Historical Images for Surface Topography Reconstruction Intercomparison eXperiment (HistoriX)

Amaury Dehecq[1], Friedrich Knuth[2,3], Joaquín M.C. Belart[4], Livia Piermattei[5], Camillo Ressl[6], Adina Racoviteanu[1], Robert McNabb[7], David Shean[8], Luc Godin[1]

[1] University Grenoble Alpes, IRD, CNRS, Grenoble INP, IGE, Grenoble, France – amaury.dehecq@univ-grenoble-alpes.fr
[2] Laboratory of Hydraulics, Hydrology and Glaciology (VAW), ETH Zurich, Zurich, Switzerland
[3] Swiss Federal Institute for Forest, Snow and Landscape Research (WSL), Birmensdorf, Switzerland
[4] Natural Science Institute of Iceland, Akranes, Iceland
[5] Department of Geography, University of Zurich, Zurich, Switzerland
[6] TU Wien, Department of Geodesy and Geoinformation, Vienna, Austria
[7] School of Geography and Environmental Sciences, Ulster University, Coleraine, UK
[8] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA

Keywords: historical images, camera calibration, bundle adjustment, registration, multi-epoch

Abstract

Historical film-based images, acquired during aerial campaigns since the 1930s and from satellite platforms since the 1960s, provide a unique opportunity to document changes in the Earth's surface over the 20th century. Yet, exploiting these rich data presents significant and specific challenges, including complex distortions in the scanned images and poorly known exterior and/or interior camera orientation. In recent years, semi- or fully-automated approaches based on photogrammetric and computer vision methods have emerged, but their performance and limitations have not yet been evaluated in a consistent way.
The ongoing project "Historical Images for Surface Topography Reconstruction Intercomparison eXperiment (HistoriX)" aims at comparing existing methods for processing stereoscopic historical images and harmonizing processing tools. Within this experiment, participants were provided with a standardized set of historical images and available metadata, then invited to produce a point cloud and estimated camera parameters. We selected two study sites and three sets of film-based images acquired coincidently in the 1970s or 80s from airborne or satellite platforms. The submitted elevation data are currently being cross-validated across different image sets and participant submissions, as well as against reference elevation data over stable terrain. The spread in the retrieved elevations is analysed with respect to image type, terrain type and processing methods to highlight the strengths and limitations of the different approaches. Here we present the study site and datasets selected for this comparison, the design of the intercomparison and evaluation metrics, as well as preliminary results.

1. Introduction

Historical film-based images, acquired by aerial sensors since the 1930s and by satellite platforms since the 1960s, provide a unique opportunity to document changes in the Earth surface over the 20th century. Yet, they present significant and specific challenges, including complex distortion in the scanned image pixel grid and poorly known camera exterior and interior orientation. In recent years, semi- or fully-automated approaches, based on photogrammetric and computer vision methods, have emerged (e.g., Knuth et al., 2023; Dehecq et al., 2020; Ghuffar et al., 2022; Dahle et al., 2024), but the performance and limitations of these methods have yet to be directly compared.
Semi- or fully-automated approaches usually bypass the need for manual Ground Control Points (GCPs) either by automating the detection of GCPs, generally leveraging deep-learning approaches (e.g., Zhang et al., 2021; Maiwald et al., 2023; Dahle et al., 2024) or by coregistering a preliminary Digital Elevation Model (DEM) with an external reference and applying the same transformation to the camera system (e.g., Dehecq et al., 2020; Knuth et al., 2023). Although these approaches have been validated individually, they are rarely compared to independent, traditional photogrammetric approaches. Additionally, these processing pipelines are often tailored to a specific dataset, in particular aerial or satellite images. All types of processing pipelines, whether manual or automated, usually rely on a different mix of commercial software (Agisoft Metashape, ERDAS Imagine, Inpho etc.) or open source research software (MicMac, COLMAP, Ames Stereo Pipeline). These software packages use different (sometimes poorly documented) algorithms, input and output formats, or terminology. In order to leverage the best of all existing tools, there is a need to either improve interoperability between software or develop a community-driven open source and well documented solution.
The main objectives of the Historical Images for Surface Topography Reconstruction Intercomparison eXperiment (HistoriX) project are to compare existing methods for processing stereoscopic historical images and harmonize processing tools. In particular, we aim to:
- Document and evaluate existing photogrammetric processing chains suited for historical satellite and aerial imagery.
- Assess the accuracy and precision of elevation data retrieved with the proposed methodologies and chosen historical datasets, with the aim of identifying systematic differences or errors among methodologies or software.
- Establish a benchmark dataset to validate future development of tools and methods.
- Centralize reproducible code/workflows to share knowledge, solve common problems, and enhance interoperability between processing chains and software.

2. Data

2.1 Image datasets and study sites

In our experiment, we compared methods suited for aerial and/or satellite film-based images. We first selected images from the American Hexagon (KeyHole-9, hereafter KH-9) reconnaissance satellite missions, which have nearly global coverage of land surfaces (Piermattei et al., 2026). Stereo images from both the mapping camera (KH-9 MC) and panoramic camera (KH-9 PC) acquired on the same day were selected. We then selected aerial images as close in time as possible to the satellite data to ultimately perform cross-validation of all results. Aerial images were selected based on quality, stereo overlap, presence of fiducial markers and available calibration reports. Details on the selected images are summarized in Table 1.

