table are for example the RMSE values computed for a particular site and for a particular candidate DEM. Their values are not related to the choice of the set of k candidate DEMs and thus the computations can be performed at any time. However, in order to apply the sorting rule and creating an opinion the evaluation table must be complete with respect to the chosen set of candidates.

10.2.1. A Short Discussion on Metrics Uncertainty

The elevation values associated with a particular DEM are uncertain therefore any metric value based on them will also be uncertain. For quantitative metrics, i.e. those with a real value output, ties are unlikely but not impossible because the outcomes need to be exact. When uncertainty is considered, a confidence interval (CI) for a given statistical confidence level can be established. Intervals as CI can be sorted, but not without potential difficulties. The sorting rule can establish an order if the CI are disjoint; otherwise, a tie will be declared. In this context, a tolerance can be identified as the length of a CI for the metric of interest, assuming that such an interval is centred on the metric value. Therefore, establishing a tolerance or a CI are somewhat equivalent alternatives to communicate uncertainty in the elevation values.

If a given tolerance is too large, then it is possible that multiple DEMs will appear at the same ranking position. For example, if the sorted metric values are the following set: {9.6, 10.5, 10.8, 11.9, 20.0, 25.0} and the tolerance was set to 1.0, the associated rankings would be: {2, 2, 2, 4, 5, 6} because the difference with respect to the immediate value is less than the tolerance. Following Friedman (1937), the ranking algorithm to be used here should be the mid-rank method (Amerise and Tarsitano, 2015) which deals with ties while keeping constant the sum of the resulting rankings. The process is not without potential paradoxes. Notice that in the example above, a tie was declared between the first three values {9.6, 10.5, 10.8}, despite the fact that the difference between the first value 9.6 and the last value of the tied set 10.8 is larger than the tolerance.

In the DEMIX context, three potential alternatives are foreseen for establishing a tolerance. On the one hand, one can just ignore uncertainty effects by setting the tolerance to 0.0. This could be chosen as the default choice especially if such information is not available, however this could result in differences where in fact what was ranked was largely noise (i.e. random). A second option would be to consider experts' opinions about reasonable values for the tolerances. The CI interval will be centred in the metric value, and the tolerance will define the length of the CI. This is the approach chosen for the wine contest described in section 4 . A third option would require further research in order to establish a propagation relationship between elevation and metric uncertainties and then derive reliable tolerance values. However, such values might be different for each DEM. An example for the case of a DEM was developed by Nadi et al. (2020) where the uncertainty of the elevation RMSE was analysed as a function of the number of ground control points available, discriminating by DEM and terrain type. Their assumption was that the error population was normal but contaminated with outliers. They also considered the case of uncertainty of both slope and aspect, thus illustrating the issues involved in elevation derived parameters. In a different study, similar requirements were analysed in a seismic context by Shafique and van der Meijde (2015). The authors used a Monte Carlo approach to estimate the uncertainty of the slope, aspect and the topographic aggravation factor (TAF) metric, again discriminating by DEM product.

10.2.2. Some Further Thoughts about Tolerances

When using numerical values from the evaluations table to build the opinions table it is reasonable to discuss whether their differences justify a distinction in the ranking. For each metric there will be some tolerance threshold to declare a tie depending on whether the difference is within or beyond that tolerance. Ideally tolerances are related to the uncertainty quantification of the magnitude under consideration. Uncertainty analysis is quite established in application areas such as hydrogeology, meteorology, etc. to discern the relative impact on the result of accuracy improvements on input datasets used by numerical models. However, providing uncertainty, especially at pixel level, is still not widely established for geospatial data.

Let's again consider the RMSE as an example. The uncertainty of the RMSE can be related to the individual uncertainty of the elevation errors, its location in space, and maybe other characteristics of the problem. The uncertainty in the RMSE can be analysed stochastically, for example using geostatistics, or deterministically, for example through functional analysis based on Taylor expansions. Formulae for the propagation of point error bounds into the RMSE can be in principle derived (see Nadi et al. 2020) and the same for MAE, LE95, etc. provided certain assumptions are applicable. Presently some guidance exists to use geostatistics to estimate the elevation error field given the error at selected samples which are sparsely located or available over a regular grid (Fatemeh and Mariethoz 2021).

Similar procedures can be extended to other point valued spatial variables such as slope, aspect, roughness, etc. once estimates of the elevation uncertainty are available. During the DEMIX activities, López-Vázquez (2022) derived deterministic uncertainty bounds for slope and aspect at the pixel level as a function of the pixel size. This work is paving the way to treat those variables with the same tools already envisioned for elevation. In the derivation of such estimates, he has neglected the uncertainty of the elevations themselves and considered only the effect of discrete sampling of the topographic surface. That is clearly a limitation.

In **Equation 5** to **Equation 8** , we present numerical estimates of the squared slope, aspect and squared roughness as a sum of two terms. The one at the left-hand side of the equation is computable (i.e. we have access to its numerical values) but the right hand side is a truncated infinite series. First term is the exact, unknown value, which we want to estimate. The following one (named principal part) involves higher order derivatives of the topographical surface multiplied by a low power of h of known exponent. In López-Vázquez (2022) it has been shown how to exploit such exponent knowledge in order to estimate numerically the principal part of the corrective term using Richardson extrapolation. Thus, we can use that result to create not just an estimate of the true value (S^2, A or R^2) but also an interval for it. Unlike a confidence interval, in this case it is a deterministic interval (i.e. with confidence level 100%). The formula of the squared roughness R^2/h^2 was limited to the case of a 5x5 window. The case of 7x7 has yet to be derived.

In the following expressions uppercase Latin letters are reserved for the computable estimates (S^2, A, R^2) while corresponding names (slope, aspect, roughness) identify the exact but unknown values.

Equation 5.

$$S^2 = slope^2 - \frac{\left(f_x + f_y\right)}{3}\left[\left(f_{xxx} + 2f_{xyy}\right)P_x^2 + \left(f_{yyy} + 2f_{xxy}\right)P_y^2\right]h^2 + O\left(h^4\right)$$

Equation 6.

$$A = aspect - \frac{30}{\pi}\left(\frac{f_y\left(f_{xxx} + 2f_{xyy}\right)P_x^2 - f_x\left(f_{yyy} + 2f_{xxy}\right)P_y^2}{f_x^2 + f_y^2}\right)h^2 + O\left(h^4\right)$$

aspect is undefined if ${f_x}^2 + {f_y}^2 = 0$

Equation 7.

$$\frac{R_{5x5}^2}{h^2} = \begin{cases} \frac{3}{4}\frac{\left[P_x^2\left(f_x f_{xx} + f_y f_{xy}\right)^2 + P_y^2\left(f_y f_{yy} + f_x f_{xy}\right)^2\right]}{f_x^2 + f_y^2} + O\left(h^2\right) & \text{if} \quad f_x^2 + f_y^2 \neq 0 \\ -\frac{1}{9}\left[\left(P+Q\right)\left(UP_x + VP_y\right) + UVP_xP_y + PQ\right] + \\ \qquad \frac{7}{24}\left(P^2 + Q^2\right) + O\left(h^2\right) & \text{if} \quad f_x^2 + f_y^2 = 0 \end{cases}$$

being

Equation 8.

$$U = \sqrt{f_{xx}^2 + f_{xy}^2}; V = \sqrt{f_{yy}^2 + f_{xy}^2}; W = f_{xy}\left(f_{xx} + f_{yy}\right)$$
$$P = \sqrt{U^2P_x^2 - 2WP_yP_x + V^2P_y^2}; Q = \sqrt{U^2P_x^2 + 2WP_yP_x + V^2P_y^2}$$
$$h = \max\left(\Delta_x, \Delta_y\right);\ P_x = \Delta_x / h;\ P_y = \Delta_y / h$$

The abovementioned expressions neglect the uncertainty of the elevation itself and only consider the discretization of the topographic surface due to the DEM resolution. Further research will be required to consider all the uncertainties at once. Since the partial derivatives of the topographic surface are a linear combination of the elevations, its uncertainty can be easily ascertained. The propagation of such uncertainties from the partial derivatives to the slope or aspect is not linear. The same happens to the roughness, now involving first and second order derivatives.

Once having uncertainty estimates at pixel level for either elevation, slope or roughness, tolerance estimates can be derived for each of the metrics of interest (RMSE, MAE, etc.) computed now at the tile level (not pixel level) (Heuvelink et al. 1989, Oksanen and Sarjakoski 2005, among others). The case of aspect has its peculiarities because of its cyclic nature, but in any case the uncertainty estimate at the pixel level is available. To be complete, each metric must provide either a value plus a tolerance (as an educated guess, or as a procedure to compute it) or a CI, despite the ranking can be anyway carried out assuming a tolerance 0.0 by default.

10.2.3. Other Criteria of Interest for DEMIX

Through the many discussions of the DEMIX SG2, it became obvious that many users are requiring information about DEMs based on a variety of aspects which were not only technical. We wish to highlight that the DEMIX wine contest framework is able to integrate such information into the intercomparison exercise and rank them as long as the chosen criteria can be ordered by the user.

Other possible criteria might be classified as weak or strong. Weak criteria describe just a preference. They can be included by adding a row to the opinion table. Such a weak criteria however, will only have a very limited effect. For example, when considering the cost of the DEM, the candidates could certainly be ordered. However, that will not prevent the winner's cost to be large and potentially outside the limits of the available budget. In such a case, the correct way to proceed is to consider the data price as a strong criterion. The set of suitable DEMs should include only those within a user's budget and therefore some of the considered DEMs will be banned from the beginning.

A user could make such a decision based on technical requirements as well especially if they need to accept some characteristic minimum requirements for the DEM application such as spatial resolution and/or vertical accuracy.

An ad-hoc possibility to increase the weight of certain criteria vs others without excluding certain candidates a priori could be to organize the opinion creation and ranking in several levels. E.g. could the final ranking of a wine contest for elevation difference be used as just one opinion in a higher-level test that includes opinions about cost, availability, or completeness. In this way data accuracy, which might be based on a few dozen to hundreds of evaluations, could be made an equally weighted factor as dataset wide and therefore single evaluated criteria such as price. This technique has however not yet been applied in the DEMIX context.

There are a number of wine contest criteria that are important to users but not included in the DEMIX wine contest. Here a few such criteria are presented and described:

1. **Data Licensing** - this criterion refers to the how the data and metadata can be used (not used) by the DEM user. Ideally it should be clearly stated and linked to type of use including research and/or business applications. Ranking can be based on openness of the licences for particular use (by type of licence where more restrictive is worse). If a strong criteria was included then any DEM product whose licence is closed would not be included in the wine contest.

2. **Data Availability** - such a criteria would be linked and ranked based on the user needs. For example, some users may want direct access to download the entire dataset quickly whereas others require graphical interfaces in order to delineate areas of interest and the service automatically extracts and delivers the data. Others may want programmatic access for services development with associated service level constraints in order to maintain high performance. In cloud environments some data might be available fast or for free, while others are served lower or at a cost. Such types of criteria would be ranked based on the user needs and would change based on their requirements and, depending on the case, could pose strong limitations in the options of the DEM candidate set.

3. **DEM Completeness and Consistency** - this dataset parameter refers to the actual coverage of the elevation data. The different remote sensing technologies used to produce the global DEMs have limitations in coverage. DEM completeness could be defined as the degree to which a dataset covers the area it is supposed to cover. Such a criteria is relevant to users not only to understand whether their area of interest (AOI) is fully covered but also whether the data within in their AOI has all been acquired by the same sensor and processed in the same manner.

This list is in no way exhaustive because the DEM user community is very large and includes diverse users across many different domains. Researching the most convenient set of criteria for specific uses is one activity that might be carried out in the future.

10.3. A Short Discussion on Reference Data

Unlike traditional wine contests, the DEMIX initiative focused on applying the technique to DEMs and therefore users do have access to many quantitative metrics to work with. For the comparison of DEMs, the most popular metrics require the use of reference data. Reference data that is applicable to the computation of DEM metrics comes in different shapes and sizes, i.e. they can be defined over a regular grid and have a finer spatial resolution as well as potentially having a higher accuracy. Reference data applicable to DEM metrics can also take the form of sparsely located data such as surveyed geodetic monuments, GNSS ground control points, LiDAR clouds, etc.

A reference can also be an orthoimage produced using a reference DEM, which can be compared to the one processed with the considered DEM. To better understand this type of metric, let's assume that we have a single image and we perform an orthorectification using the reference DEM and each candidate DEM keeping the software and all other parameters of the orthorectification process the same. The closeness of the resulting images to the reference image can be measured using planimetric metrics. For example, measuring planimetric distance between roads observed in the reference image and candidate image. This is an example of f(g(DEM)-g(R)), where g() is the process of orthorectification and f() are one of the metrics valid for polylines (see metrics section 8). This technique provides a quantitative criterion which could be applied on a per tile basis similar to the ones used in (Bielski et al, 2024).

Reference data can also be a linear feature such as a lineament, a road, runway, etc. described using GNSS tracks and discernible within an orthorectified image. In such cases, the metric will take the form of the discrepancy of a function of the DEM and the function of the reference data. The first cases are examples of general metrics that can be associated with the DEM producer whereas the others are tightly connected with DEM use cases (López-Vázquez and Ariza-López 2023). Another case where this could be applied are watersheds, river networks and similar hydrographic features.

To be declared as a suitable reference data set, the reference or fiducial data must be significantly more accurate than the candidate. Standards such as the National Standard for Spatial Data Accuracy (FGDC 1998) require that the accuracy of the reference data must be at least three times better than the data to which it is being compared. In the case of elevation data, this accuracy requirement can be tested with a well-established and well documented procedure. However, for other DEM derived parameters such as slope, aspect, roughness, etc. this requirement is not easily checked. The DEM derived parameters of slope, aspect, roughness, etc. that are computed with a higher accuracy DEM reference dataset do not need to have such a narrow CI to be considered as a reference. Note that using a higher accuracy elevation dataset alone does not necessarily lead to high accuracy of the derived parameters. To illustrate this particular issue the following example is presented: a low resolution DEM might produce a low accuracy value for derived parameters irrespective of its elevation accuracy. In particular, the relationship between the accuracy and the pixel size for slope and aspect has been addressed by López-Vázquez (2022), while some clues for the effect based on roughness has been reported in the 4.3 Section in this document. Other parameters of interest will (in theory) require a similar analysis. Within the DEMIX activities, we have used finer spatial resolution and higher accuracy reference datasets to build the reference DEM, and only afterwards the derived parameters were obtained and applied within the wine contest.

10.4. A Short Discussion on Test Areas

We define test areas as those regions of the world that have suitable reference data for assessing the accuracy of elevation values and their derivatives. In principle these test areas might substantially differ in both, location and extent. By using the DEMIX tile organisation (see section 3.7) we (almost) equalised the size of each unit used within the test area thus mitigating concerns with respect to imbalance when combining results from a variety of tiles. However, it is not necessary to use the DEMIX tiles in order to apply the wine contest and therefore if different test settings would deem more suitable one may consider other spatial sampling schemes.

