

GEM-Qt: Generating embeddings for meteorological reanalysis data using a quadtree-based approach

Gabriela Czibula, **Andrei Mihai**, Sîrbu Alexandru-Gabriel,
Istvan Gergely Czibula, Eugen Mihuleț, and Ioan-Ștefan Gabrian
Babeș-Bolyai University

WeADL 2026 Workshop

The workshop is organized under the umbrella of WinDMiL, project funded by CCCDI-UEFISCDI, project number [PN-IV-P7-7.1-PED-2024-0121](#), within PNCDI IV

Goal – Embeddings of Meteorological Data

- Compress a meteorological image → generate embeddings that capture relevant information
 - using a quad-tree based approach

Goal – Embeddings of Meteorological Data

- Compress a meteorological image → generate embeddings that capture relevant information
 - using a quad-tree based approach
- Qualitative evaluation of the embeddings
 - in terms of clustering
 - compare to embeddings generated using other baseline models

Reanalysis Dataset

Data: ECMWF **Re**Analysis, 5th generation (ERA5)

Source: Copernicus Climate Data Store catalog

- Geographical area covering Southeastern Europe, centered on Romania
- Spatial resolution ~ 31 km ($0.25^\circ \times 0.25^\circ$ latitude–longitude)
- Temporal resolution – 1 hour, from 2020 to 2024

Labeling: by general atmospheric circulation using the WLK method
each timestep was associated 1 of 40 labels

Products used

We chose to use the following products:

Land variables:

- 2 meters temperature (t_{2m})
- u and v components of the wind (u_{10} and v_{10})
- mean sea level pressure (msl)

Free atmosphere variables:

- air temperature at 850 and 500 hPa (t_{850} and t_{500})
- u and v components of the wind at 850 hPa (u_{850} and v_{850})
- geopotential height at 500 hPa (g_{500})
- relative humidity at 700 hPa (rh_{700})

① 10 products, originally selected:

t2m, u10, v10, msl, t850, t500, u850, v850, g500, rh700

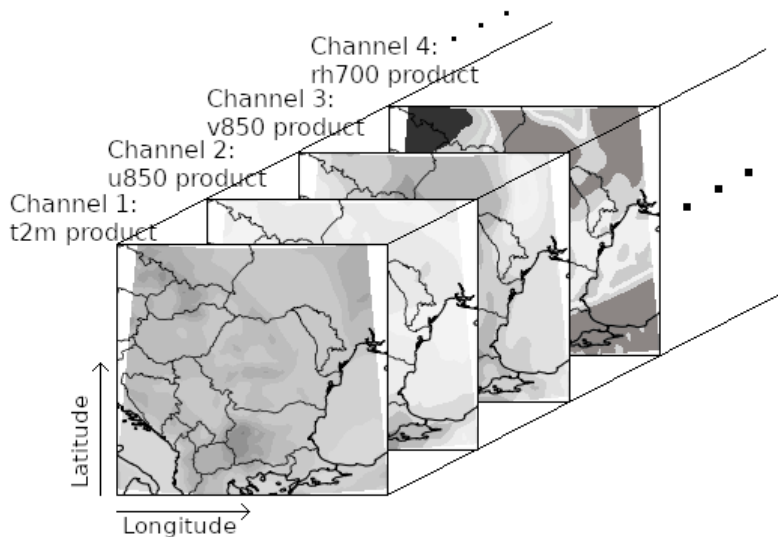
② 5 products selection:

msl, u850, v850, g500, rh700

③ Corelation-based 5 products selection:

u10, v10, msl, t500, rh700

Data representation



Quadtree Definition

A *quadtree* is a data structure that recursively partitions an input domain into four equally sized quadrants and sub-quadrants, where each node represents a specific region of the original space.

Quadtree use in GEM-Qt

- If the meteorological image contains sufficient complex information, it should be subdivided into four equal subimages.
- Obtain embeddings for each subimage then combine them
- If the data is relevant enough (a criterion is met) for the (sub)image, do not subdivide, and generate the embedding

Criterion used:

- 1 If the image is less than a specific size
- 2 If the image has more than 25% missing values

Quadtree approach

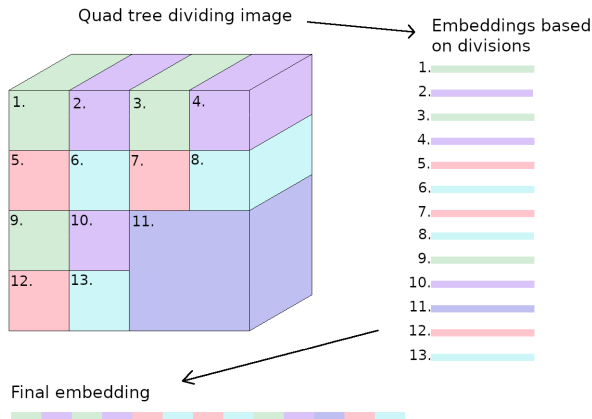


Figure: Example of applying the quadtree-based algorithm to generate embeddings. Continuously dividing the image into 4 subdivisions until a certain criteria is met, then creating embeddings for each subdivision and finally combining the divisions.