We also aimed at comparing method performance in different geographical contexts and for different applications. Therefore, we selected two study sites (Fig. 1):

1) The Casa Grande Photogrammetric Test Range in Arizona, USA, featuring semi-arid landscapes with vast areas of bare soil and rock, low vegetation, and localized urban and agricultural areas. It was selected because of the relatively stable surface conditions (slow evolving vegetation, limited urbanization, little to no elevation change over the past decades) facilitating multi-epoch tie point matching and potential coregistration, and due to the presence of a dense array of 256 photo-identifiable control point targets with known 3D coordinates over a 26 x 26 km^2 area (Fig. 1; Merchant, 1972).

2) Southern Iceland, featuring a mix of bare soil and rock, low vegetation, glaciers, lava flows, and coastal areas. This area is representative of large environmental changes that can occur due to glacier retreat, erosion, volcanic activity, and other natural processes.

Data availability also largely dictated the choice of study sites.

In each study site, we delineated two bounding boxes: the "zoom" area that encompasses the aerial dataset and the "large" area that encompasses all datasets. The "zoom" areas are less than 100 km in size, while the "large" areas are up to 500 km long.

Site	Dataset	Acquisition date	Image count	Image GSD (m)
Casa Grande	Aerial	1978/09/06	37	0.7
	KH-9 MC	1978/03/25	4	6-9
	KH-9 PC	1978/03/25	6	0.6-1.2
Iceland	Aerial	1980/08/22	125	0.9
	KH-9 MC	1980/08/22	4	6-9
	KH-9 PC	1980/08/22	6	0.6-1.2

Table 1. Dataset selected for the experiment. GSD = Ground Sampling Distance.

2.2 Satellite image sets

Satellite images were downloaded from the US Geological Survey (USGS) data collection hosted at www.earthexplorer.usgs.gov. The KH-9 MC (resp. KH-9 PC), images are part of the "Declassified 2" (resp. "Declassified 3") datasets (EROS 2017a, 2017b). The raw panchromatic images are scanned and delivered as TIF images with a 7 µm resolution and an 8-bit encoding (256 gray levels). The large images, sometimes 2-3 m long, do not fit in photogrammetric scanners and are hence scanned in multiple pieces (2 to 8 depending on the film) with a ~1 cm overlap between pieces.

The USGS provides limited metadata for each image, from which we extracted the following: image ID, acquisition date, approximate image footprint corner and center coordinates. The provided coordinates have limited precision and are usually within a kilometer of the true location (e.g., Dehecq et al., 2020). We assumed a nominal focal length of 304.8 mm and 1524 mm for the MC and PC cameras respectively and an approximate orbit height of 160 km, based on declassified documentation and scientific literature (Burnett et al., 2012; Piermattei et al., 2026).

2.3 Aerial images sets

The aerial images for Casa Grande were also downloaded from the USGS EarthExplorer dataset "Aerial Photo Single Frame" (EROS 2017c). The raw images are scanned and delivered as TIF images with a 25 µm resolution and a single 8-bit band for panchromatic images or three 8-bit bands for color images. As for the satellite images, we extracted the following metadata from the catalog: image ID, acquisition date, approximate image corner and center coordinates in latitude and longitude.

The Iceland images were collected by the Icelandic Institute of Natural Sciences (former National Land Survey of Iceland) and are open access at https://loftmyndasja.gis.is/. Examples of the photographs of the Icelandic archives are described in Belart et al. (2019). The raw B/W images were scanned and delivered as TIF images with a 20 µm resolution and 8-bit encoding. These were accompanied with the same metadata as the other datasets. The image center and corner coordinates were manually generated using a pre-existing, approximate location of the camera centers to mirror the USGS metadata.

For all datasets, we retrieved camera calibration reports from which we extracted the measured interior orientation, including focal length and physical location of the fiducial markers.

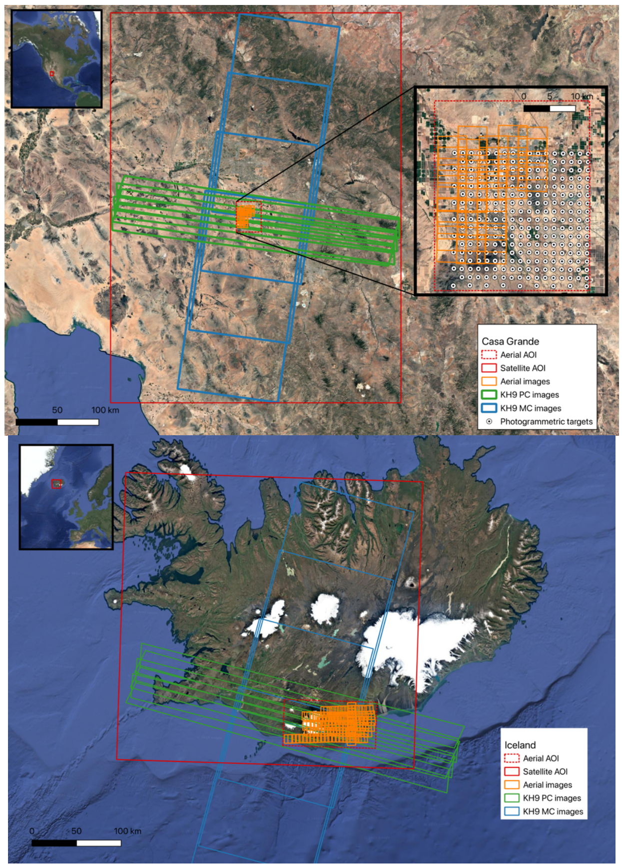

Figure 1: Map of the two test sites: Casa Grande (top), Iceland (bottom). The red rectangles indicate the areas of interest for the aerial (dotted line, so called "zoom") and satellite (solid line, so-called "large") datasets where reference data are provided. Image footprints are shown for each dataset. For Casa Grande, the inset shows the locations of the photogrammetric targets. Background image © Google.