10.4.1. Sample Size Considerations for Naïve Sampling

If for practical reasons the test areas are a priori limited in number, one should implement an appropriate sampling design which is described below. At the other extreme, if the reference data is available for the entire world there is no need to perform any sampling because data is available everywhere. In such an unlikely case, and despite some exceptions, one can compute the metric over the entire population, i.e. all DEMIX tiles. Such a case is not only of theoretical interest because global reference datasets exist. For example, ICESat or GEDI data coverage are (almost) global and thus every sizable portion of the land masses has at least a handful of reference elevations to proceed. In this case, any traditional metric that analyses sparsely located elevation differences can be applied. Such computations were performed during the activities for three global DEMs, and they are presented elsewhere in this document, for the case of ICESat-1. Another example of a potentially global reference dataset was illustrated in Franks and Rengarajan (2023) for Landsat Collection-2 processing, where each DEM was used to orthorectify two Landsat images of the same overlapping region acquired with different view angles. The discrepancies of the resulting orthoimages are attributed to the errors in the DEM and some metric of image differences can be applied as an indirect indicator. Due to the fact that Landsat imagery covers a significant part of the entire world, the procedure can be (in principle) applied globally.

If the metric of interest requires other types of reference data not available for the whole globe, or if the amount of processing and generated data is deemed excessive, sampling will be required. Simple sampling is the naive strategy whereby a number of individual tiles are selected at random ignoring location, terrain characteristics, etc. Once the sampling locations are chosen, reference data should be acquired for each of those tiles. It implicitly assumes that reference data for every tile exists and can be accessed, something that needs to be confirmed for each chosen sampling area. In our case, the approach is not very practical because the user may not have access to reference data, or the reference data may not exist for the selected tiles. Furthermore, due to the naive sampling strategy, the minimum number of tiles to be included may be too large.

The number T of DEMIX tiles to be considered has a non-obvious effect on the credibility of the wine contest results because it indirectly can control the power of the statistical tests. When applying a statistical test, the user is placed in one of the situations described in **Table 9** . While Type I and II errors are unavoidable, the test parameters should be chosen to keep such errors low. A Type I error is defined as the probability to reject H_0 despite it being true. It is controlled by the confidence level α that is usually prescribed in advance with a typical value of 5% or 0.05. If DEM_A is the best, choosing DEM_B as the winner will produce a Type I error. A Type II error is the probability to accept H_0 despite it being false and is usually denoted with the greek letter β but is not commonly used as such. Instead, it is more popular to talk about the statistical power of the test defined as 1-β. The statistical power of the test is a function of α, the sample size and other properties of the used statistics. A desirable value to achieve might be as high as 0.80 or 80%. The power of the test is increased by increasing the sample size and by decreasing α. In our setting, it is measured by the probability that, despite being the best, we reject that DEM_A is the best.

Table 9. Illustration of the meaning of the Type I and II errors.

		Reality	
		H_0 True	H_0 False
Estimation	Fail to reject H_0	Correct	Type I error ≤ α False positive
	Reject H_0	Type II error ≤ β False negative	Correct

Source: Carlos López-Vázquez

Given α, the sample size T needs to be estimated to reach a prescribed statistical power. If the sample size is too low, the power of the test will be low and any conclusion can (should) be questioned. On the other hand, if sample size is too large the cost of the test might become prohibitive. The goal is to find a number T of DEMIX tiles with suitable reference data available that are distributed around the globe and are representative of the wanted landforms and landscapes.

The *a priori* definition of the sample size T has not yet been considered. Unlike other sampling experiments, the outcome of the DEMIX wine contest is a ranking, i.e., a discrete output, so the relationship of its uncertainty to the number T is yet to be fully investigated. This particular problem however is not so simple because the standard statistical theory requires some estimates of the distribution of the variable to control implying that the minimum T may be related to the criterion applied, i.e., minimum T may not be the same for RMSE and MAE. To improve our understanding of the sampling effects described, an experiment could be envisioned using global ICESat-2 or GEDI data as reference for the entire population of DEMIX tiles and applying different metrics such as the RMSE computed for sparse locations as the criterion. By subsampling the DEMIX tile universe with varying sample size T, we can ascertain the frequency when the winner of the contest is the one obtained from the reference case computed using the full set of DEMIX tiles. For the time being, DEMIX uses the largest T as possible which is directly limited by the existence and access to reference data.

10.4.2. Stratified Sampling

In the context of the DEMIX wine contest, it was not possible to choose the DEMIX tiles first and then collect the associated reference data because there was no way to guarantee that reference data was available. Therefore, it was necessary to proceed in the opposite way, i.e. collect reference data first and then associate the DEMIX tile sampling to the regions where reference data is available. In this manner, DEMIX was able to collect from members and open sources as much reference data as possible and then fit them to the DEMIX tiles. Then it was possible to perform the necessary sampling over this available subset. Using such an approach includes a certain risk of bias because some terrain types might not necessarily be well represented in the set. Consequently, two situations can be envisioned due to this fact. On the one hand, if the DEMIX wine contest user is interested in just specific types of terrain, e.g. mountainous, then it is reasonable to apply a sampling strategy from within the available DEMIX tile set that only includes such a type of terrain. On the other hand, if the user is interested in global applications, then the assumption is that all types of terrain should be included, and the sampling should be representative of the proportion of terrain types found around the globe. This process is known as stratified sampling and was integrated into the DEMIX wine contest through filtering by DEMIX tile characteristics and are presented in section 4.7.

Assuming that the sample size T is chosen freely, to achieve a valid conclusion for the entire globe one may proceed as follows:

- Select one tile classification criteria, able to organize the whole set into a finite number of classes. One possibility is using landforms and its associated classes. Another is to use classes as depicted by K-means classification.
- Compute the prevalence of each of such classes for the whole world. The output could look similar to class #1, 18.5%; #2, 17.4%, … ,#15, 2.3%, etc.
- With both outputs in hand, and after selecting a suitable sample size T, the number of cases for each class will be defined. Afterwards, perform naive sampling within each class.

Stratified sampling is a well-known procedure that has been validated over a number of different geospatial use cases. The power of this sampling technique is that requires fewer samples than applying blindly the simple sampling procedure. Furthermore, the described approach can be applied at different scales considering for example, the heterogeneity within the DEMIX wine contest user's region of interest. When a user needs to select the best global DEM for their particular region, they may first compute the distribution of Maybeck’s classes (Maybeck et al. 2001) (or some other aspect of interest) within it. Note that such a distribution does not need to replicate the global distribution. Furthermore, the data set allows estimation of how well the various DEMIX approaches capture representative topography providing users with a valuable source to check representativeness of a certain set with respect to the globe or some part of it (Guth 2026).

10.5. A Short Discussion on the DEMIX Wine Contest Protocol

Selecting the best DEM is a tricky endeavour and once a DEM has been chosen it is assumed that the choice is significant because once you make the selection, it will be integrated into your workflows and analysis which is not easily undone. For this particular discussion, we discard the possibility of selecting the best DEM for different continents, for example the best DEM for Africa and then the best DEM for Europe, which potentially could be different. Therefore, one wants to be as sure as possible that their 'best' selection based on the DEMIX wine contest ranking is valid.

Much of the DEMIX wine contest procedure as it was applied, was described in a previous section (see Section 3). Here we dive deeper into the details to highlight certain aspects of the intercomparison. For each DEMIX tile, a ranking for each criterion comparing all the k DEMs available is produced. This takes the form of a matrix of k columns and C rows. The majority of the literature only reports a single metric (the RMSE) thus C equals 1. If there are T DEMIX tiles in our wine contest, then after collecting all the outcomes a matrix (evaluation table) can be built holding all the opinions. This opinion table is composed of k columns, and N=C*T rows. If a stratified sampling approach was considered, then T will be smaller.

As mentioned, the described DEMIX wine contest procedure is a fairly standard procedure within the empirical sciences. For the DEMIX wine contest, the application of the non-parametric Friedman's test circumvents the data requirements of following specific distributions such as a normal distribution or uniform distribution for the wine contest metric outcomes while still being powerful enough to deliver on the requirements of the DEM intercomparison within DEMIX. Furthermore, the use of the Friedman test does not require a large k or N to be applicable. In the presented DEMIX problem of the intercomparison of up to six global DEMs, k is moderate and can be less than 6. At the same time, there is no bound on the number of criteria C or the number of tiles T to be considered within the DEMIX wine contest. To reject or not reject the null hypothesis, the needed statistical tables are available for low k and N values, whereas for large k and N there exists an asymptotic analytical estimate using the chi-square distribution. For an intermediate number of k or N values, specialised tables are required. The tables are also different depending on whether ties in the ranking are accepted or not. The traditional Friedman's test does not consider ties whereas the DEMIX application was designed to specifically address them because the group realised that this is of value for the intercomparison of global DEMs. The problem was analysed in more detail recently, significantly extending the statistical tables for intermediate values of k and N for both the case with ties (López-Vázquez and Hochsztain 2020, López-Vázquez et al. 2021) and without ties (López-Vázquez and Hochsztain 2019).

11. DEMIX Associated Studies and Platform Activities

Section contributors: Serge Riazanoff and Axel Corseaux

11.1. Introduction

This section highlights work performed within DEMIX Subgroup 3 “Platforms and Processing”. This group primarily focuses on the visualisation and processing of DEMs (data retrieval, analysis and transformations) made broadly available through public tools and web applications.

11.2. Background and Set-Up – The EDAP / EDAP+ Project

Within the framework of its Earthnet Data Assessment Project[3] (EDAP) and later the EDAP+ activity, as part of the EARTHNET program, the European Space Agency (ESA) supports DEMIX through a number of dedicated studies and developments. These DEM related activities are carried out by VisioTerra, a company specialised in EO studies, platform development, expertise and training.

VisioTerra on behalf of ESA engaged primarily in Sub-Group 3 (SG3) by managing the group and providing dedicated DEM access, processing, and visualisation services on its cloud hosted VtWeb[4] platform. Beyond that it performed on request of ESA a series of related analytical and comparative DEM studies which were shared and discussed within DEMIX. All the respective technical notes are publicly available on the “EDAP+ multi-missions & DEM web page”.

The VtWeb platform (**Figure 50**) is a client-server application allowing:

- access to a wide variety of datasets,
- processing them on-the-fly,
- analysing them,
- sharing layer stacks or
- exporting all or part of the rendered layers.

This platform developed by VisioTerra since 2011 has enabled the design of dedicated platforms within the framework of various ESA projects. On-the-fly processing is performed using the POF-ML language (Processing On-the-Fly Macro-Language) and allows for example to calculate the difference between two DEMs at all scales. POF-ML has been presented to CEOS WGISS-52[5] . VtWeb allows us to create interfaces for specialized operations, such as DEM4S2[6] to integrate DEMs and Sentinel-2 imagery, and the DEMIX Operations Platform (Section 11.6).

A large number of public DEMs (ALOS World 3D, ASTER GDEM, CopDEM EEA-10 GLO-30 GLO-90, EU-DEM, FABDEM, MERIT DEM, NASADEM, SRTM) and their versions have been stored in the VtWeb infrastructure at VisioTerra premises. New VHR (Very High Resolution) DEMs like BRA-ES, GBR-ENG,

[3] https://earth.esa.int/eogateway/activities/edap

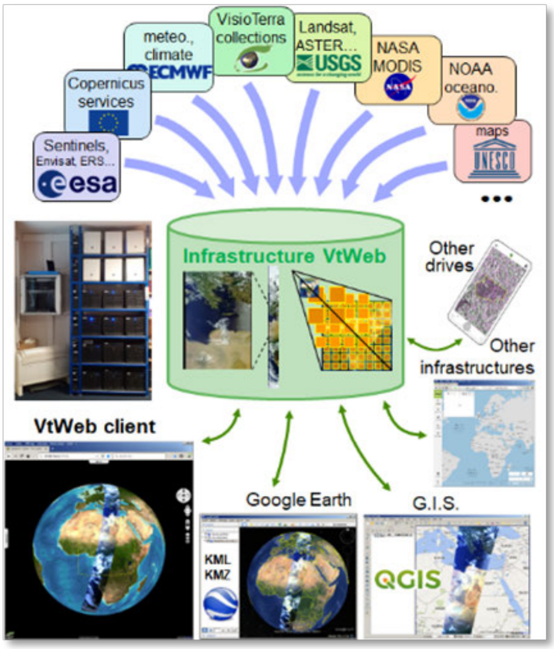
[4] https://visioterra.org/VtWeb/

[5] https://ceos-dev.ceos.org/meetings/wgiss-52/

[6] https://visioterra.org/DEM4S2/

ITA-TRE, RGE Alti or URY have been added to serve as reference. All these candidate and reference DEMs were resampled on a common grid[7] defined by the Open Geospatial Consortium[8] (OGC). To allow visualisation and analysis at all scales, each DEM is organised into a quadtree by mean aggregation (2x2, 4x4, 8x8, 16x16, etc.) from the base level of the DEM in the quadtree. This base level corresponds to the level just below the original ground sampling distance (GSD) of the DEM. Thus, a DEM at 1'' arc (approximately 30 m at the equator) is oversampled at ⅔'' arc (approximately 20 m). Another invention of VtWeb is the sharing of hyperlooks allowing to pool a workspace or to build galleries of use cases. A hyperlook is a rich URL identifying the server, the layers, the processing parameters, the viewing geometry. Activating a hyperlook enables to interactively navigate / zoom in / zoom out / select layers / change POF-ML parameters / re-share the results, and many others. The list of documents found at the VisioTerra website are called "hyperlook documents" showing images annotated by hyperlooks of 2D, 3D views or animations.

Figure 50. Illustration of the VtWeb Platform.



Source: Visioterra

The comparison of DEMs in VtWeb makes it possible to highlight the differences between two DEMs or two versions of the same DEM but does not make it possible to decide which one of the two DEMs is better. At most, we can suspect observation defects coinciding for example with the orbits of the satellites used to generate the DEM or even holes or salient peaks, geometric shapes that suggest defects that do not look like natural topographic patterns.

This way of directly comparing two DEMs without one being previously designated as the reference is different from the "Wine contest" strategy (see section 4) offering a series of tests to be applied DEM by DEM, computing scores and ranking. This method also overlays the interpolated DEMs,

[7] http://docs.opengeospatial.org/is/17-083r2/17-083r2.html#63

[8] https://www.ogc.org/

which is particularly challenging when the two DEMs have a different geometric basis or when the grid spacing increases at higher latitudes, most notably between CopDEM and ALOS World 3D.