Embeddings - Histogram based

Statistical based compression algorithm.

Histogram embedding algorithm:

- 1 Split into its separate products.
- 2 Compute histogram for each product.
 - 20-binned histogram
 - range of bins decided based on min and max for that product for the entire dataset
- 3 Concatenate histograms to create final embedding

Embeddings - JPEG based

Visually aware compression algorithm.

- Based on the JPEG 2000 compression algorithm.
- To generate uniform sized embeddings bit rate was controlled by changing:
 - the compression ratio on every layer
 - the number of decomposition levels
- Minimal padding was added to reach desired fixed size length.

Embeddings - Convolutional Autoencoder (CAE) based

Machine learning based compression algorithm.

- *Autoencoder* – a neural network comprised of 2 parts: an encoder and a decoder
 - input is the same as output, it learns how to encode then decode the same image
- Embeddings – the output of the encoder's final layer
- Encoder architecture:
 - 3 convolutional layers with kernel size 3 and stride 2.
 - each followed by a SELU activation.
 - Feature map flattening.
 - Fully connected layer
- The Decoder mirrors the encoder
- Loss function: Mean Square Error (MSE)

Clustering Based Evaluation

- Qualitative evaluation of the resulted embeddings.
- Applied *k*-means clustering algorithm on the generated embeddings datasets.
 - using 40 clusters, to match the number of labels
- Evaluate the quality of the generated clusters.
 - using internal evaluation metrics
 - using external evaluation metrics, based on the preexisting labels

Internal validation metrics

- **Silhouette Score** – measures intra-cluster cohesion and inter-cluster separation
 - values range from -1 to 1
 - higher scores indicate more compact and well-separated clusters
- **Davies-Bouldin (DB)** – measures the average similarity between each cluster and its most similar cluster
 - considering both cluster compactness and separation
 - a better clustering is expressed by a lower DB value

Clustering measures

External validation metrics

- **Homogeneity** – measures how mixed are the clusters (in term of labels)
 - values range from 0 to 1
 - 1 indicates that each cluster contains only members of a single class
- **Completeness** (DB) – measures the extent to which all data instances of a particular class are assigned to the same cluster
 - values range from 0 to 1
 - 1 indicates that all members of each class appear in only one cluster
- **V-measure** – harmonic mean of homogeneity and completeness

Results - Table

generateFeatureVector	Meteorological products	Internal evaluation		External evaluation		
		Silhouette $\uparrow$	DB index $\downarrow$	Homogeneity $\uparrow$	Completeness $\uparrow$	V-measure $\uparrow$
Histogram-based	10 products	0.04	3.62	0.27	0.21	0.24
	5 products	0.04	3.66	0.29	0.23	0.26
	Correlation-based selection	0.05	3.31	0.24	0.19	0.21
JPEG-based	10 products	0.01	20.92	0.01	0.01	0.01
	5 products	0.01	14.09	0.01	0.01	0.01
	Correlation-based selection	0.01	13.77	0.01	0.01	0.01
Autoencoder-based	10 products	0.09	2.05	0.29	0.24	0.26
	5 products	0.08	2.14	0.29	0.24	0.26
	Correlation-based selection	0.07	2.29	0.26	0.21	0.23

Table: Results obtained using the k -means clustering method and the Euclidean distance metric. The best values obtained for each evaluation metric are highlighted.

Comparison to related Work

DL Model	Internal evaluation		External evaluation		
	Silhouette $\uparrow$	DB $\downarrow$ index	Homo-geneity $\uparrow$	Completeness $\uparrow$	V-measure $\uparrow$
Inception-v3	0.08	2.24	0.26	0.21	0.23
Efficient-b0	0.07	2.48	0.12	0.10	0.11
CAE	0.08	2.12	0.29	0.23	0.26
GEM-Qt	0.09	2.05	0.29	0.24	0.26

Table: Comparison to related work. The best values for the performance metrics are marked in bold.

Conclusions

Introduced GEM-Qt – a proof-of-concept quadtree-based method for generating embeddings for meteorological data.

- Three embedding approaches are compared: histogram-based, JPEG-based, CAE based
- CAEs were empirically found to be the best method
- Generating embeddings using quadtrees led to more qualitative embeddings without the quadtree-based data split.

Future work:

- Evaluate embedding methods based on other types of neural networks.
- Extend the experimental evaluation on satellite imagery and radar data.
- Reduce the spatial extent of the area of interest.

Thank you! Questions?

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI-UEFISCDI, project number PN-IV-P7-7.1-PED-2024-0121, within PNCDI IV.