2.4 Auxiliary data

Along with the image sets, we collected any additional data that can be leveraged for the photogrammetric processing, including: a reference DEM and mask of stable ground, manually identified Ground Control Points (GCPs), modern orthomosaics and multi-epoch images. The source of each of these datasets is detailed in the following paragraphs, while the data preparation is presented in section 3.2. We prepared the raster data (reference DEM, stable mask and modern orthomosaics) from a different source and at a different spatial resolution for the "zoom" and "large" areas (see section 2.1).

- **Reference DEMs:** the data for the "large" areas were collected from the 30 m Copernicus Global DEM (COP30), generated from TanDEM-X data acquired between 2011 and 2015 (European Space Agency, 2024), and downloaded from the Open Topography Global Datasets API (https://portal.opentopography.org). COP30 has a vertical RMSE < 1.4 m (Wessel et al., 2018) as estimated from GPS measurements. The "zoom" reference DEM for Casa Grande was prepared from the USGS 3DEP lidar point clouds for the AZ MaricopaPinal 1 2020 project[1]. A 1-m DSM and DTM were prepared using the lidar_tools package (https://github.com/uw-cryo/lidar_tools). It has a vertical accuracy of 0.05 m RMSE and horizontal accuracy of 0.2-0.3 m. The "zoom" reference DEM for Iceland was collected from the IslandsDEM v1, a seamless and bias-corrected mosaic from ArcticDEM (Porter et al., 2018) and lidar on glaciers and their vicinity (Jóhannesson et al., 2013) from the Icelandic Institute of Natural Sciences (https://dem.gis.is/). The IslandsDEM has a horizontal accuracy better than 2 m and vertical accuracy better than 0.5 m, according to the metadata, based on numerous comparisons with lidar and drone surveys.
- **Stable masks:** a raster mask was generated on the same grid as the reference DEMs by masking 1) all landcover classes but bare ground and low vegetation in the ESA WorldCover 10 m 2021 dataset (ESA CCI-LC, 2017), 2) groundwater/aquifer subsiding areas on the Casa Grande study site (Arizona Department of Water Resources[2]), and 3) glaciers in Iceland, as identified in the Randolph Glacier Inventory 7.0 outlines (RGI Consortium, 2023).
- **Reference orthomosaics:** For manual and automated GCP selection, we generated a single reference orthomosaic for each site from globally available, modern basemaps. For Casa Grande, we used the Google Satellite basemap (accessed November 2025) to generate a grayscale mosaic on the same bounds as the study areas, with a resolution of 0.25 m and 5 m for the "zoom" and "large" areas, respectively. For Iceland, we used the ESRI World Imagery basemap to generate a grayscale mosaic with a resolution of 0.5 m and 5 m for the "zoom" and "large" areas, respectively. The choice of input dataset was based on a visual comparison of the horizontal alignment with a hillshade of the high-resolution reference DEMs. The spatial resolution was chosen to be lower than the corresponding image GSD (aerial and KH-9 MC for the "zoom" and "large" rasters) while keeping manageable data volume. The original RGB mosaics were converted to grayscale using the ITU-R 601-2 luma transform to be visually closer to the historical grayscale images, and compressed with 80% JPEG compression.
- **Multi-epoch aerial datasets:** Some stereo processing approaches can benefit from using a multi-epoch bundle-block adjustment (e.g., Feurer & Vinatier, 2018); therefore we provided several additional aerial image sets for both sites acquired before or after the primary image datasets. We extracted and processed other film-based aerial archives for the following dates: 1980/07/28, 1986/07/17, 1996/07/27 (Iceland), 1971/01/19, 1971/01/24, 1971/01/25, 1990/10/31 (Casa Grande).

3. Methods

3.1 Experiment design overview

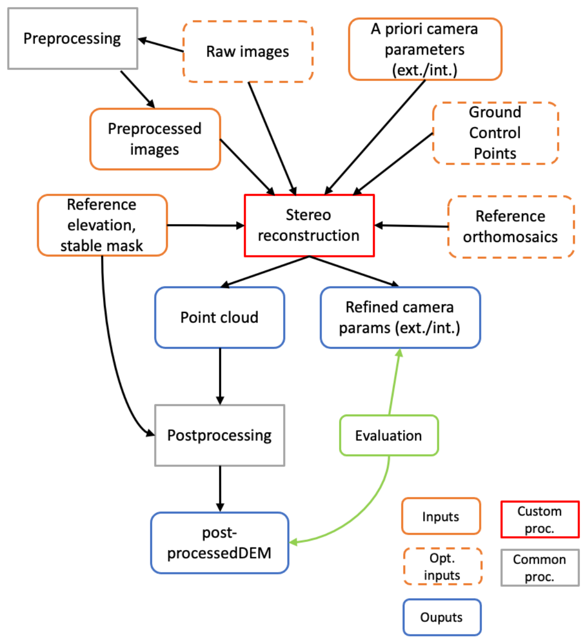

Figure 2: Diagram illustrating the intercomparison design. The stereo processing step (red box) to be evaluated takes as mandatory input (plain orange boxes): preprocessed images, a priori camera parameters and reference stable elevation, and optional inputs (dashed orange boxes). It outputs a point cloud and refined camera parameters. Pre- and post-processing steps (gray boxes) are shared among submissions, to avoid evaluating these steps.