11.3. Product Quality Evaluations Matrix

As done for SRTM, ASTER GDEM and ALOS World 3D, an important work of research has been carried out to retrieve the documentation of Copernicus DEMs. The richness of this documentation has been assessed complying with the EDAP QC standards (see EDAP Best Practice Guidelines[9]). **Table 10** compares the evaluation matrix of Copernicus DEMs against the three other global DEMs. Only one matrix has been created for the three instances of Copernicus DEM, as they share the same generation procedures and the same documentation.

The documentation of Copernicus DEMs is significantly better than the three other global DEMs. One may note seven "Excellent" grades have been given to Copernicus DEMs, as opposed to only one for SRTM, ASTER GDEM and ALOS World 3D.

Table 10. Product quality evaluation table including SRTM, ASTER GDEM, ALOS World 3D and Copernicus DEMs.

	SRTM	**ASTER GDEM**	**ALOS World 3D**	**Copernicus DEMs**
Product Details	Excellent	Excellent	Excellent	Excellent
Availability & Accessibility	Good	Good	Good	Excellent
Product Format	Basic	Good	Good	Excellent
User Documentation	Intermediate	Intermediate	Intermediate	Intermediate
Metrological Traceability Documentation	Intermediate	Not Assessable	Not Assessable	Intermediate
Sensor Calibration & Characterisation Pre-Flight	Intermediate	Good	Good	Good
Sensor Calibration & Characterisation Post-Launch	Intermediate	Good	Good	Good
Retrieval Algorithm Method	Intermediate	Good	Good	Excellent

[9] https://earth.esa.int/eogateway/activities/edap/edap-best-practice-guidelines

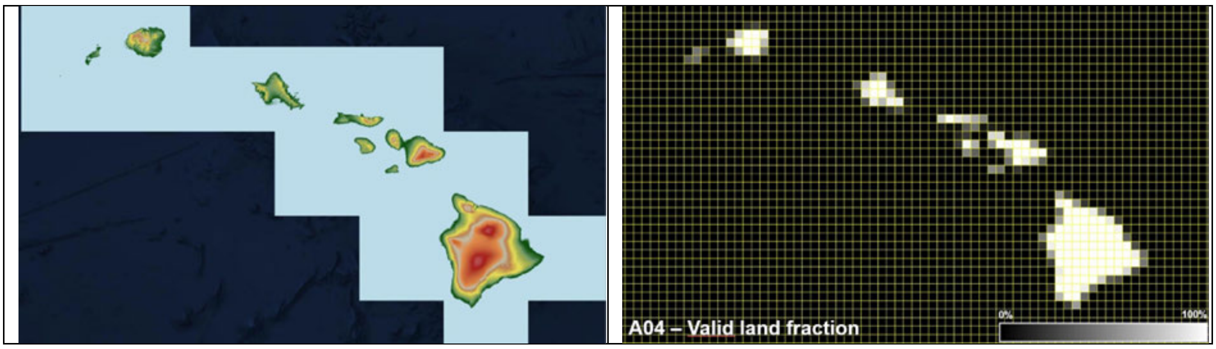
Retrieval Algorithm Tuning	Intermediate	Intermediate	Not Assessable	Intermediate
Additional Processing	Intermediate	Not Assessable	Not Assessable	Excellent
Product Flags	Good	Not Assessable	Good	Excellent
Ancillary Data	Good	Good	Good	Excellent
Uncertainty Characterisation Method	Basic	Basic	Not Assessable	Intermediate
Uncertainty Sources Included	Intermediate	Intermediate	Basic	Good
Uncertainty Values Provided	Intermediate	Intermediate	Intermediate	Good
Geolocation Uncertainty	Intermediate	Intermediate	Not Assessable	Basic
Reference Data Representativeness	Good	Good	Good	Intermediate
Reference Data Quality	Intermediate	Intermediate	Intermediate	Intermediate
Validation Method	Intermediate	Intermediate	Intermediate	Intermediate
Validation Results	Good	Intermediate	Good	Intermediate

Source: Visioterra

11.4. Global Criteria Assessments

DEMIX proposed five completeness criteria (A01 to A05) and we computed them for all the candidate DEMs, over all the existing DEMIX tiles.

Figure 51. Global criteria assessments methods - from the DEM (left) to the scores of "A04 - Valid land fraction" per DEMIX tile image (right).



Source: Visioterra

The completeness criteria have been applied to the following DEMs: ALOS World 3D, ASTER GDEM, Copernicus DEM GLO-30 and GLO-90, EU-DEM, NASADEM and SRTM. For any DEM, DEMIX tile and criterion given, the pixels matching the area of the DEMIX tile are filled with the local score of the DEM for this criterion. The dimensions of the output image are chosen to allow a perfect matching between the DEMIX grid and the pixels of the image (**Figure 51**).

The images for each of the five criteria are saved as GeoTIFFs in the DEMIX repository (https://visioterra.fr/telechargement/P317_DEMIX/global_scores/) and on a Zenodo dataset repository (https://doi.org/10.5281/zenodo.11389297).

11.5. DEM Transformation Studies

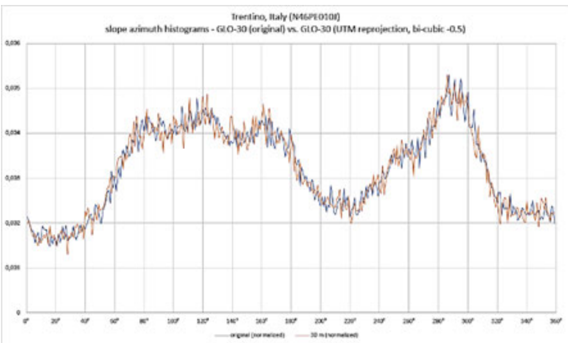
11.5.1. Study 1 – Impact of the Reprojection on DEM Features

We have defined seven features (geomorphometric measurements) of the DEMs: height, slope norm, slope azimuth, horizontal curvature, vertical curvature, tangential curvature and local standard deviation of the slope norms. These last four measurements are considered as small-extent roughness estimators, i.e. over a small area consisting of a 3x3 window around the pixel to be estimated. To overcome complex calculations of geodesic arcs in the geographic (GEO) coordinates reference system (CRS), users may request that the DEMs be projected onto a map. We have chosen the local UTM (i.e. the UTM whose hemisphere and horizontal zone are automatically calculated according to the longitude and latitude of the centre of the tile). For each of the 7 features, we calculate the difference between the histogram of the feature in UTM minus the histogram of the feature in GEO. These differences are computed in three (3) areas: Trentino (Alps) presenting mountains and a wide valley (**Figure 52**), -La Reunion (Indian Ocean) corresponding to the northern part of a volcanic cone, -Les Landes (South-Western France) corresponding to a very flat area close to the Atlantic Ocean.

As conclusions, the reprojection has:

- a stronger (relative) impact in quasi-flat regions
- a greater impact in vertical and tangential curvature calculations.

Figure 52. Computed differences for the Trentino region.



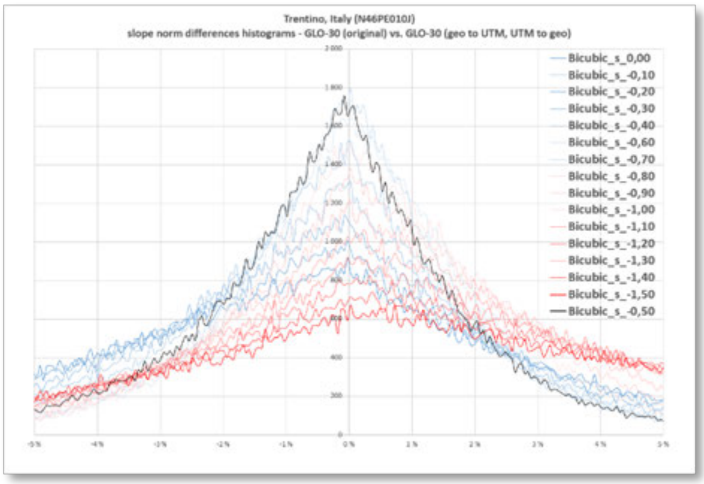
Source: Visioterra

11.5.2. Study 2 – Impact of the Resampling Method on DEM Features

During the reprojection, an interpolation of the values of the candidate DEM is used according to the bicubic method. This method uses the 4x4 pixels around the point to be interpolated by calculating a weighted average according to a "bicubic weight function". There are several weight functions depending on the value of the derivative at the point (1,0) which makes it possible to obtain a bilinear smoothing (derivative at 0) up to a strong value of the edge enhancement effect (value at -1.5). By varying these values of the derivative, we show that the best bicubic, i.e. those minimizing the difference between the histogram of the GEO image and the histogram of the UTM image (see study 1), are not the same depending on the feature that is sought to be preserved. These tests were applied to the three (3) zones of Trentino (Italy) (**Figure 53**), La Réunion (France) and in the Landes (France).

We note that the quasi-flat zones are better interpolated with bicubics with a very strong Laplacian effect (derivative equal to -1.5) for all features except the “slope azimuths” which are better preserved using a derivative equal to 0.

Figure 53. Slope norm differences for the Trentino (IT) region.



Source: Visioterra

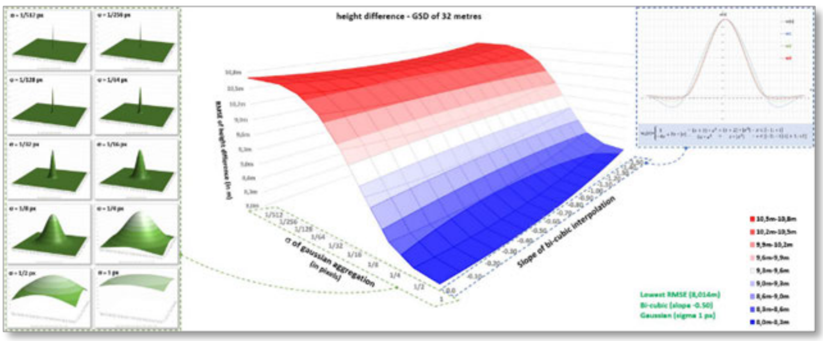
11.5.3. Study 3 – DEM Multiscale Analysis

The ground sampling distance (GSD) of VHR DEMs is very different from that of candidate DEMs; from 1 to 2 meters for the first ones against 30 meters for the candidate DEMs. The objective of this study is to statistically compare the images of the features of the reference DEM VHR and those of the candidate DEM that have been projected in the same CRS with the same GSD. We will vary these GSD values from 4m, 8m, 12m, 16m… up to 72m. The 4 features considered are heights, slope norms, slope azimuths, and local standard deviation of slope norms. To bring the VHR DEM to a given GSD, several aggregation methods are used by varying the standard deviation of a reference Gaussian centred on the pixel around which to aggregate and defining the weight of the neighbours in the aggregation window. For example, the aggregation window includes (GSD+1)x(GSD+1) pixels if VHR DEM is at 1 meter. The standard deviation of the Gaussians of aggregations is expressed in pixels according to the GSD of the output image: 1 (quasi-flat surface equivalent to an average), 1/2, 1/4, 1/8, 1/16… up to 1/512. The last standard deviation values correspond to dominant weight affected to the central point, that is equivalent to a simple decimation considering only the central pixel (nearest neighbour logic). Conversely, the oversampling of the candidate DEM from coarse GSD to finer GSD is carried out according to the bicubic interpolation of which the derivative will vary from 0 to -1.5 with a step of -0.1 (see study 2). For each couple (standard deviation of aggregation, derivative of interpolation), we calculate the quadratic mean of the differences for the feature considered, for each of the 3 zones (Trentino, La Réunion, Les Landes) and for the different values of GSD.

In conclusion, the search for combinations showing the lowest mean squared errors shows that:

- the largest GSDs (72m) decrease the difference between references and candidates,
- the average aggregation (standard deviation of the Gaussian at 1 or 1/2) is often the best except for the slopes in the Landes,
- the bicubic interpolation with a derivative close to 0 (often at -0.3) is often the best.

Figure 54. Results of comparing height differences.



Source: Visioterra

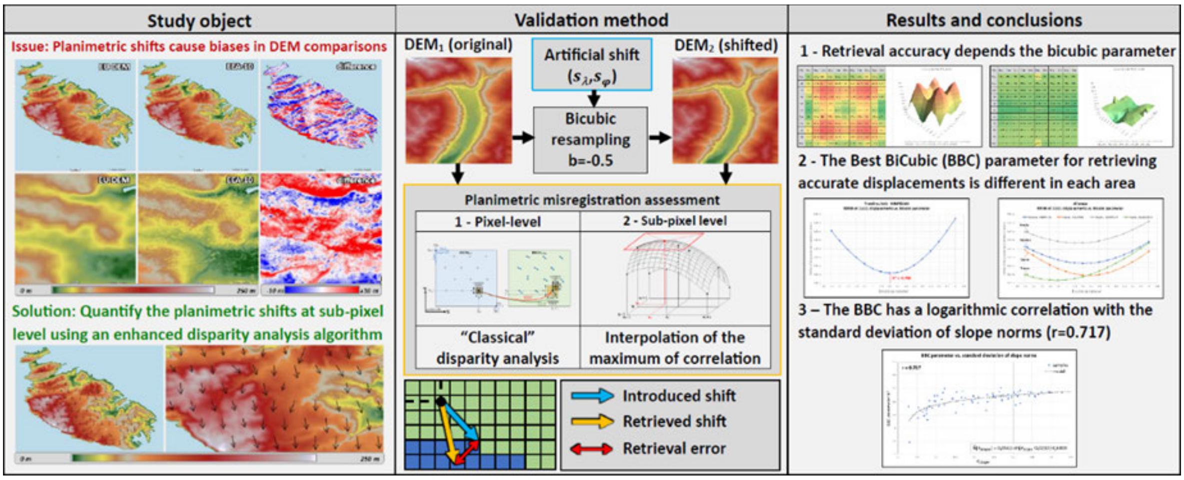
11.5.4. Study 4 – DEM Planimetric Mis-registration Assessment (article)

In this article Riazanoff et al. (2024), the negative effects of planimetric mis-registrations (or shifts, displacements) on Digital Elevation Models (DEMs) are highlighted, especially in the case of DEM comparisons (elevation differences). A novel method (**Figure 55**) allowing to quickly and accurately retrieve these displacements at a sub-pixel level is presented. As this algorithm requires the DEMs to be co-gridded, one of the input DEMs should be resampled to match the pixels of the other DEM.

It has been discovered the following:

- the accuracy of the displacement retrieval depends on the resampling parameters (bicubic parameter) used to resample the input DEMs,
- the Best BiCubic (BBC) parameter is different depending on the study area (low on flat areas, high in mountainous areas),
- a logarithmic correlation coefficient of r=0.717 has been found between the BBC parameter and the standard deviation of slope norms of the DEM.

Figure 55. Graphical abstract of the planimetric mis-registration assessment methodology.



