Multi-frame cloud prediction from all-sky images: RGB vs. Segmented masks

J. Gatón[1,2], R. Román[1,2], C. Guzman[3], D. González-Fernández[1,2], B. Longarela[1,2], C. Herrero del Barrio[1,2], S. Herrero-Anta[1,2], R. González[1,2], C. Toledano[1,2]

(1) Grupo de Óptica Atmosférica (GOA), Universidad de Valladolid, Valladolid, 47011, Spain
(2) Laboratory of Disruptive Interdisciplinary Science (LADIS), Universidad de Valladolid, Valladolid, 47011, Spain
(3) DriMT AI Space, Valga, 68204, Estonia

gaton@goa.uva.es

- Many cloud nowcasting methods rely on all-sky imagery
- Most deep learning approaches operate on raw RGB frames
- Semantic structure is not explicitly exploited

Does input representation (RGB vs. semantic masks) affect short-term cloud prediction stability?

OBSERVATIONS AND LABELS:

Data: SKIPP'D dataset
- 57,803 all-sky videos · 31 frames · 64x64 · 1-minute resolution

Semantic Representation: GOA-UVa All-Sky segmentation U-Net model
- Pixel-wise sky segmentation
- 5 classes:
  - Cloud-free
  - Cloud
  - Low-confidence cloud
  - Sun
  - Not-sky

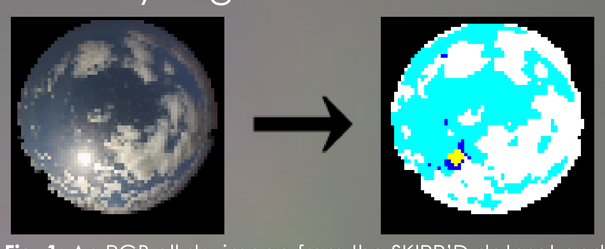

Fig. 1: An RGB all-sky image from the SKIPP'D dataset and its semantically segmented mask obtained with the GOA-UVa All-Sky segmentation U-Net model

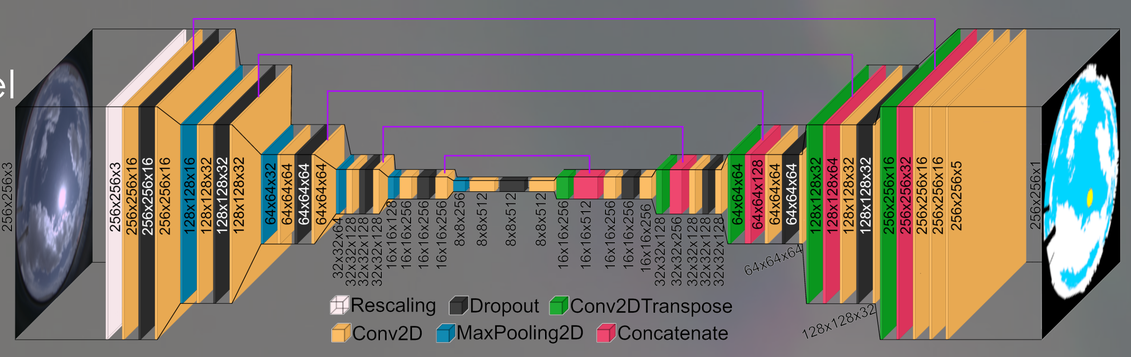

Fig. 2: Architecture of the GOA-UVa All-Sky segmentation U-Net model

WHAT IS COMPARED?: Two configs of the same ConvLSTM backbone. They differ in input & output representation:

- **RGBConvLSTM**: Operates on **RGB** frames (64x64x**3**)
- **MaskConvLSTM**: Operates on **semantic** maps (64x64x**5**)

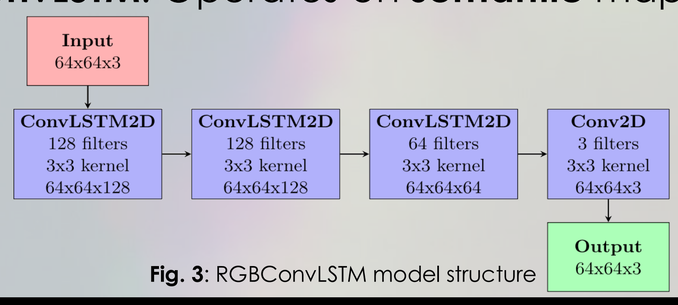

Fig. 3: RGBConvLSTM model structure

PREDICTION BACKBONE:

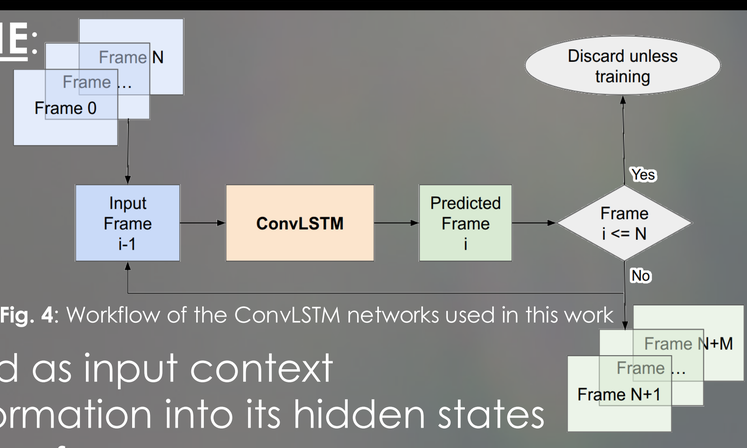

Fig. 4: Workflow of the ConvLSTM networks used in this work

- First 16 frames are used as input context
  - Extracting their information into its hidden states
- Model predicts 15 future frames
- **Autoregressive**: Each predicted frame is fed back as input

EVALUATION: In a shared semantic label space

- RGB and masks are not directly comparable
- Outputs are **mapped** into the **same** semantic **space**
- Enables fair, representation-focused comparison

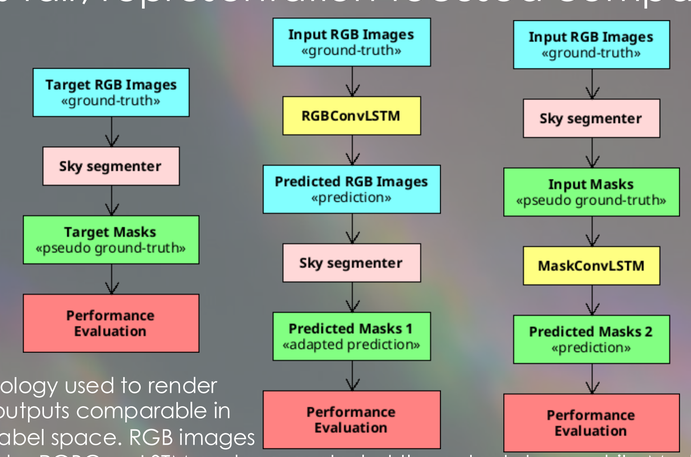

Fig. 5: Methodology used to render both models' outputs comparable in the semantic label space. RGB images are processed by RGBConvLSTM and segmented at the output stage, while MaskConvLSTM operates on segmented inputs. Target RGB images are segmented to obtain the proxy reference masks.

QUALITATIVE COMPARISON EXAMPLE:

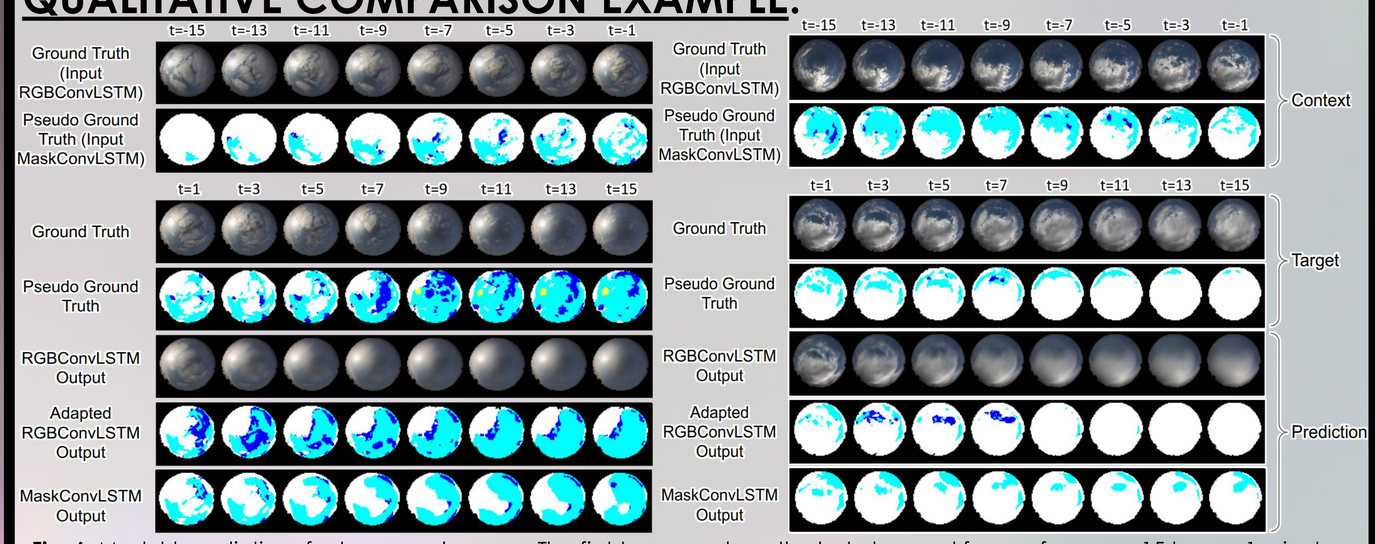

Fig. 6: Models' predictions for two sample cases. The first two rows show the last observed frames from $t = -15$ to $t = -1$ minutes, sampled every 2 minutes. The following two rows present the ground truth future frames from $t = +1$ to $t = +15$ minutes, also sampled every 2 minutes. The last three rows display the models' predictions for the same forecast horizon.

MAIN FINDINGS:

- Semantic masks improve short-term prediction stability
  - IoU: +0.49%
  - Dice Score: +0.94%
- Strongest gains in early forecast horizons
- Trade-off: Less recall · More precision

Fig. 7: Row-normalized confusion matrices averaged across all time steps. Each matrix shows the distribution of predicted classes given the ground truth labels, corresponding to per-class recall.

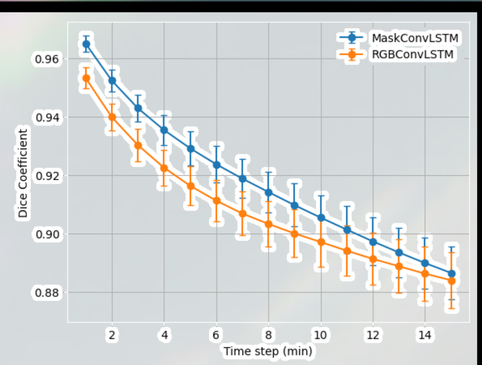

Fig. 8: Mean Dice coefficient across all test sequences for a given timestep. Error bars indicate the 95% confidence interval of the mean.

CONCLUSIONS:

- Input representation affects cloud prediction stability
- Semantic masks yield more stable short-term forecasts
- Improvements are class-dependent
- Segmentation quality is a key limiting factor

REFERENCE: Gatón, J., Román, R., Guzman, C., González-Fernández, D., Longarela, B., Toledano, C., & González, R. (2026). Multi-frame cloud prediction in all-sky images from RGB images and segmented masks. *Solar Energy, 311*, 114515 https://doi.org/10.1016/j.solener.2026.114515

ACKNOWLEDGEMENTS: This work was supported by the Ministerio de Ciencia e Innovacion (MICINN), with the grant no. PID2024-157697OB-I00. This work is part of the project TED2021-131211B-I00375 funded by MCIN/AEI/10.13039/501100011033 and European Union, "NextGenerationEU"/PRTR and is based on work from COST Action CA21119 HARMONIA. Financial support of the Department of Education, Junta de Castilla y León, and FEDER Funds is gratefully acknowledged (Reference: CLU-2023-1-05). The authors acknowledge the support of the Spanish Ministry for Science and Innovation to ACTRIS ERIC. This work was supported as part of EUBURN-RISK (S2/2.4/F0327), an Interreg Sudoe Programme project co-funded by the European Union.