In this experiment, participants were provided with a set of historical images and available metadata and invited to return an elevation point cloud and estimated (i.e., refined) camera parameters. The submitted elevation datasets were then cross-validated between the different image sets and submissions, as well as against reference elevation data on stable terrain. The experiment aimed to limit the impact of pre- and post-processing methods to focus on the core photogrammetric steps leading from the preprocessed images to the raw point cloud (Figure 2). Therefore all other processing steps and input data were strictly shared among all participants, in particular:
- We provided a set of preprocessed images to avoid evaluating the effect of preprocessing alongside the photogrammetric processing.
- We provided all auxiliary data needed for the processing, to avoid evaluating the impact of input data quality.

[1] https://prd-tnm.s3.amazonaws.com/StagedProducts/Elevation/metadata/AZ_MaricopaPinal_2020_B20/AZ_MaricopaPinal_1_2020/reports/312020359_USGS_Maricopa_Lidar_LidarMappingReport.pdf (last accessed 30 Mar 2026)

[2] https://www.azwater.gov/hydrology/land-subsidence/ tucson-land-subsidence-feature (last accessed 30 Mar 2026)

- Participants were asked to submit a raw point cloud, to avoid introducing differences in postprocessing. The exact same postprocessing was then applied to all submissions to derive the final DEM.
- Participants were encouraged to present, document and share their processing workflows.

3.2 Experiment input data

3.2.1 Overview of input data: Input data were split between “mandatory” and “optional” input data. The use of any of the mandatory data was considered implicit, as most approaches rely on those, while the use of optional data was to be specified with an experiment naming convention, in order to facilitate the analysis of processing strategies. Mandatory input data included: preprocessed images, calibrated or assumed (for satellites) intrinsic parameters, approximate extrinsic parameters and/or image footprints, full calibration report if available, modern reference elevation and stable terrain mask. Optional input data included: raw images, manual GCPs and modern reference orthomosaics. Inputs were strictly enforced, i.e. participants were not allowed to rely on any other data, such as self-picked GCPs or external orthophotos, to avoid evaluating differences in data quality. Raw images were preprocessed using documented and standard approaches (see section 3.2.2), but participants could optionally apply their own custom preprocessing as an additional experiment, to analyze the impact of preprocessing. Participants were also provided with a document summarizing the experiment design, input data and deliverables. The following sections describe the preparation of the input data.

3.2.2 Image preprocessing: The historical images were preprocessed or standardized 1) to stitch the image pieces in the case of the satellite images, 2) to remove any distortion introduced during scanning and possibly during film handling and 3) cropped such that image fiducial markers are aligned and dimensions are identical between images. The preprocessing is specific to each dataset but based on published literature.

For aerial images, the preprocessing method mainly followed Knuth et al. (2023). In brief:
- The images were first rotated to match the orientation in the calibration report and color images were converted to grayscale.
- Fiducial markers were detected using template matching with a manually selected template from one of the images. Outliers were discarded using a threshold on the correlation score and the angle between markers (which should be close to 90º at corners).
- An affine transformation was computed and applied to the image to align the detected markers with the calibration report positions.
- The images were cropped at fixed dimensions around the outermost markers.
- Contrast limited adaptive histogram equalization (CLAHE) was applied to improve image contrast.

For KH-9 MC images, the preprocessing followed Dehecq et al. (2020). In brief:
- The position of the reseau markers (regularly spaced dark crosses) was detected at subpixel resolution using a convolution with a cross-like kernel.
- Outliers were rejected by removing markers that are not regularly spaced (1 cm).
- A degree 3 polynomial transformation was first applied to remove rotation and scaling, then a Thin Plate Spline interpolation was used to calculate the non affine distortion at each pixel.
- The two pieces were stitched together by matching the areas of overlap using normalized cross-correlation.
- The image was cropped to a fixed distance from the outermost markers to keep a constant image dimension among the images
- The reseau markers were filled with inpainting.

For KH-9 PC images, the preprocessing method was similar to that of Ghuffar et al. (2023) with a few additional steps. In brief:
- The image pieces were joined into a single composite image using feature-matching to detect tie points in areas of overlap and an affine transformation was calculated to align the pieces together.
- The image collimation lines (bright lines at the top and bottom of the film, printed during acquisition) were detected and a Thin-Plate Spline transformation was used to align these lines with the horizontal direction and keep a fixed distance between the two. The top and bottom edges were then cropped to a fixed distance from these lines.
- The image left and right edges were detected automatically and a vertical line was fit to crop the left and right borders.

3.2.3 Auxiliary data preparation: The auxiliary data were prepared in order to provide a standardized dataset. First, all geographic data were reprojected to a local, metric coordinate system (EPSG:32627 for Iceland, EPSG:6341 for Casa Grande). The relevant camera intrinsic (focal length, pixel size, physical position of the fiducial markers with respect to the optical center) and extrinsic (a priori camera center location and flying height) parameters were provided in a standardized CSV format. The “large” reference DEMs were converted to the same vertical datum as the “zoom” reference DEMs and horizontally aligned using the method proposed by Nuth & Kääb (2011) to remove any horizontal offset between the datasets.