Source: Visioterra

11.5.5. Study 5 – DEM Datums and Transformations

Systematic planimetric shifts have been noticed when trying to compare candidate DEMs to VHR DEMs of North America and England. VisioTerra has carried-out an extensive study on these DEMs to understand the transformations required to express these VHR DEMs in the DEMIX standard reference systems (WGS84/EGM2008). It has been shown that the DEMs of Alaska[10] (NOAA) and USA[11] (USGS) are expressed in local horizontal and vertical datums (NAD83 and NAVD88) which are not equivalent to the international ones (WGS84/ITRS and EGM2008).

In such cases, the study has shown that the four following datum characteristics are of crucial importance in order to perform correct DEM transformations:

- **Name** – The name of the datum (e.g., WGS84 or ITRS)
- **Ellipsoid** – The reference ellipsoid of the datum (e.g., WGS84 or GRS80)
- **Realization** – The realization of the datum (e.g., WGS84_G1150 or ITRF97), which can be seen as a datum "version".
- **Epoch** – The epoch used for the datum, which is generally standard or within the acquisition period of the data used to generate the DEM.

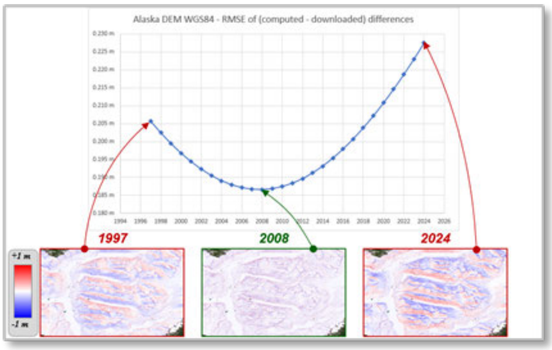
Note: In the examples provided (**Figure 56**), one can see that the name "WGS84" can refer to a datum or to the ellipsoid of this datum depending on the context.

With these metadata, both Alaska and USA DEMs have been expressed in the WGS84/EGM2008 reference systems, with no major shifts visible in differences with candidate DEMs (excluding local anomalies).

[10] https://coast.noaa.gov/dataviewer/#/lidar/search/

[11] https://apps.nationalmap.gov/downloader/#/

Figure 56. Example of the Alaska DEM differences.



Source: Visioterra

11.6. Cloud Based Platforms

Many of the SG3 studies include views of DEM processing performed on the VtWeb platform. While the VtWeb platform is powerful, it is not specifically designed to support the DEMIX activities. For this reason, the "VT-DEMIX Jupyter Notebook" and the "DEMIX Operations Platform" were developed and released.

The VT-DEMIX Jupyter Notebook[12] is a tool primarily targeting scientists, used to retrieve DEM elevations and ancillary data using the Python programming language. The DEMIX Operations Platform targets a broader public than the VT-DEMIX Jupyter Notebook. With a few clicks, users may tune and run the two main functions, which are the following:

- Export DEMIX tiles - This function allows the export of DEM elevations as an archive of GeoTIFFs. Users may select multiple DEMIX tiles on the 3D globe, a DEM of interest and output parameters (CRS, VRS, GSD, resampling method and pixel-type).
- Compute criteria and ranking - This function allows to import DEM scores. A DEM ranking is then computed and rendered, according to the DEMIX SG2 principles.

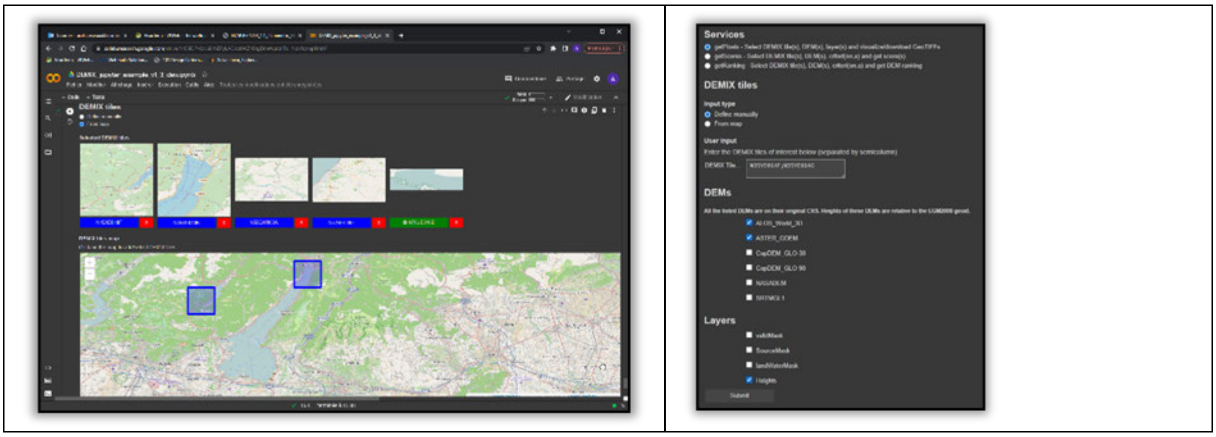
These two solutions are further described in the following subsections.

11.6.1. VT-DEMIX Jupyter Notebook

Hereafter are shown screenshots of the latest "VT-DEMIX Jupyter Notebook" version 1.3 (downloaded at https://visioterra.fr/telechargement/P317_DEMIX/notebook/releases/), running on the Google Colab cloud environment (https://colab.research.google.com/).

[12] https://visioterra.fr/telechargement/P317_DEMIX/notebook/releases/

Figure 57. VT-DEMIX Jupyter Notebook in the Google Colab environment.



Source: Visioterra

Download and Set-up: The VT-DEMIX Jupyter Notebook consists of a Jupyter Notebook file (.ipynb extension), including cells of script (in the Python language) and cells of formatting / documentation (in the Markdown format).

The Jupyter Notebook files need a proper environment to be executed. Such environments can be installed locally on any machine (see https://jupyter.org/). Alternatively, services such as Google Colab allow users to run notebooks on the cloud, with no prior installation required. This last option is chosen for the next steps.

In order to run the VT-DEMIX Jupyter Notebook, the following must be done:

1. Download the latest version of the notebook (DEMIX_jupyter_example_v1_3.ipynb),
2. Open an execution environment such as Colab https://colab.research.google.com/ ,
3. Go to “File” -> “Open notebook” -> “Upload” tab,
4. Upload the “DEMIX_jupyter_example_v1_3.ipynb” to Colab,
5. Click on “Runtime” -> “Run all”.

This last step will run all the Python cells of code, individually followed by their own output. In case of any error during the installation / import of the “DEMIX lib”, follow these additional steps:

1. Click on Runtime -> Restart runtime and Yes,
2. Click on Runtime -> Run all.

As previously stated, the VT-DEMIX Jupyter Notebook relies on the DEMIX lib, which is a Python library developed specifically for the needs of the DEMIX activities. Most sections of the notebook can be seen as tutorials and examples for using the DEMIX lib.

- **DEMIX lib functions** - This section presents the main functions of the DEMIX lib. Examples of this section include retrieving the name and polygon of a DEMIX tile based on provided coordinates (longitude and latitude), computing criteria scores on a selected DEM and retrieving DEM heights or ancillary data over a selected DEMIX tile,
- **Available DEMs and criteria** - This section presents the DEMIX lib functions used to retrieve the lists of supported DEMs and criteria,

- **Interactive examples** - This section provides advanced examples of usage for the DEMIX lib, including the visualisation of DEM heights and ancillary layers as well as the definition of custom criteria in Python.

- **User interface** – This last example provides user-friendly access to the functions of the DEMIX lib (no coding required). Users may export GeoTIFFs of DEM layers, compute criteria and compute DEM ranking by tuning parameters in this example's interface. This interface can be seen as the first iteration of the "DEMIX Operations Platform", which is described in the next section.

11.6.2. DEMIX Operations Platform

The DEMIX Operations Platform (**Figure 58**) is a DEM processing tool allowing users to export custom GeoTIFF images of Medium Resolution (MR) and Very High Resolution (VHR) DEM elevations, as well as ranking DEMs according to several criteria based-on the DEMIX SG2 RCBD principles.

The latest version 2.2 of the DEMIX Operations Platform is available at https://visioterra.org/DemixOperationsPlatform/ . A detailed list of changes for each version can be seen in the changelog[13]. The main features of this platform are currently the following (ordered by operation type).

1. **Export DEMIX tiles:**

- Selection of multiple DEMIX tiles on a 3D globe,
- Selection of a DEM (including VHR DEMs) and its version,
- Access to DEMs characteristics tables,
- Selection of export parameters, which are the CRS (Geographic or UTM), VRS (EGM2008, EGM96 or WGS84), GSD, resampling method (among nearest neighbour, bilinear and bicubic for interpolation, mean and gaussian for aggregation) and pixel-type (point or area),
- Export of the selected DEMIX tile as GeoTIFF, compliant with the parameters specified by the user.

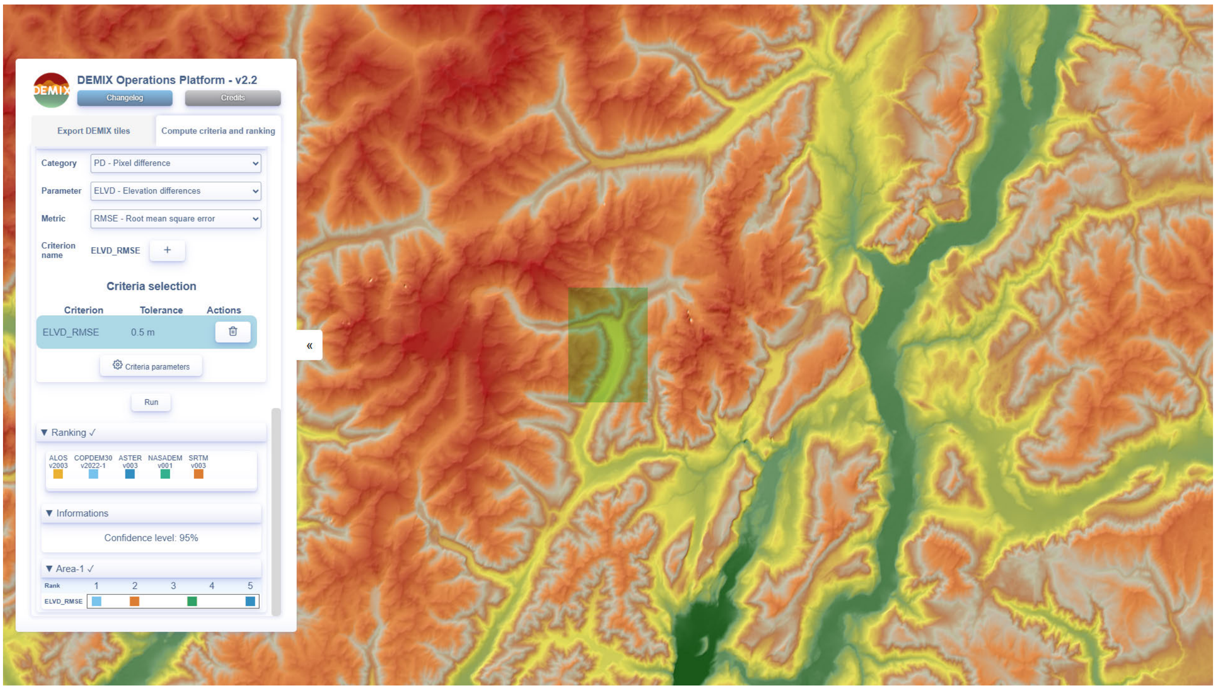
2. **Compute criteria and ranking:**

- Selection of DEMIX tiles and DEMs (similarly to the first function)
- Selection of criteria, among the completeness and pixel difference categories
- Computation of scores based on each (DEMIX tile, DEM, criterion) tuple
- Progressive computation and display of a DEM ranking, compliant with DEMIX SG2 principles

[13] https://visioterra.org/DemixOperationsPlatform/changelog.html

As the latest version 2.2 allows to compute a DEM ranking compliant with the RCBD principle of DEMIX SG2, the future updates will mainly be focused on user feedback and/or future criteria and methodologies of DEMIX.

Figure 58. View of the DEMIX Operations Platform version 2.2 interface.



Source: Visioterra

12. Latest Developments Following the DEMIX Report

Section Contributors: Peter Guth, Carlos López, Conrad Bielski and Peter Strobl

Since the completion of the DEMIX report in late 2023, a number of contributions by members of the DEMIX working group and associated members have extended both the methodological framework and the theoretical understanding of wine contest. These developments reinforce and broaden the scope of the original work. This section presents the main published developments based on the DEMIX sub-group 2 results.

While this report was being prepared (main work concluding at the end of summer 2023), three preprint versions of the DEMIX wine contest research paper were already available[14]. The final open journal publication of Bielski et al. (2024) was authored by twelve contributors. Accordingly, while Bielski et al. (2024) documents the DEMIX state-of-the-art as of late 2023, it does not capture the full scope of the work carried out within SG2, which is more comprehensively presented in this report.

In addition, DEMIX has become an active component of the Geomorphometry conference series (https://www.geomorphometry.org/), providing a dedicated forum for researchers and practitioners to present results, exchange ideas, and foster new collaborations.

The following subsections highlight the expanded and published works since the completion of the DEMIX report.

12.1. Publications from 2023

López-Vázquez and Ariza-López (2023) published the open journal article entitled Global Digital Elevation Model Comparison Criteria: An Evident Need to Consider Their Application. The following highlights are extracted and summarised from it.

The contribution is a meta-analysis of global digital elevation model (GDEM) comparison practices based on a corpus of published literature. The study is based on a dataset comprising 313 scientific papers retrieved from bibliographic databases and web searches. It investigates how GDEMs are compared in scientific studies, with the aim of identifying common methodologies, reference data, and evaluation criteria. The objective of the work was to identify:

- which DEMs are most frequently compared
- what reference data are used
- how accuracy is evaluated
- whether comparison methodologies are consistent

The paper shows that GDEM comparison studies are widespread but methodologically inconsistent, and it calls for global standardization to make results comparable, reproducible, and meaningful. The authors argue that most studies focus on isolated evaluations rather than consistent comparative frameworks. Unsurprisingly, most studied DEMs are SRTM and ASTER GDEM while

[14] https://arxiv.org/abs/2302.08425

newer DEMs (Copernicus, ALOS, NASADEM, FABDEM) appear but less systematically. The paper identifies as a major problem that reference data is inconsistent. In particular often reference accuracy is not justified or not even reported. This undermines comparability across studies.

The RMSE of elevation differences is the standard metric, while other metrics (slope, aspect, functional performance) are rare. This leads to a data-centric rather than application-centric evaluation culture. The study finds that sampling strategies vary widely, statistical assumptions are often unchecked, error distributions are rarely validated and outliers and autocorrelation are mostly ignored. As a result, studies are not statistically homogeneous.