3.2.4 Manual selection of GCPs: For non-automated approaches, we created a unique set of GCPs for the aerial and for the KH-9 MC datasets only (participants interested in processing the KH-9 PC data did not raise interest in using GCPs). GCPs were manually identified by one (resp. two) operators for the aerial (resp. KH-9 MC) datasets using the reference orthomosaics and DEMs provided as auxiliary data. The image coordinates were obtained using either the “Multi Photo Measurement Tool” (in the ApplicationsMaster Software by Inpho/Trimble) or the “Point Measurement tool” (in ERDAS Imagine LPS). A few GCPs were first selected to create approximately georeferenced historical orthophotos, to then facilitate the selection and evaluation of further GCPs.
The strategy for GCP picking differed between operators, illustrating the variability in GCP selection approaches. Operator 1 focused on group features (such as groups of trees) to limit the risk of selecting single objects that may appear similar but have moved, and the final selected feature was usually a round one, such as a bush, to reduce the sensitivity to light change conditions. Operator 2 focused on intersections (crossroads, road/river, bridges), manmade features or stable terrain landforms such as boulders. Both outputs were merged and evaluated by an independent bundle-block adjustment (BBA) to ensure consistency.
This resulted in a total of 44 aerial and 66 KH-9 MC GCPs for Casa Grande and 70 aerial and 38 KH-9 MC GCPs for Iceland. The easting and northing coordinates of the final GCPs were extracted from the manually identified corresponding point in the reference orthomosaic. The elevation at these locations was then extracted from the given reference DEM (“zoom” for aerial, “large” for KH-9 MC).

3.3 Experiment output/deliverables

All participants were requested to process at least one dataset for both study sites, using the preprocessed images (mandatory experiment) or raw images (optional). The mandatory deliverables included: the sparse point cloud (generated from tie points used in the BBA), the dense point cloud, estimated/refined camera intrinsic and extrinsic parameters. Optional deliverables included a postprocessed DEM and an orthomosaic. Generated point clouds were provided as compressed LAS files, rasters in GTiff format and tabular data in CSV format. Each group was also requested to provide a processing report describing the participants contribution, software used, the methodology and computing resources used. Each group was allowed to submit several results from the same input dataset, obtained with different setup or algorithms. It was requested that the dense point cloud be provided without interpolation or complex filtering involving auxiliary data, since the same postprocessing was applied to all submissions prior to evaluation. However, participants were allowed to provide a postprocessed DEM if they considered that they could provide a higher quality product with a custom postprocessing. In which case, both raw point cloud and postprocessed DEM were evaluated individually to analyze the impact of stereo processing and postprocessing strategies separately. Orthoimages were collected if provided, but no systematic evaluation was planned for the moment.

3.4 Output postprocessing

Point clouds were post-processed with the following standardized steps:
- Rasterization on the same grid as the corresponding reference DEM using inverse distance weighting interpolation (PDAL implementation).
- Correction of residual horizontal and vertical rigid shifts with respect to the reference DEM on assumed stable terrain, following the Nuth and Kääb (2011) methodology implemented in xDEM (xDEM contributors, 2025). Residual rotation or scale errors were not corrected.
- No outlier filtering was applied for this preliminary analysis, but mild filtering, e.g., residual threshold with respect to the reference DEM, will be considered in the final analysis in case of strong/obvious outliers.

3.5 Evaluation

The evaluation of the output for this preliminary round focused primarily on the quality of the retrieved elevations, but the evaluation of the sparse point clouds and/or camera parameters is envisioned in order to understand differences that occur between submissions. For example, the number of points or residuals of the sparse point cloud (with respect to the reference elevation) may explain strengths/weaknesses in the tie point detection step or performance of the BBA. The spread of values in camera extrinsic or intrinsic parameters may help identify faulty BBA or georeferencing. The next subsections focus on the comparison and evaluation of the final DEM that is generated from the dense point clouds using the standard postprocessing approach described in section 3.4. The evaluation of the results is currently ongoing as the final submissions are expected for April 2026, so the evaluation procedure is still subject to changes.

3.5.1 Comparison to reference elevation

First, the retrieved elevations will be compared against the reference elevation (modern Lidar or satellite data, see section 2.4) on assumed stable terrain. Basic statistics (mean, median, standard deviation, Normalized Median Absolute Deviation or NMAD) will be calculated to assess the accuracy and precision of the historical elevations. These statistics will be computed for the entire domain, but also binned by point density, terrain slope, curvature and by landcover type to assess differences in uncertainty distributions across the domain, i.e., heteroscedasticity (Bater and Coops, 2009; Hugonnet et al., 2022). Spatial statistics, and in particular variogram analysis, will be used to assess spatially correlated errors (Rolstad et al., 2009; Hugonnet et al., 2022), which may originate from image noise, artifacts, unresolved camera distortions or dense matching assumptions. This step quantifies the range of biases and random errors in elevation that are commonly present in products derived from stereoscopic images. Changing terrain types (glaciers, vegetation, water bodies…) are excluded from this evaluation, given that our reference elevation is temporally separated by several decades.