The core problem identified is the absence of a widely standardized framework for comparing GDEMs in common applications, which leads to non-interoperable and often non-reproducible results. Although standards such as NSSDA were published a long time ago, they do not appear to be widely adopted within the community. This affects ranking of DEMs, meta-analysis of the literature, as well as reproducibility of conclusions. As its main conclusion the paper calls for widely accepted standardized comparison methodologies, consistent reference data definition, unified statistical treatment, clearer reporting protocols and inclusion of application-oriented (fitness-for-use) criteria. In line with some post-DEMIX initiatives, it also suggests developing a global reference framework for DEM evaluation.

12.2. Publications from 2024

Guth et al. (2024) published the open journal article entitled Ranking of 10 Global One-Arc-Second DEMs Reveals Limitations in Terrain Morphology Representation. This publication made a number of enhancements to the Bielski et al. (2024) paper's data and methodology. The following highlights are extracted and summarised in bullet points:

- Included 124 test areas and 3462 DEMIX tiles, an order of magnitude increase over Bielski et al. (2024).
- Added 4 additional global DEMs: DSM TanDEM-X, and DTMs CoastalDEM, DiluviumDEM, and DeltaDTM.
- Used 16 different derived land surface parameters (LSPs) in addition to elevation to compare grids. This compares to 3 LSPs used in Bielski et al. (2024), where the 5 criteria for each LSP are highly correlated. The new DEMIX GIS database version 3[15] includes the original 15 criteria for anyone who still wants to use them.
- Introduced fraction of unexplained variance (FUV), ranging from 0 (best) to 1 (worst), as a more differentiated criterion compared to the statistics of the difference distribution for the LSPs. This allows visually comparing various criteria and illustrates in detail the degree of correspondence correspondence with the reference DEMs.
- Used 2 different raster classifications, based on geomorphometry, of every pixel in the DEMs, also with a range from 0 to 1.

[15] https://zenodo.org/records/13331458

- Expanded the channel network criterion from one test area to the entire data set, also with a range from 0 to 1.

- Used the correspondences rather than the rankings as a more nuanced way to express the differences among the global DEMs, decreasing the effects of having to set a tolerance for ties. No Friedman statistics were performed, as they argued them to be less insightful than the proposed FUV metrics and graphics.

- Due to the severe challenges of applying subjective criteria at scale to hundreds or thousands of test tiles, the feasibility of the full RCBD (wine contest) statistics becomes an issue.

The key results from this publication (Guth et al., 2024) include:

- CopDEM remains the best performing DEM, even when compared to a reference DTM or and against the edited DTM derivatives of CopDEM to produce a DTM.

- The edited DTMs (FABDEM, CoastalDEM, DiluviumDEM, and DeltaDTM) generally improve on elevation metrics compared to CopDEM but are worse for the derived LSPs because they were only trained for elevation to remove vegetation and they can hallucinate.

- The LSPs can be ranked in terms of their FUV and its effect on the sensitivity of the LSPs to DEM quality; curvature measures are the hardest for the global DEMs to match the reference.

- None of the global DEMs, even the edited ones designed for low elevation coastal areas, perform well in areas with average slopes under 5%.

- Subjective criteria for hundreds or thousands of test tiles not feasible

- Edited DTMs improve on elevation metrics but worse for derived LSPs (hallucinate).

- Confirmation that a number of DEMs should be retired: SRTM, NASADEM, TanDEM-X, and ASTER GDEM.

- Introduced two pixel origin models, named for ALOS and SRTM. They are based on more than just the pixel-is-point and pixel-is-area characteristics. These determine whether the two grids align, or if they have a ½ pixel offset. They are based on pixel geometry, and not sampling to create the grids.

Riazanoff et al. (2024) published an open journal article entitled Best BiCubic Method to Compute the Planimetric Misregistration between Images with Sub-Pixel Accuracy: Application to Digital Elevation Models. The following highlights from this publication are extracted and summarised:

The paper shows that sub-pixel horizontal misregistration between DEMs can be robustly estimated using an improved disparity analysis method, although accuracy strongly depends on the bicubic resampling parameter, which should be adapted to terrain roughness. Planimetric misregistration refers to sub-pixel horizontal shifts between DEMs and their dependence on resampling choices. Even small shifts can significantly bias elevation differences and derived products such as slope and curvature. The study improves a disparity analysis method to:

- detect DEM-to-DEM horizontal shifts

- refine them to sub-pixel accuracy

- and analyse how bicubic resampling parameters affect results.

A central insight is that resampling is not neutral: it directly affects displacement accuracy. Sub-pixel displacement estimates show a significant improvement in precision compared to pixel-level matching. The study introduces a systematic analysis of the bicubic interpolation parameter b used during DEM resampling. It shows that different b values produce different displacement errors, and that the optimal b depends on terrain roughness. The study uses Copernicus DEM (30 m resolution), in which artificial shifts ranging from 0 to 1 pixel (in 0.1 pixel increments) are introduced. Multiple DEMIX tiles representing different terrain types, including flat and mountainous regions, are analysed. The proposed method shows that sub-pixel displacement is achievable, with artificial shifts recovered with errors typically around 10–20% of a pixel size, confirming the robustness of the sub-pixel estimation approach. A "Best Bicubic (BBC)" parameter is identified that minimizes displacement retrieval error for each terrain type and can be estimated via regression on terrain roughness. In summary, the paper provides:

- A fast sub-pixel DEM misregistration estimator
- A systematic evaluation of resampling effects on displacement accuracy
- A method to estimate an optimal bicubic parameter based on terrain roughness

A central conclusion is that DEM comparison accuracy is not only controlled by correlation methods, but also strongly by the interpolation strategy used before comparison.

12.3. Publications from 2025

Guth et al. (2025) published the open journal article entitled Benchmarking Elevation Plus Land Surface Parameters Finds FathomDEM and Copernicus DEM Win as Best Global DEMs. The following highlights from this publication are extracted and summarised in bullet points:

- Analysed the latest FathomDEM (Uhe et al. 2025), which is generally the best performing of the global DEMs if compared against a DSM and if a user can accept its restricted license.
- Analysed the GEDTM (Ho et al. 2025), which is not recommended for use with the version available at the time of publication.
- Based the DEMIX tiles on the delivery tiles from the national mapping agencies. In some cases these are based on tiles in the local projected coordinate system (generally 10x10 km or 1x1 km which we aggregate to 10x10km) or are based on the tiles in the historical national maps where are roughly 10 x 10 km). This makes processing much simpler, provides a consistent sampling system, which unlike the older system of geographic tiles unlike across national boundaries.
- Emphasized the role of the tile's characteristics in the performance of the test DEMs. The most useful classification uses average tile slope and the percentage of either barrenness or forest coverage.
- Discussion of pixel geometry for DEMs, which requires consideration of two Geotiff tags, GTRasterTypeGeoKey (#1025) and ModelTiepointTag (#33922). These refer only to the geometric interpretation of the pixels, and not the sampling used to create the DEM (or other Geotiff raster grid).

The Geomorphometry 2025 conference[16] was organised in Perugia, Italy where several presentations of DEMIX related work was presented. These have been published within a book of proceedings (Alvioli et al., 2026). The following presentations were published as part of these proceedings:

- Corseaux, A., Gross, K., & Riazanoff, S. (2025, April 14). Activities of the DEMIX Subgroup 3 "Platforms and Processing". Geomorphometry 2025, Perugia, Italy[17].
- Guth, P. (2025, June 26). Update on MICRODEM in DEMIX. Geomorphometry 2025, Perugia, Italy[18]. This paper has a discussion of pixel geometry for DEMs, which requires consideration of two Geotiff tags, GTRasterTypeGeoKey (#1025) and ModelTiepointTag (#33922).
- López Vázquez, C. (2025, April 6). Formal weaknesses of some definitions of roughness, and a solution proposal. Geomorphometry 2025, Perugia, Italy[19]. It shows that roughness has direct connection with first and second order partial derivatives of the topographic surface, that can be reduced to second order after detrending the surface.
- López Vázquez, C. (2025). Update on some accuracy and uncertainty topics in DEMIX[20].

Two presentations from the conference were not published within the conference proceedings. One was a review of this DEMIX report for the Geomorphometry audience and the second described work developing methodologies to extend DEMIX to 5 m (0.15 arc second) Global DEMs.

12.4. Publications from 2026

The open journal article entitled A new field-based approach to satisfying NSSDA accuracy requirements in digital slope model assessment by López-Vázquez and Méndez-Baíllo (2026) was accepted. The following highlights are extracted and summarised in the bullet points:

- The goal of the paper is to describe a fieldwork procedure capable of measuring elevation partial derivatives with sufficient accuracy to validate those derived from a DEM.
- Traditional methods for measuring slope exist, but none are sufficiently accurate. Partial derivatives of elevation are not directly observable quantities; they depend not only on elevation error but also on measurement scale and sampling configuration.
- The authors show that measurements taken over a regular grid of equilateral triangles can be used to estimate partial derivatives. Provided that local elevation measurements are error-free, the error induced by the finite triangle side is significantly smaller than that associated with a regular DEM. The triangle side should be chosen equal to the DEM cell size.
- It is shown that the resulting error satisfies NSSDA (FGDC, 1998) requirements when compared with the fourth-order expression of Florinsky (2009) for partial derivative

[16] https://www.geomorphometry2025.org/index.php/

[17] https://doi.org/10.5281/zenodo.15213919

[18] https://doi.org/10.5281/zenodo.15755639

[19] https://doi.org/10.5281/zenodo.15161390

[20] https://www.geomorphometry2025.org/images/geomorphometry_2025_postDEMIXwork_lopez.pdf

estimation; therefore, the method can be used with commonly available second-order formulations (e.g., those of Young and Evans, Horn, and Zevenbergen and Thorne, among others).

- The paper ultimately defines how what “ground truth partial derivatives” can be assessed by in-situ field work means in a way that is consistent with DEM-derived ones. Without such a definition, validation results may be misleading or meaningless.
- In the DEMIX context, this implies that reference partial derivatives (and slope) values could be obtained by first resampling high-resolution, high-accuracy local DEMs onto a local, regular grid of equilateral triangles, and only then estimating slope.

The article entitled Analytical Estimate of DEM Accuracy (MSE) as a Function of Cell Size and Dataset Elevation Error by López-Vázquez (2026) was published in Photogrammetric Engineering and Remote Sensing. The paper addresses a long-standing problem in digital elevation models (DEMs): although cell size and accuracy are usually related, there has been no analytical (closed-form) relationship between them—only empirical formulas. The main contribution is to derive such a relationship analytically, using Mean Square Error (MSE) as the accuracy metric.

The paper proves that DEM accuracy (as MSE) can be estimated as the sum of three types of terms:

- Dataset-dependent term: Depends only on the input elevation MSE measured at grid nodes and is independent of cell size
- Mixed terms: Combine dataset elevation error and cell size h. They appear as monomials involving powers of h
- Pure discretization term: Depends only on cell size h and is thus independent of dataset error.

For a fixed interpolation method the powers of h are constant (method-dependent but topography-independent). Some coefficients depend on terrain characteristics and can be estimated. The author explains why previous empirical relationships worked—but also where they fail. The result shows that DEM error comes from two interacting sources:

- Measurement error (input dataset accuracy)
- Discretization/interpolation error (cell size)

Even for perfect elevation data (i.e. no uncertainty in measurement is assumed) the model captures the impact of finite resolution. Already existing empirical formulas are incomplete because they do not capture the full structure of error and certainly miss interaction terms. The described model provides a theoretical foundation for those approximations.

Aside from these considerations, the author gives a design tool for DEM production. To achieve a target accuracy, the DEM producer can either choose an appropriate cell size for a given sensor accuracy, or determine the required sensor accuracy for a chosen resolution. In other words, it enables trade-off decisions between resolution and measurement quality.

In summary, the contribution turns a previously empirical relationship into a rigorous analytical framework, showing that DEM accuracy is governed by a structured combination of dataset error, resolution, and their interaction.

12.5. Continuing to Move Forward with DEMIX

These new contributions to the DEM intercomparison experiment do not alter the conclusions of the DEMIX report but they significantly strengthen its foundations and extend its applicability. They also point to future directions, including the incorporation of uncertainty propagation into DEM evaluation, the development of standardized field validation procedures for other derived quantities, and the continued refinement of scale-aware metrics for comparing digital terrain and surface models to field values.

13. Conclusions and Recommendations

Section contributors: Conrad Bielski, Carlos H. Grohmann, Peter Guth, Carlos López-Vázquez, Laurence Hawker, Dean Gesch, Sebastiano Trevisani, Virginia Herrera-Cruz, Serge Riazanoff, Hannes I. Reuter, Peter Strobl

Drawing conclusion from DEMIX is not a straightforward task as it tackled very different aspects of DEM assessment and comparison leading to a number of results and findings that go well beyond assessing the quality of a number of DEMs. This chapter is therefore split into four parts, beginning with highlighting the main results and how they could be achieved. Thereafter, a number of recommendations are rolled out to users and to producers. The chapter is closed with an outlook on possible future work and challenges.

13.1. Retrospect of DEMIX

Since almost a decade CEOS organises so called 'intercomparison exercises' to evaluate different approaches to the generation of higher level EO products. In most cases the focus is on methods and algorithms used to create a specific outcome, such as atmospheric correction to arrive at 'surface reflectance'. In these, different groups assume a task, usually starting from the same input, to produce a certain type of product which then will be compared. However, in some cases a lack of definition and parametrisation of the product and incomplete test results inhibit a direct quantitative comparison and consequently clear rankings of the different approaches are mostly avoided (Doxani et al., 2018, Doxani et al., 2023). DEM validations and comparisons on the other hand are ongoing since respective products exist but as outlined in section 8.3, were often hampered by limitations in geographical scope, parameters used, or amount and quality of reference data. Some authors presented rankings, but no study so far provided statistics-based confidence levels regarding these ranks.

DEMIX set out and exceeded in overcoming many of these shortfalls. The first important milestone was the in-depth discussion and convergence on a common terminology and analytical basis. This laid the foundation for an elaborated, innovative, and flexible methodological framework which spans from the harmonised formulation of benchmarking criteria to a statistically sound ranking of the tested alternatives. Because rankings require less interpretation than tables filled with numbers or comments, the results are clear and easy to convey not only to experts in the field. Rankings also allow to integrate results from quantitative and qualitative criteria. The developed framework was implemented in a detailed while large scale comprehensive study incorporating and unprecedented amount of high-quality (fiducial) reference data covering a broad range of globally relevant landforms and surface types. It does not rely on comparison of elevation alone but includes slope and roughness, two main geomorphometric parameters, all of which were retrieved from the original data without any resampling or interpolation. As a result, DEMIX could not only come forward with a clear recommendation regarding the comparative quality of all major 1-arc second free and open DEMs, it also offers full transparency on how these results were achieved, why they are considered globally relevant and how they can be tailored to specific user requests if desired.