3.5.2 Intercomparison of elevation results

To overcome the lack of validation data for historical datasets, we take advantage of (near) synchronous acquisitions across submissions and datasets in this intercomparison. First, the different dense point clouds will be analyzed in terms of point density and fraction of data gaps, to assess the spatial coverage that can be achieved for a given dataset and processing methodology. Then, the different datasets will be cross-compared and the same statistics as calculated in section 3.5.1 will be analyzed to identify possible systematic differences in accuracy or precision for specific datasets, methodologies (e.g., manual vs automated) or software. Finally, the central tendency (median) and spread (NMAD) of elevation data across all submissions of a same dataset will be calculated at each pixel in order to assess the spatial variability of the differences. The spread will be analyzed in particular with regard to terrain type, stereo overlap and distance to nearest GCP. This analysis may reveal biases for certain terrain types (including changing terrain types), algorithms and/or help define a set of processing recommendations.

4. Preliminary results

4.1 Submissions

At this stage, 20 groups, representing ~30 persons, with published workflows for stereo historical image processing were invited to participate. We aimed at having a diversity of methods (traditional photogrammetry, SfM, AI-based), software and scientific expertise (glaciology, forestry, volcanology, etc). All participants discussed choices on data, study sites, mandatory and optional input data, as well as deliverables during online meetings and one in-person workshop in September 2025.

	Aerial	KH-9 MC	KH-9 PC	Total
Casa Grande	18	6	2	26
Iceland	9	4	1	14
Total	27	10	3	40

Table 2: Number of submissions received for each study site and dataset for the preliminary round.

Participants were invited to submit preliminary results, for at least a subset of the study cases, in October 2025. A total of 18

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-9-2026 | © Author(s) 2026. CC BY 4.0 License.

different methods were used to submit 40 results, sometimes with different parameter settings. The number of submissions for each site and data type is summarized in Table 2. The approaches used can be qualified as traditional photogrammetry with manual GCPs, automated approaches based on multi-epoch tie point matching or based on point-cloud/DEM co-registration.

4.2 Preliminary comparison

Our preliminary evaluation indicated that about half of the submissions have uncertainties in the expected range (e.g., the NMAD of differences to the stable reference terrain is less than the image GSD), but some submissions did not achieve satisfying results. Reasons for that may be a lack of time for this preliminary round or potential issues in the data or experiment design. Such issues have been discussed and adjustments have been made (as described in this article). We expect the next round of submissions to better reflect the performance of each processing methodology.
With these limitations in mind, we illustrate the analysis that is planned on Figures 3 and 4. Figure 3 shows the spread (standard deviation) of elevation values between all submissions for the aerial dataset of Casa Grande, after removing gross outliers. The apparent "edge" or "seam" artifacts are likely related to the distribution of input image footprints, while areas of larger variability are likely correlated with the lack of GCP constraint or low image overlap on the edges. Figure 4 shows the spread of elevation values for each study site and dataset, as a function of the landcover type. It highlights the differences in uncertainty between datasets, which is roughly proportional to the image GSD, i.e., the spread between the submissions is smallest for aerial images, followed by KH-9 PC and KH-9 MC. It must be noted however that the number of KH-9 PC submissions (2) for this round was limited and may not yet reflect the variability of methods. More submissions are expected. For the aerial datasets, no clear differences are found for the different terrain types, since as discussed above, the main factors of variability are likely the distribution of GCPs and image overlap. For the satellite datasets, there is a tendency towards higher disagreement on forests, urban areas and water bodies.

4.3 Open source code

The project aims at encouraging open and reproducible science. Therefore, all codes used for preprocessing the input data, postprocessing the deliverables and evaluation are publicly available since the beginning of the project. We initiated a set of Python packages aiming at centralizing code for historical image processing. A package named "Historial Image Pre-processing" or HIPP was created and is the basis for all image preprocessing used in this experiment. It is available at https://github.com/shippp/hipp. It is largely based on previous developments by Knuth et al. (2023), McNabb et al (2020) and Dehecq et al. (2020). All codes and scripts used to prepare the experiment data, as well as for the postprocessing and evaluation are centralized in the following repository: https://github.com/shippp/history. Any additional software outcome of this project will be made publicly available in a shared GitHub organization "Stereo Historical Image Processing Python Packages" or SHIPPP: https://github.com/shippp.

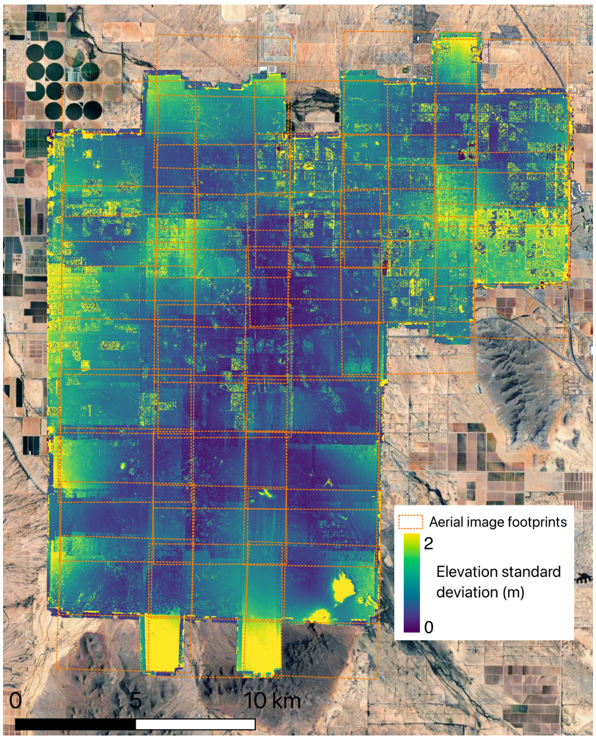

Figure 3: Standard deviation of elevations from all submissions, calculated at each pixel of the DEM grid, for the aerial dataset of Casa Grande (18 submissions in total).

5. Conclusions

This preliminary exercise allowed us to identify common points and differences in input data requirements, such as the number and quality of GCPs or reference orthomosaics, and to agree on an updated set of input data (as described in this article). The updated dataset was sent to all participants in January 2026, with a submission deadline in April 2026. A final evaluation of the results is planned by the end of 2026.
The analysis of the submissions will hopefully help us address the following research questions:
- What is the best achievable accuracy and precision of elevations retrieved from historical aerial, KH-9 MC and KH-9 PC images?
- How does it vary with landcover and terrain type?
- Are there systematic differences or biases in existing methods?

By the end of the project, we plan to release the experiment dataset on a public data repository which can serve as a benchmark dataset for future exercises. These could focus, for example, on better evaluating the impact of the pre- or post-processing steps, the impact of GCP number and quality, or the impact of the reference orthomosaic or DEM. We hope to strengthen the set of open source code released under the SHIPPP GitHub organization, in order to facilitate interoperability between processing chains and foster community-driven software and reproducible science.

References

Bater, C.W., Coops, N.C., 2009. Evaluating error associated with lidar-derived DEM interpolation. Computers & Geosciences 35, 289–300. https://doi.org/10.1016/j.cageo.2008.09.001

Belart, J.M.C., Magnússon, E., Berthier, E., Pálsson, F., Aðalgeirsdóttir, Gu., Jóhannesson, T., 2019. The geodetic mass balance of Eyjafjallajökull ice cap for 1945–2014: processing guidelines and relation to climate. Journal of Glaciology 65, 395–409. https://doi.org/10.1017/jog.2019.16

Burnett, M., 2012. Hexagon (KH-9) mapping camera program and evolution. National Reconnaissance Office (NRO), Center for the Study of National Reconnaissance (CSNR), Chantilly, VA, USA.

Dahle, F., Lindenbergh, R., Wouters, B., 2024. Polar perspectives: a deep dive into geo-referencing historical Antarctic photos. International Journal of Digital Earth 17, 2406384. https://doi.org/10.1080/17538947.2024.2406384

Dehecq, A., Gardner, A.S., Alexandrov, O., McMichael, S., Hugonnet, R., Shean, D., Marty, M., 2020. Automated Processing of Declassified KH-9 Hexagon Satellite Images for Global Elevation Change Analysis Since the 1970s. Front. Earth Sci. 8. https://doi.org/10.3389/feart.2020.566802

EROS, Earth Resources Observation And Science Center (2017a) Declassified Satellite Imagery - 2. doi:10.5066/F74X5684.

EROS, Earth Resources Observation And Science Center (2017b) Declassified Satellite Imagery - 3. doi:10.5066/F7WD3Z10.

Earth Resources Observation And Science (EROS) Center (2017c) Aerial Photography Single Frame. doi:10.5066/F7610XKM.

ESA CCI-LC, CCI Land Cover time series version 2. Technical Report (2017). Available at maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf. Accessed 25 Mar 2026.

European Space Agency (2024). Copernicus Global Digital Elevation Model. Distributed by OpenTopography. https://doi.org/10.5069/G9028PQB. Accessed 2026-03-25

Feurer, D., Vinatier, F., 2018. Joining multi-epoch archival aerial images in a single SfM block allows 3-D change detection with almost exclusively image information. ISPRS Journal of Photogrammetry and Remote Sensing 146, 495–506. https://doi.org/10.1016/j.isprsjprs.2018.10.016.

Ghuffar, S., Bolch, T., Rupnik, E., Bhattacharya, A., 2022. A pipeline for automated processing of declassified Corona KH-4 (1962-1972) stereo imagery. IEEE Trans. Geosci. Remote Sensing 1–1. https://doi.org/10.1109/TGRS.2022.3200151.

Hugonnet, R., Brun, F., Berthier, E., Dehecq, A., Mannerfelt, E.S., Eckert, N., Farinotti, D., 2022. Uncertainty analysis of digital elevation models by spatial inference from stable terrain. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 1–17. https://doi.org/10.1109/JSTARS.2022.3188922

Jóhannesson, T., Björnsson, H., Magnússon, E., Guðmundsson, S., Pálsson, F., Sigurðsson, O., Thorsteinsson, T., Berthier, E., 2013. Ice-volume changes, bias estimation of mass-balance measurements and changes in subglacial lakes derived by lidar mapping of the surface of Icelandic glaciers. Annals of Glaciology 54, 63–74. https://doi.org/10.3189/2013AoG63A422.

Knuth, F., Shean, D., Bhushan, S., Schwat, E., Alexandrov, O., McNeil, C., Dehecq, A., Florentine, C., O'Neel, S., 2023. Historical Structure from Motion (HSfM): Automated processing of historical aerial photographs for long-term topographic change analysis. Remote Sensing of Environment 285, 113379. https://doi.org/10.1016/j.rse.2022.113379.

Maiwald, F., Feurer, D., Eltner, A., 2023. Solving photogrammetric cold cases using AI-based image matching: New potential for monitoring the past with historical aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 206, 184–200. https://doi.org/10.1016/j.isprsjprs.2023.11.008.

McNabb, R., Girod, L., Nuth, C., Kääb, A., 2020. An open-source toolset for automated processing of historic spy photos: sPyMicMac. https://doi.org/10.5194/egusphere-egu2020-11150.