13.2. Recommendations for Users

DEMIX was primarily designed to inform users. DEM users should consult the results of DEMIX in order to make informed choices when needing a DEM for their work. Where the presented overall result seems too general, they should use the variety of test areas and parameters ('opinions') to extract the most relevant combination for their case. With little effort they are able to investigate by their own whether and which influence their selection of tests has on the ranking. To facilitate and steer this, experienced users and experts should, e.g. for a number of common applications, create test scenarios with the most relevant metrics, landforms, and geographic areas and share these through DEMIX.

Users with requirements not well covered by the current tests are invited to propose parameters and metrics for respective tests which can then be integrated to the DEMIX database.

The DEMIX approach can accommodate quantitative or qualitative evaluations without compromising statistical rigor, thus opening the door to consider a variety of criteria with direct impact on the intended uses.

Finally, by embracing a common benchmarking and ranking system like DEMIX users can gain considerable influence on product development. Popular metrics are likely to spur the improvement of DEMs given that specific aspect and consensus on sets of relevant quality metrics will help improve DEM quality overall.

13.3. Recommendations for Producers

A major difficulty in the comparison of the various global DEMs was their varying formats, data and meta-data contents. In order to enhance the quality and interoperability of global DEMs, it is crucial for data producers to adhere to a common set of standards. This will greatly benefit both producers and users, as it streamlines the exchange and utilization of DEM data.

To promote interoperability, all global DEMs should adopt a standardized grid layout. Individual values should represent the entire grid cell in a way that avoids the risk of aliasing allowing interpolation and computation of derivatives. Grid definitions and encodings must be complete and follow ISO/OGC rules. Standardization is essential for consistent and accurate results, especially in applications that involve pixel-by-pixel comparisons.

Metadata plays a crucial role in understanding and utilizing DEM data effectively. DEM producers should incorporate vertical datum information in GeoTIFF files. While this is permitted by current standards, it is frequently omitted. Including the vertical datum ensures that users have essential information for accurate geospatial analysis. Units for the stored parameters should be reported in the GeoTIFF file. While all the global DEMs use meters for elevation, many other parameters are possible, including slope in degrees or percent. This may necessitate adjustments to the GeoTIFF standard and subsequent implementation by DEM producers. MICRODEM's approach of storing this information for various parameters can serve as a reference.

The product version should be clearly indicated in file names, and this information should be updated for all tiles in a data set. This allows users to easily identify the version they are working with, and whether it is the most up to date one, without the need to navigate through layers of metadata files. Machine readable lists and checksums should be used to easily identify files which have not been changed since the last version to avoid unnecessary download. Ensuring that this information is readily available is particularly important for maintained datasets like CopDEM.

Producers of DEM data are strongly encouraged to participate in DEMIX rounds. DEMIX serves as a neutral platform for rigorously evaluating potential products before their release. One of its key features is the ability to openly share evaluation results without revealing raw data, allowing for independent confirmation by customers. By involving users in the evaluation process, DEMIX fosters transparency and quality improvement.

It is recommended that DEM products be paired with access to high-quality, finer resolution, accurate, and multi-temporal elevation data that can be used as reference. This reference data should ideally be authoritative and adhere to prescribed quality standards and technical requirements, including the provision of extensive metadata. The combination of the DEMIX ranking approach and reference data repository can establish transparent quality standards for global DEM production and quality assessments.

Incorporating these recommendations into the production of future DEMs will not only improve the quality and utility of the data but also promote transparency, interoperability, and overall user satisfaction. This collaborative effort between producers and users can lead to significant advancements in global DEM production and assessment.

13.4. Outlook on Future Intercomparisons

To begin with, it is important to anticipate the emergence of finer-resolution global DEMs, potentially at ½" or 1/3" resolutions, going beyond CopDEM and FABDEM. We are already seeing this today and the DEMIX state-of-the-art continues to evolve, and newer comparisons have been published by authors of this report (See previous Section 12). This will require a simplified approach for evaluating new 1" DEMs by comparing them directly with current contest winners. One of the challenges in comparing DEMs of different resolutions is the variation in geomorphometric parameters. Therefore, it is necessary to acknowledge these differences and adapt the evaluation methodology accordingly. In addition to evaluating the superiority of new DEMs, it is important to consider non-inferiority tests to assess whether replacing an existing DEM with a contest-winning new DEM is worthwhile. This will provide valuable insights into the potential benefits and drawbacks of switching to a new DEM.

Uncertainty estimates for each metric used in the evaluation should be derived and validated whenever possible. These estimates will provide a solid basis for establishing ties between DEMs and will be essential for both non-inferiority tests and reception tests for new DEMs.

In terms of a future cloud based DEMIX evaluation platform, it is first and foremost important to ensure an implementation resembling entirely the functionality and results of the current exercise. The platform should be able to handle different projections and grid types, as well as changes in longitudinal spacing at higher latitudes correctly and without requiring user interference. A fully open, transparent, and repeatable workflow to incorporate new criteria, add new reference DEMs, and add new global DEM candidates should then be added. Finally, an intuitive user interface, based on standard technologies like the Jupyter notebooks, aiding in configuration and analysis of tailored rankings should attract and retain a large user community.

References

Abrams, M., Bailey, B., Tsu, H., Hato, M., *The ASTER Global DEM.* Photogrammetric Engineering and Remote Sensing 76 (4), 344ñ348, 2010.

Addelman, S., *The Generalized Randomized Block Design*. The American Statistician 23 (4), 35, 1969. http://dx.doi.org/10.2307/2681737

Alvioli M., L. Melelli, I. Marchesini eds., *Proceedings of Geomorphometry 2025*, 9-13 June, Perugia, Italy. CNR Edizioni, Rome, 2026. ISBN 978-88-8080-765-0. https://doi.org/10.30437/gmft25pg

Amerise, I. L., Tarsitano, A., *Correction methods for ties in rank correlations*. Journal of Applied Statistics 42 (12), 2584–2596, 2015. http://dx.doi.org/10.1080/02664763.2015.1043870

Becek, K., *Assessing Global Digital Elevation Models Using the Runway Method: The Advanced Spaceborne Thermal Emission and Reflection Radiometer Versus the Shuttle Radar Topography Mission Case*. IEEE Trans. Geosci. Remote Sensing, 52, 4823–4831, 2014. doi:10.1109/TGRS.2013.2285187.

Becker, D. de Andrés-Herrero, M. Willmes, C.; Weniger, G.C. & Bareth, G., *Investigating the influence of different DEMs on GIS-based cost distance modeling for site catchment analysis of prehistoric sites in Andalusia*. ISPRS International journal of geo-information, 6(2), p.36, 2017.

Benavoli, A.; Corani, G. & Mangili, F., *Should we really use post-hoc tests based on mean-ranks?*. The Journal of Machine Learning Research, 17(1), pp.152-161, 2016.

Bian, G.; McAleer, M. & Wong, W.K., *A trinomial test for paired data when there are many ties*. Mathematics and Computers in Simulation, 81(6), pp.1153-1160, 2011.

Bielski, C.; López-Vázquez, C.; Grohmann, C.H.; Guth, P.L.; Hawker, L.; Gesch, D.; Trevisani, S.; Herrera-Cruz, V.; Riazanoff, S.; Corseaux, A. & Reuter, H.I., Novel approach for ranking DEMs: Copernicus DEM improves one arc second open global topography. IEEE transactions on geoscience and remote sensing, 62, pp.1-22, 2024. doi: 10.1109/TGRS.2024.3368015.

Brock, J.; Schratz, P.; Petschko, H.; Muenchow, J.; Micu, M. & Brenning, A., *The performance of landslide susceptibility models critically depends on the quality of digital elevation models*. Geomatics, Natural Hazards and Risk, 11(1), pp.1075-1092, 2020.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., Smets,B., *Copernicus Global Land Cover Layers—Collection 2*. Remote Sensing 12 (6), 1044, https://doi.org/10.3390/rs12061044 ,2020.

Burrough PA, McDonnel RA, *Principles of Geographical Information Systems*. Oxford University Press, Oxford, 2000.

Carabajal, C. C. and Harding, D. J., *ICESat validation of SRTM C-band digital elevation models*. Geophysical Research Letters 32 (22), L22S01, 2005.

Conover, W. J., *Practical Nonparametric Statistics 2nd. Edition*, Wiley, 462 pp, 1980.

Corder G. W. and Foreman D. I., *Nonparametric Statistics for Non-Statisticians: A Step-By-Step Approach*. Wiley, 2014.

Corder, G. W. & Foreman, D. I., *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach.* John Wiley & Sons, 2009.

Courty, L.G.; Soriano-Monzalvo, J.C. & Pedrozo-Acuña, A., *Evaluation of open-access global digital elevation models (AW3D30, SRTM, and ASTER) for flood modelling purposes*. Journal of Flood Risk Management, 12, p.e12550, 2019.

Crippen, R., Buckley, S., Agram, P., Belz, E., Gurrola, E., Hensley, S., Kobrick, M., Lavalle, M., Martin, J., Neumann, M., Nguyen, Q., Rosen, P., Shimada, J., Simard, M., Tung, W., *NASADEM Global Elevation Model: Methods and Progress*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLI-B4, 125ñ128, 2016.

Crosetto, M., Solari, L., Balasis-Levinsen, J., Casagli, N., Frei, M., Oyen, A., and Moldestad, D. A., *Ground Deformation Monitoring at Continental Scale: The European Ground Motion Service.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2020, 293ñ298. https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-293-2020 , 2020.

Daniel, W. W. & Cross, C.L., *Biostatistics: a foundation for analysis in the health sciences.* Wiley, 10th edition, 2013.

Di Gregorio, A., and Jansen, L. J. M., *Land Cover Classification System (LCCS): Classification Concepts and User Manual.* FAO, Rome, 2000.

Dobre, B.; Kovács, I. P, & Bugya, T., *Comparison of digital elevation models through the analysis of geomorphic surface remnants in the Desatoya Mountains, Nevada*. Transactions in GIS, 25, 2262–2282, 2021. doi:10.1111/tgis.12819

Doxani, G.; Vermote, E.; Roger, J.-C.; Gascon, F.; Adriaensen, S.; Frantz, D.; Hagolle, O.; Hollstein, A.; Kirches, G.; Li, F.; et al. *Atmospheric Correction Inter-Comparison Exercise*. Remote Sens. 2018, 10, 352. https://doi.org/10.3390/rs10020352

Doxani, G. E. Vermote, J-C Roger, S. Skakun, F. Gascon, A. Collison, L. De Keukelaere, C. Desjardins, et al. , *Atmospheric Correction Inter-comparison eXercise, ACIX-II Land: An assessment of atmospheric correction processors for Landsat 8 and Sentinel-2 over land*, Remote Sensing of Environment, Volume 285, 2023, 113412, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2022.113412

Dunn, O.J., *Multiple Comparisons among Means*. Jour. of the Am. Statistical Association 56,52-64, 1961. https://doi.org/10.2307/2282330

Ebinne, E.S.; Apeh, O.I.; Moka, E.C. & Abah, E.J., *Comparative analysis of freely available digital elevation models for applications in multi-criteria environmental modeling over data limited regions*. Remote Sensing Applications: Society and Environment, 27, p.100795, 2022.

Embarcadero, *Delphi & C++Builder Community Edition FAQs*, 2023, https://www.embarcadero.com/products/delphi/starter/faq

Evans, I. S., *An integrated system of terrain analysis and slope mapping*. Zeitschrift für Geomorphologie Suppl.-Bd. 36, 274–295, 1980.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., *The Shuttle Radar Topography Mission*. Review of Geophysics 45, RG2004, 2007.

Federal Geographic Data Committee (FGDC), *Geospatial Positioning Accuracy Standards: Part 3, National Standard for Spatial Data Accuracy*. Washington, D.C., Federal Geographic Data Committee Report No STD-007.3-1998, 1998.

Fey, M., Clarke, K. A., *Consistency of choice in nonparametric multiple comparisons*. Journal of Nonparametric Statistics 24 (2), 531–541, 2012. http://dx.doi.org/10.1080/10485252.2012.675436

Fijałkowska, A., *Analysis of the Influence of DTM Source Data on the LS Factors of the Soil Water Erosion Model Values with the Use of GIS Technology*. Remote Sensing. 13(4):678, 2021. doi:10.3390/rs13040678

Florinsky, I. V., *Computation of the third-order partial derivatives from a digital elevation model.* International Journal of Geographical Information Science, 23 (2), 213-231. https://doi.org/10.1080/13658810802527499 , 2009.

Florinsky, I. V.; Skrypitsyna, T. N.; Trevisani, S. and Romaikin, S. V., *Statistical and visual quality assessment of nearly-global and continental digital elevation models of Trentino, Italy*, Remote Sensing Letters, 10:8, 726-735, 2019. doi:10.1080/2150704X.2019.1602790

Franks, S. & Rengarajan, R., *Evaluation of Copernicus DEM and Comparison to the DEM Used for Landsat Collection-2 Processing*. Remote Sens. 15, 2509, 2023. https://doi.org/10.3390/rs15102509

Friedman, M., *The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance*. Journal of the American Statistical Association 32, 675-701, 1937. https://doi.org/10.2307/2279372

Gdulová, K.; Marešová, J. & Moudrý, V., *Accuracy assessment of the global TanDEM-X digital elevation model in a mountain environment*. Remote Sensing of Environment, 241, p.111724, https://doi.org/10.1016/j.rse.2020.111724 ,2020.

Gesch, D. B., Verdin, K. L., and Greenlee, S. K., *New land surface digital elevation model covers the Earth. Eos,* Transactions of the American Geophysical Union, 80(6), 69ñ70. https://doi.org/10.1029/99EO00050 , 1999.

Gesch, D. B., Muller, J.-P., & Farr, T. G., *The Shuttle Radar Topography Mission: Data Validation and Applications.* Photogrammetric Engineering & Remote Sensing, 72(3), 233ñ235, 2006.

Gesch, D. B., Oimoen, M. J., Greenlee, S. K., Nelson, C. A., Steuck, M. J., & Tyler, D. J., *The National Elevation Dataset.* Photogrammetric Engineering & Remote Sensing, 80(12), 1075-1091, 2014.

Gesch, D. B., *Best Practices for Elevation-Based Assessments of Sea-Level Rise and Coastal Flooding Exposure*. Frontiers in Earth Science, 6, Article 230. https://doi.org/10.3389/feart.2018.00230 , 2018.

Gesch, D. B., *Assessing Global Elevation Models for Mapping the Low Elevation Coastal Zone.* In Geomorphometry 2023 Proceedings, Geomorphometry 2023 Conference, Iasi, Romania, 10ñ14 July 2023. International Society for Geomorphometry. https://doi.org/10.5281/zenodo.8011577 2023.

Rodriguez, E., Morris, C. H., Belz, J. E., 2006. A global assessment of the SRTM performance. Photogrammetric Engineering and Remote Sensing 72, 249ñ260.

González-Moradas, M. del R. & Viveen, W., *Evaluation of ASTER GDEM2, SRTMv3.0, ALOS AW3D30 and TanDEM-X DEMs for the Peruvian Andes against highly accurate GNSS ground control points and geomorphological-hydrological metrics*, Remote Sensing of Environment, Volume 237, 111509, ISSN 0034-4257, 2020. doi:10.1016/j.rse.2019.111509

González-Moradas, M. del R.; Viveen, W.; Vidal-Villalobos, R.A. & Villegas-Lanza, J.C., *A performance comparison of SRTM v. 3.0, AW3D30, ASTER GDEM3, Copernicus and TanDEM-X for tectonogeomorphic analysis in the South American Andes*. Catena, 228, p.107160, 2023.

Google, *Get insights quickly, with Google Forms*: https://www.google.com/forms/about/ Accessed 12 June 2023.

Goudie, A., *Geomorphological techniques*. Routledge, 2003.

Grohmann, C. H., López-Vázquez, C., Guth, P. L., & Bielski, C., *The DEMIX Wine Contest: a summary*. Proceedings of the Geomorphometry2023 Conference, 2023 (a). https://doi.org/10.5281/zenodo.8066530

Grohmann, C.H., *The DEMIX Wine Contest Jupyter notebook*. Proceedings of the Geomorphometry2023 Conference, 2023 (b). https://doi.org/10.5281/zenodo.8169460

Grohmann, C. H., Smith, M. J., Riccomini, C., *Multiscale Analysis of Topographic Surface Roughness in the Midland Valley, Scotland*. Geoscience and Remote Sensing, IEEE Transactions on 49 (4), 1200–1213, 2011. http://dx.doi.org/10.1109/TGRS.2010.2053546

Grohmann, C.H., *Comparative Analysis of Global Digital Elevation Models and Ultra-Prominent Mountain Peaks*. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., III–4, 17–23, 859, 2016. doi:10.5194/isprsannals-III-4-17-2016.

Grohmann, C.H., Steiner, S.S., *SRTM resample with short distance-low nugget kriging.* International Journal of Geographical Information Science, 22:8, 895-906, 2008. DOI:10.1080/13658810701730152

Grohmann, C.H., *Evaluation of TanDEM-X DEMs on selected Brazilian sites: comparison with SRTM, ASTER GDEM and ALOS AW3D30*. Remote Sensing of Environment, 212C:121-133, 2018.

Guth, P. L., Ressler, E.K., and Bacastow, T.S., *Microcomputer program for manipulating large digital terrain models.* Computers & Geosciences, vol.13, no.3, p.209-213, 1987.

Guth, P. L., *Geomorphometry from SRTM: Comparison to NED.* Photogrammetric Engineering & Remote Sensing, special issue based on Shuttle Radar Topography Mission—Data Validation and Applications Workshop, Reston, VA, 14 June 2005, vol.72, no.3, p.269-277, 2006.

Guth, P. L., *Geomorphometry in MICRODEM*, In Hengl, T., Reuter, H.I. (eds), Geomorphometry: concepts, software, applications. Developments in Soil Science Series, Elsevier, ISBN-13: 978-0-12-374345-9, p.351-366, 2009.

Guth, P. L., *Geomorphometric Comparison of ASTER GDEM and SRTM.* ASPRS/CaGIS 2010 Fall Specialty Conference, Orlando, FL, November 15-19, 10 p, 2010. http://www.asprs.org/a/publications/proceedings/orlando2010/files/Guth.pdf

Guth P. L., *What should a bare earth digital terrain model (DTM) portray?* PeerJ Preprints 6:e27053v1, 2018a. https://doi.org/10.7287/peerj.preprints.27053v1

Guth, P. L., *Castles in the Clouds: LiDAR for Historical Study and Terrain Analysis.* Scientia Militaria - South African Journal of Military Studies, Vol 46, No 1, p.79-95, 2018b. https://doi.org/10.5787/46-1-1226

Guth, P. L., Geoffroy, T. M., *LiDAR point cloud and ICESat-2 evaluation of 1 second global digital elevation models: Copernicus wins.* Transactions in GIS, 2245–2261, 2021a.

Guth, P. L., Kane, M., *Slope, aspect, and hillshade algorithms for non-square digital elevation models*. Transactions in GIS 25 (5), 2309ñ2332. URL https://doi.org/10.1111/tgis.12852 , 2021b.

Guth, P. L., Van Niekerk, A., Grohmann, C. H., Muller, J., Hawker, L., Florinsky, I. V., Gesch, D., Reuter, H. I., Herrera-Cruz, V., Riazanoff, S., López-Vázquez, C., Carabajal, C. C., Albinet, C., Strobl, P., *Digital Elevation Models: Terminology and Definitions*. Remote Sensing. 13 (18), 3581, 2021c. http://dx.doi.org/10.3390/rs13183581

Guth, P. L., Strobl, P., Gross, K. and Riazanoff, S., *DEMIX 10k Tile Data Set (1.0)* [Data set]. Zenodo, 2023a. https://doi.org/10.5281/zenodo.7504791

Guth, P. L., *DEMIX GIS Database Version 2 (2.0)* [Data set]. Zenodo, 2023b. https://doi.org/10.5281/zenodo.8062008

Guth, P. L., *DEMIX Difference Distribution Histograms (1.0)* [Data set]. Zenodo, 2023c. https://doi.org/10.5281/zenodo.8087804

Guth, P. L., Riazanoff, S., Corseaux, A., Grohmann, C., Trevisani, S., and López-Vázquez, C., *DEMIX 1" Reference DEMs version 2.0 (2.0)* [Data set]. Zenodo, 2023d. https://doi.org/10.5281/zenodo.8086806

Guth, P. L., Grohmann, C.H., Trevisani, S., *Subjective Criterion for the DEMIX Wine Contest: Hillshade Maps*. Proceedings of the Geomorphometry2023 Conference, 2023e. https://doi.org/10.5281/zenodo.8030735

Guth, P. L., Grohmann, C.H., Trevisani, S., *Geomorphometric and Geospatial Patterns in Differences Between ALOS and CopDEM.* Geomorphometry2023. Proceedings of the Geomorphometry2023 Conference, 2023f. https://doi.org/10.5281/zenodo.8057703

Guth, P. L., Grohmann, C. H., Trevisani, S., and Bielski, C., *Going Forward:. Extending and Exploiting the DEMIX Database With MICRODEM*. Proceedings of the Geomorphometry2023 Conference, 2023g. https://doi.org/10.5281/zenodo.8092218

Guth, P. L., *GIT-MICRODEM* [Delphi source code, archived installation versions], 2023h. https://github.com/prof-pguth/git_microdem . last access: 2023-02-10.

Guth, P. L., *MICRODEM: Open-source GIS with a focus on Geomorphometry* [download latest Win64 executable and CHM help file] URL, 2023i. https://microdem.org/

Guth, P. L.; Trevisani, S.; Grohmann, C.H.; Lindsay, J.; Gesch, D.; Hawker, L.; Bielski, C. *Ranking of 10 Global One-Arc-Second DEMs Reveals Limitations in Terrain Morphology Representation*. Remote Sens., 16, 3273, 2024. https://doi.org/10.3390/rs16173273

Guth, P. L.; Trevisani, S.; Grohmann, C.H.; Lindsay, J.B.; Reuter, H.I., *Benchmarking Elevation Plus Land Surface Parameters Finds FathomDEM and Copernicus DEM Win as Best Global DEMs*. Remote Sens., 17, 3919, 2025. https://doi.org/10.3390/rs17233919

Guth, P., *Key Characteristics of Global DEMIX Tiles* [Data set]. Zenodo, 2026. https://doi.org/10.5281/zenodo.18331917

Gross, K., Corseaux, A., Riazanoff, S., and Albinet, C., *Copernicus DEMs Quality Assessment Summary.* EDAP.REP.039, Issue 1.2, European Space Agency (ESA), 29 July 2021.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T. E., *Array programming with numpy*. Nature 585 (7825), 357–362, 2020. https://doi.org/10.1038/s41586-020-2649-2

Hasan, R. H., *Evaluation of the Accuracy of Digital Elevation Model Produced from Different Open Source Data.* Journal of Engineering, 25(8), pp. 100–112, 2019. doi:10.31026/j.eng.2019.08.07

Hawker, L., Uhe, P., Paulo, L., Sosa, J., Savage, J., Sampson, C., Neal, J., *A 30 m global map of elevation with forests and buildings removed.* Environmental Research Letters 17 (2), 024016, 2022.

Hayakawa, Y. S., Oguchi, T. and Lin, Z., *Comparison of new and existing global digital elevation models: ASTER G-DEM and SRTM-3*. Geophysical Research Letters, 35, L17404. https://doi.org/10.1029/2008GL035036 , 2008.

Hengl, T., *Finding the right pixel size*. Computers & Geosciences, Volume 32, Issue 9, 1283-1298, 2006.

Heuvelink, G.B.; Burrough, P.A. & Stein, A., *Propagation of errors in spatial modelling with GIS*. International Journal of Geographical Information System, 3(4), pp.303-322, 1989.

Hnila, P. & Elicker, J., *Quality Assessment of Digital Elevation Models in a Treeless High-Mountainous Landscape. A Case Study from Mount Aragats.* Armenia. Magazén, 2(1), 71-102, 2021. doi:10.30687/mag/2724-3923/2021/03/003

Ho, Y.-F., Hengl, T., and Parente, L., *Ensemble Digital Terrain Model (EDTM) of the World* (Version 1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7676373 , 2023.

Ho Y, Grohmann CH, Lindsay J, Reuter HI, Parente L, Witjes M, Hengl T., *GEDTM30: global ensemble digital terrain model at 30 m and derived multiscale terrain variables*. PeerJ 13:e19673, 2025. https://doi.org/10.7717/peerj.19673

Hochberg, Y., *A sharper Bonferroni procedure for multiple tests of significance.* Biometrika 75, 800–803, 1988.

Hofton, M.; Dubayah, R.; Blair, J.B. & Rabine, D., *Validation of SRTM elevations over vegetated and non-vegetated terrain using medium footprint lidar*. Photogrammetric Engineering & Remote Sensing, 72(3), pp.279-285, 2006.

Holm, S., *A simple sequentially rejective multiple test procedure.* Scandinavian Journal of Statistics 6, 65–70, 1979.

Hommel, G., *A stagewise rejective multiple test procedure based on a modified Bonferroni test.* Biometrika 75, 383–386, 1988.

Hunter, J.D., *Matplotlib: A 2D graphics environment.* Computing. Science & Engineering 9(3),90-95, 2007. https://doi.org/10.1109/MCSE.2007.55

Iman, R. L. & Davenport, J. M., *Approximations of the critical region of the Friedman statistic.* Communications in Statistics - Theory and Methods, 9:6, 571-595, 1980. DOI: 10.1080/03610928008827904

Isaaks E H, Srivastava RM, *An Introduction to Applied Geostatistics.* Oxford University Press, New York, 1989.

Iwahashi, J. & Pike, R.J., *Automated classifications of topography from DEMs by an unsupervised nested-means algorithm and a three-part geometric signature.* Geomorphology, 86(3-4), pp.409-440, 2007.

Jasiewicz, J. & Stepinski, T.F., *Geomorphons—a pattern recognition approach to classification and mapping of landforms*. Geomorphology, 182, pp.147-156, 2013.

Józsa, E.; Fábián, S. Á. & Kovács, M., *An evaluation of EU-DEM in comparison with ASTER GDEM, SRTM and contour-based DEMs over the Eastern Mecsek Mountains*, Hungarian Geographical Bulletin, 63(4), pp. 401-423, 2014. doi:10.15201/hungeobull.63.4.3

Karlson, M.; Bastviken, D. & Reese, H., *Error characteristics of pan-arctic digital elevation models and elevation derivatives in Northern Sweden*. Remote Sensing, 13(22), p.4653, 2021.

Kluyver T., Ragan-Kelley B., Pérez F., Granger B.E., Bussonnier M., Frederic J., Kelley K., Hamrick J.B., Grout J., Corlay S. et al., *Jupyter notebooks - a publishing format for reproducible computational workflows*. In: Loizides F., Scmidt B. (eds) Positioning and Power in Academic Publishing: Players, Agents and Agenda. IOS Press, pp 87–90, 2016. http://dx.doi.org/10.3233/978-1-61499-649-1-87

Knuth, D.E., *Literate Programming*, The Computer Journal, 27:2, 97–111, 1984. https://doi.org/10.1093/comjnl/27.2.97

Kramm, T. & Hoffmeister, D., *Comprehensive vertical accuracy analysis of freely available DEMs for different landscape types of the Ruhr catchment, Germany*, Geocarto International, https://doi.org/10.1080/10106049.2021.1984588 ,2022.

Lehmann, E.L., *Nonparametrics: Statistical Methods Based on Ranks*, NewYork: Springer Science, 2006.

López-Vázquez, C., *Uncertainty interval estimates for computing slope and aspect from a gridded digital elevation model.* International Journal of Geographical Information Science, 36(8), pp.1601-1628, 2022.

López-Vázquez, C. & Ariza-López, F.J., *Global Digital Elevation Model Comparison Criteria: An Evident Need to Consider Their Application*. ISPRS Int. J. Geo-Inf. 12, 337, 2023. doi:10.3390/ijgi12080337

López-Vázquez, C. and Hochsztain, E., *Extended and updated tables for the Friedman rank test.* Communications in Statistics-Theory and Methods, 48(2), pp.268-281, 2019.

López-Vázquez, C. and Hochsztain, E., *Tables for the Friedman's Test with ties*, 2020. Last access 22-Jun-2023. https://dspace.ort.edu.uy/handle/20.500.11968/4208

López Vázquez, C., and Méndez-Baillo, R., A new field-based approach to satisfying NSSDA accuracy requirements in digital slope model assessment. Revista Cartográfica, (113), 159–178, 2026. https://doi.org/10.35424/rcarto.i113.6207

López-Vázquez, C., Tasistro, A., Hochsztain, E., *Exact tables for the Friedman rank test: Case with ties*. Chilean Journal of Statistics (ChJS) 12 (1), 2021. http://soche.cl/chjs/volumes/12/ChJS-12-01-02.pdf

McKinney, W., *pandas: a Foundational Python Library for Data Analysis and Statistics.* In: Python for High Performance and Scientific Computing, 2011. https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf

Meybeck, M., Green, P., Vörösmarty, C., *A New Typology for Mountains and Other Relief Classes*. Mountain Research and Development 21 (1), 34–45, 2001. http://dx.doi.org/10.1659/0276-4741(2001)021[0034:ANTFMA]2.0.CO;2

Mouratidis, A. and Ampatzidis, D., *European Digital Elevation Model Validation against Extensive Global Navigation Satellite Systems Data and Comparison with SRTM DEM and ASTER GDEM in Central Macedonia (Greece)*. ISPRS International Journal of Geo-Information. 2019; 8(3):108, 2019. doi:10.3390/ijgi8030108

Nadi, S.; Shojaei, D. and Ghiasi, Y., *Accuracy assessment of DEMs in different topographic complexity based on an optimum number of GCP formulation and error propagation analysis.* Journal of Surveying Engineering, 146(1), p.04019019, 2020.