Merchant, D.C., 1972. Metric Calibration of the Aerial Photographic System by the Method of Mixed Ranges. The Ohio State University Research Foundation. https://apps.dtic.mil/sti/tr/pdf/AD0748812.pdf (Last accessed 30 Mar 2026)

Nuth, C., Kääb, A., 2011. Co-registration and bias corrections of satellite elevation data sets for quantifying glacier thickness change. The Cryosphere 5, 271–290. https://doi.org/10.5194/tc-5-271-2011.

Piermattei, L., McNabb, R., Elias, M., Ressl, C., Dehecq, A., Girod, L., Dewez, T., Eltner, A., 2026. Unlocking the Past: A review of digital processing of historical aerial and satellite stereo analog imagery for geoscience applications. IEEE Geoscience and Remote Sensing Magazine 2–33. https://doi.org/10.1109/MGRS.2025.3645144.

Porter, C., Morin, P., Howat, I., Noh, M.-J., Bates, B., Peterman, K., Keesey, S., Schlenk, M., Gardiner, J., Tomko, K., Willis, M., Kelleher, C., Cloutier, M., Husby, E., Foga, S., Nakamura, H., Platson, M., Wethington, M., Williamson, C., Bauer, G., Enos, J., Arnold, G., Kramer, W., Becker, P., Doshi, A., D'Souza, C., Cummens, P., Laurier, F., Bojesen, M., 2018. ArcticDEM. https://doi.org/10.7910/DVN/OHHUKH.

RGI Consortium, 2023. Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 7.0. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. https://doi.org/10.5067/f6jmovy5navz

Rolstad, C., Haug, T., Denby, B., 2009. Spatially integrated geodetic glacier mass balance and its uncertainty based on geostatistical analysis: application to the western Svartisen ice cap, Norway. Journal of Glaciology 55, 666–680. https://doi.org/10.3189/002214309789470950.

Wessel, B., Huber, M., Wohlfart, C., Marschalk, U., Kosmann, D., Roth, A., 2018. Accuracy assessment of the global TanDEM-X Digital Elevation Model with GPS data. ISPRS Journal of Photogrammetry and Remote Sensing 139, 171–182. https://doi.org/10.1016/j.isprsjprs.2018.02.017.

xDEM contributors, 2025. xDEM (v0.1.6). Zenodo. https://doi.org/10.5281/zenodo.4809697.

Zhang, L., Rupnik, E., Pierrot-Deseilligny, M., 2021. Feature matching for multi-epoch historical aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 182, 176–189. https://doi.org/10.1016/j.isprsjprs.2021.10.008.

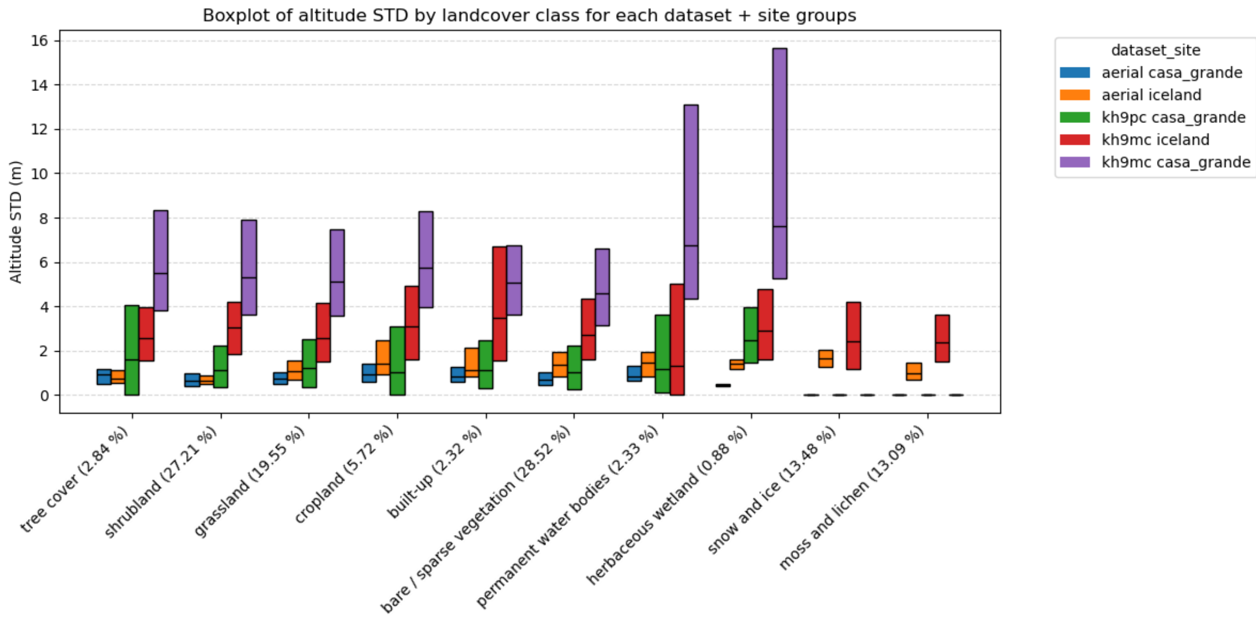

Figure 4: Boxplots (median and interquartile range) from all pixels belonging to the same landcover type, of the spread (standard deviation) of elevations among all submissions (Table 2). Each color represents a study site and dataset. The blue boxes are calculated from the results shown on Figure 3.