Nemenyi, P., *Distribution-free multiple comparisons*. Ph.D. thesis, Princeton University, 1963.

Oksanen, J. and Sarjakoski, T., *Error propagation of DEM-based surface derivatives.* Computers & Geosciences, 31(8), pp.1015-1027, 2005.

Parolini, G., *The Emergence of Modern Statistics in Agricultural Science: Analysis of Variance, Experimental Design and the Reshaping of Research at Rothamsted Experimental Station, 1919–1933*. J Hist Biol 48, 301–335, 2015. doi:10.1007/s10739-014-9394-z

Pereira, D. G., Afonso, A., Medeiros, F. M., *Overview of Friedman's Test and Post-hoc Analysis*. Communications in Statistics Simulation and Computation 44 (10), 2636–2653, 2014.

Pimentel, J.F., Murta, L., Braganholo, V. et al., *Understanding and improving the quality and reproducibility of Jupyter notebooks*. Empirical Software Engineering 26:65, 2021. https://doi.org/10.1007/s10664-021-09961-9

Pratt, J.W. & Gibbons, J.D., *One-Sample and Paired-Sample Inferences Based on Signed Ranks.* In: Concepts of Nonparametric Theory. Springer Series in Statistics. Springer, New York, NY, 1981. doi:10.1007/978-1-4612-5931-2_3

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P., *Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3rd Edition*. Cambridge University Press, 2007. http://numerical.recipes/

Purinton, B. and Bookhagen, B., *Beyond vertical point accuracy: Assessing inter-pixel consistency in 30 m global DEMs for the arid central Andes*. Frontiers in Earth Science, 9, p.758606, 2021.

Python Software Foundation, *Python Programming Language, version 3.10.x,* 2021. http://www.python.org/

qgrid python library. 2020. https://github.com/lukewys/qgrid

Quade, D., *Using weighted rankings in the analysis of complete blocks with additive block effects*, Journal of the American Statistical Association, 74, 680–683, 1979.

Rabby, Y.W.; Ishtiaque, A. and Rahman, M.S., *Evaluating the Effects of Digital Elevation Models in Landslide Susceptibility Mapping in Rangamati District, Bangladesh*. Remote Sensing. 12(17):2718, 2020. doi:10.3390/rs12172718

Rawat, K.S.; Kumar, S.; Mishra, A.K. and Singh, S.K., *Assessing the Accuracy of Open Source Altitude Data for the Hilly Area in Tehri Garhwal District of Uttarakhand, India.* In: Kolhe, M.L., Jaju, S.B., Diagavane, P.M. (eds) Smart Technologies for Energy, Environment and Sustainable Development, Vol 2. Springer Proceedings in Energy. Springer, Singapore, 2022. doi:10.1007/978-981-16-6879-1_17

Reuter, H. I., Nelson, A. and Jarvis, A., *An evaluation of void-filling interpolation methods for SRTM data.* International Journal of Geographical Information Science, 21(9), 983ñ1008, 2009.

Rexer, M. & Hirt, C., *Comparison of free high-resolution digital elevation data sets (ASTER GDEM2, SRTM v2.1/v4.1) and validation against accurate heights from the Australian National Gravity Database.* Australian Journal of Earth Sciences, pp 1-15, 2014. doi:10.1080/08120099.2014.884983

Riazanoff, S.; Corseaux, A.; Albinet, C.; Strobl, P.A.; López-Vázquez, C.; Guth, P.L.; Tadono, T. *Best BiCubic Method to Compute the Planimetric Misregistration between Images with Sub-Pixel Accuracy: Application to Digital Elevation Models.* ISPRS Int. J. Geo-Inf. 2024, 13, 96. https://doi.org/10.3390/ijgi13030096

Rizzoli, P., Martone, M., Gonzalez, C., Wecklich, C., Tridon, D. B., Brautigam, B., Bachmann, M., Schulze, D., Fritz, T., Huber, M., Wessel, B., Krieger, G., Zink, M., Moreira, A., *Generation and performance assessment of the global TanDEM-X digital elevation model.* ISPRS Journal of Photogrammetry and Remote Sensing 132, 119-139, 2017.

Romshoo, S. A.; Yousuf, A.; Altaf, S. & Amin, M., *Evaluation of Various DEMs for Quantifying Soil Erosion Under Changing Land Use and Land Cover in the Himalaya*. Front. Earth Sci. 9:782128, 2021. doi:10.3389/feart.2021.782128

Rule A, Birmingham A, Zuniga C, Altintas I, Huang S-C, Knight R, et al., *Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks*. PLoS Comput Biol 15(7): e1007007, 2019. https://doi.org/10.1371/journal.pcbi.1007007

Shafique, M. and van der Meijde, M., *Impact of uncertainty in remote sensing DEMs on topographic amplification of seismic response and Vs 30*. Arabian Journal of Geosciences, 8(4), pp.2237-2245, 2015.

Shen, H., *Interactive notebooks: Sharing the code*. Nature 515, 151–152, 2014. https://doi.org/10.1038/515151a

Siegel, S., *Nonparametric statistics for the Behavioral Sciences*. Mc Graw-Hill Book Co., 1956.

Siegel, S. & Castellan, N. J., Jr., *Nonparametric statistics for the behavioral sciences (2nd ed.)*. New York, NY, England: Mcgraw-Hill Book Company, 1988.

Slater, J. A., Heady, B., Kroenung, G., Curtis, W., Haase, J., Hoegemann, D., Shockley, C. and Tracy, K., *Global assessment of the new ASTER Global Digital Elevation Model.* Photogrammetric Engineering & Remote Sensing, 77(4), 335ñ349, 2011.

Strobl, P., *The new Copernicus digital elevation model.* GSICS Quarterly 14 (1), 2020.

Strobl, P.A., Bielski, C., Guth, P.L., Grohmann, C.H., Muller, J.P., López-Vázquez C., Gesch, D.B., Amatulli, G., Riazanoff, S. and Carabajal, C., *The digital elevation model intercomparison experiment DEMIX, a community-based approach at global DEM benchmarking.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 43(B4), pp.395-400, 2021.

Szumska, E.; Jurecki, R. and Pawelczyk, M., *Assessment of Total Costs of Ownership for Midsize Passenger Cars with Conventional and Alternative Drive Trains*. Communications - Scientific Letters of the University of Zilina, 21(3), 21-27, 2019. doi: 10.26552/com.C.2019.3.21-27

Tadono, T., Nagai, H., Ishida, H., Oda, F., Naito, S., Minakawa, K., Iwamoto, H., *Generation of the 30 M-Mesh Global Digital Surface Model by ALOS PRISM.* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLI-B4, 157ñ162, 2016.

Tavares da Costa, R.; Mazzoli, P. and Bagli, S., *Limitations Posed by Free DEMs in Watershed Studies: The Case of River Tanaro in Italy*. Front. Earth Sci. 7:141, 2019. doi:10.3389/feart.2019.00141

Toothaker, L. E. and Newman, D., *Nonparametric Competitors to the Two-Way ANOVA.* Journal of Educational Statistics, 19(3), 237–273, 1994. https://doi.org/10.3102/10769986019003237

Trevisani, S., Skrypitsyna, T.N., Florinsky, I.V., *Global digital elevation models for terrain morphology analysis in mountain environments: insights on Copernicus GLO-30 and ALOS AW3D30 for a large Alpine area.* Environmental Earth Sciences 82. https://doi.org/10.1007/s12665-023-10882-7 , 2023.

Trevisani, S.; Teza, G. & Guth, P.L., *A simplified geostatistical approach for characterizing key aspects of short-range roughness.* Catena, 223, p.106927, 2023.

Trevisani, S.; Guth, P.L. *Surface Roughness in Geomorphometry: From Basic Metrics Toward a Coherent Framework*. Remote Sens., 17, 3864, 2025. https://doi.org/10.3390/rs17233864

Uhe, P., Lucas, C., Hawker, L., Brine, M., Wilkinson, H., Cooper, A., & Sampson, C., *FathomDEM: an improved global terrain map using a hybrid vision transformer model*. Environmental Research Letters, 20(3), 034002, 2025. https://doi.org/10.1088/1748-9326/ada972

Uuemaa, E., Ahi, S., Montibeller, B., Muru, M., and Kmoch, A., *Vertical Accuracy of Freely Available Global Digital Elevation Models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM).* Remote Sensing, 12(21), 3482. https://doi.org/10.3390/rs12213482 , 2020.

Waskom, M.L., *seaborn: statistical data visualization*. Journal of Open Source Software, 6(60), 3021, 2021. https://doi.org/10.21105/joss.03021

Watson, C.S.; Kargel, J.S. & Tiruwa, B., *UAV-Derived Himalayan Topography: Hazard Assessments and Comparison with Global DEM Products*. Drones. 3(1):18, 2019. doi:10.3390/drones3010018

Wessel, B., *TanDEM-X Ground Segment ñ DEM Products Specification Document.* Tech. rep., EOC, DLR, public Document TD-GS-PS-0021, Issue 3.1. URL https://tandemx-science.dlr.de/ , 2016.

Yamazaki, D.; Ikeshima, D.; Neal, J.C.; O'Loughlin, F.; Sampson, C.C.; Kanae, S. & Bates, P.D., *MERIT DEM: A new high-accuracy global digital elevation model and its merit to global hydrodynamic modeling.* In AGU fall meeting abstracts (Vol. 2017, pp. H12C-04), 2017.

Young, A.; Brunsden, D. and Thornes, J.B., *Slope profile survey.* BGRG Tech.Bull. 11. Norwich: Geo Abstracts, University of East Anglia, 1974.

Zakeri, F. and Mariethoz, G., *A review of geostatistical simulation models applied to satellite remote sensing: Methods and applications.* Remote Sensing of Environment, 259, p.112381, https://doi.org/10.1016/j.rse.2021.112381L ,2021.

Zhang, K., Gann, D., Ross, M., Robertson, Q., Sarmiento, J., Aguilar de Santana, S., Rhome, J., and Fritz, C., *Accuracy assessment of ASTER, SRTM, ALOS, and TDX DEMs for Hispaniola and implications for mapping vulnerability to coastal flooding.* Remote Sensing of Environment, 225, 290ñ306. https://doi.org/10.1016/j.rse.2019.02.028 , 2019.

Zhidkin, A.P.; Golosov, V.N. and Dob, A.S., *Evaluation of applicability of digital terrain models for soil erosion modeling (on the example of a small catchment area in the Kursk region) (in russian).* Modern problems of remote sensing of the Earth from space. 2021. Vol. 18. No. 5. pp. 133–144, 2021.

Zingaro, M.; La Salandra, M.; Colacicco, R.; Roseto, R.; Petio, P. and Capolongo, D., *Suitability assessment of global, continental and national digital elevation models for geomorphological analyses in Italy*. Transactions in GIS, 25, 2283– 2308, 2021. doi:10.1111/tgis.12845

Zylshal, Z.; Bayanuddin, A.A.; Nugroho, F.S. and Munawar, S.T.A., *Correcting the Topographic Effect on SPOT-6/7 multispectral imageries: a comparison of different Digital Elevation Models*. Geographia Technica, Vol. 16, Special Issue, pp 163 to 179, 2021. doi:10.21163/GT_2021.163.13

List of abbreviations and definitions

Abbreviations	Definitions
AI	Artificial Intelligence
ALOS	Advanced Land Observing Satellite
ARD	Analysis Ready Data
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
ASTER GDEM	ASTER Global Digital Elevation Model
AVD	Average Deviation
AW3D30	ALOS World 3D - 30m global digital surface model
CEOS	Committee on Earth Observation Satellites
CI	Confidence Interval
COG	Cloud Optimized GeoTIFF
CopDEM	Copernicus DEM
CRS	Coordinate Reference System
DEM	Digital Elevation Model
DGED	Defence Gridded Elevation Data
DoD	U.S. Department of Defence
DTED	Digital Terrain Elevation Model
DTM	Digital Terrain Model
DSM	Digital Surface Model
DEMIX	Digital Elevation Model Intercomparison eXercise
EC	European Commission

Abbreviations	Definitions
EDAP	Earthnet Data Assessment Pilot
EGM2008	Earth Gravitational Model 2008
EGM96	Earth Gravitational Model 1996
EO	Earth Observation
ESA	European Space Agency
ETRS89	European Terrestrial Reference System 1989
EU	European Union
EU-DEM	European Digital Elevation Model
EuroGEO	Europe's part of the Group on Earth Observations (GEO)
FABDEM	Forest And Building removed Copernicus DEM
GDAL	Geospatial Data Abstraction Library
GEDI	Global Ecosystem Dynamics Investigation
GEO	Group on Earth Observations
GEOSS	Global Earth Observation System of Systems
GNSS	Global Navigation Satellite System
GSD	Ground Sampling Distance
GUI	Graphical User Interface
ICEsat	Ice, Cloud, and land Elevation Satellite
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organisation for Standardisation
JRC	Joint Research Centre
LAEA	Lambert Azimuthal Equal-Area

Abbreviations	Definitions
LE90	Linear Error at 90% confidence
LE95	Linear Error at 95% confidence
LIDAR	Light Detection and Ranging
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MERIT DEM	Multi-Error-Removed Improved-Terrain DEM
MICRODEM	Freeware microcomputer mapping program written by Professor Peter Guth of the Oceanography Department, U.S. Naval Academy
ML	Machine Learning
NSSDA	National Standard for Spatial Data Accuracy
OSGeo4W	A binary distribution of a broad set of open source geospatial software for Windows environments
NAD83	North American Datum of 1983
NASADEM	A modernization of the DEM and associated products generated from the Shuttle Radar Topography Mission data
RCBD	Randomized Complete Block Design
RMSE	Root Mean Square Error
SRTM	Shuttle Radar Topography Mission
STD	Standard Deviation
TPI	Topographic Position Index
TRI	Topographic Ruggedness Index
USGS	United States Geological Survey
UTM	Universal Transverse Mercator

Abbreviations	Definitions
VRS	Vertical Reference System
WG	Working Group
WGCV	Working Group on Calibration & Validation
WGISS	Working Group on Information Systems and Services
WGS84	World Geodetic System 1984

List of boxes

List of figures

List of tables

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

— by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
— at the following standard number: +32 22999696,
— via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/write-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

Open data from the EU

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society

Scan the QR code to visit:

Joint Research Centre
https://joint-research-centre.ec.europa.eu

Publications Office
of the European Union