Supplementary Material for "Predicting partially observable dynamical systems via diffusion models with a multiscale inference scheme"

Rudy Morel[*,1], **Francesco Pio Ramunno**[2,3], **Jeff Shen**[4], **Alberto Bietti**[1], **Kyunghyun Cho**[5], **Miles Cranmer**[6], **Siavash Golkar**[1,5], **Olexandr Gugnin**[7], **Geraud Krawezik**[1], **Tanya Marwah**[1], **Michael McCabe**[1,5], **Lucas Meyer**[1], **Payel Mukhopadhyay**[6,8], **Ruben Ohana**[1], **Liam Parker**[1,8], **Helen Qu**[1], **François Rozet**[9], **K.D. Leka**[10,11], **François Lanusse**[1,12], **David Fouhey**[5], **Shirley Ho**[1,4,5]

The Polymathic AI Collaboration

[1]Flatiron Institute, [2]University of Geneva, [3]FHNW, [4]Princeton University, [5]New York University, [6]University of Cambridge, [7]University of Kyiv, [8]University of California, Berkeley, [9]University of Liège, [10]NorthWest Research Associates, [11]Nagoya University, [12]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM.

Contents

[*]Contact: rmorel@flatironinstitute.org

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

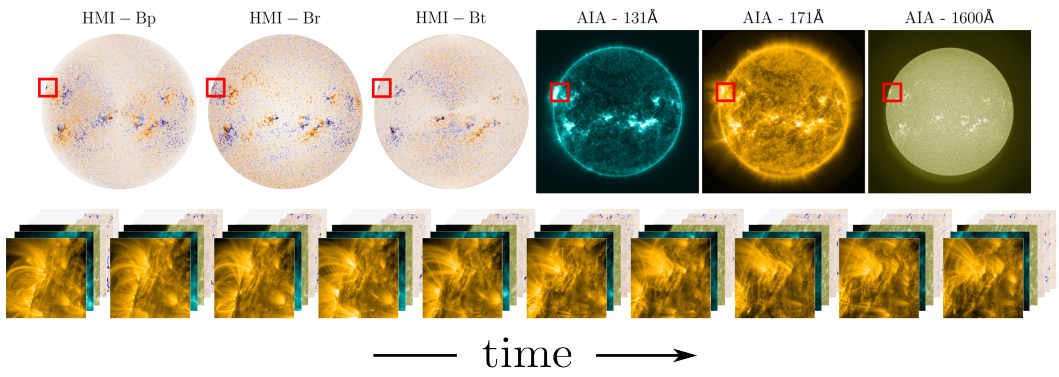


Figure 1: **Solar dataset. (Top):** Full-disk examples of the input channels used in our training dataset. The first three images show the components of the HMI vector magnetic field: the poloidal component B_{p}, the radial component B_{r}, and the toroidal component B_{t}, respectively. The remaining three images correspond to AIA EUV and UV channels at 131 Å, 171 Å, and 1600 Å. These full-disk images are from NASA SDO [16] and preprocessed according to [4]. **(Bottom):** Sample from our curated active region dataset. It shows a spatially aligned sequence of magnetograms and atmospheric images extracted from tracked active regions over a 48-hour window.

1 Solar dataset

1.1 Full-disk images

We construct our dataset from full-disk observations captured by NASA's Solar Dynamics Observatory (SDO) [16], which has been in continuous operation since 2010. In particular, we use two instruments aboard SDO: the Helioseismic and Magnetic Imager [17] and the Atmospheric Imaging Assembly [10], which together provide a comprehensive view of solar activity across the surface and atmosphere.

Magnetograms (surface of the Sun). The Helioseismic and Magnetic Imager [17, HMI] captures vector magnetograms of the solar photosphere, measuring the magnetic field in three orthogonal components. These observations are acquired at a cadence of 12 minutes and a spatial resolution of 1 arcsecond per pixel, producing 4096×4096 pixel full-disk images. The magnetic field values span an average dynamic range from -3000 to $+3000$ Gauss. Since direct measurement of magnetic fields in the solar corona, the Sun's outermost atmospheric layer, HMI magnetograms serve as the primary constraint on the magnetic environment of the outer solar atmosphere. They are thus essential for studying the magnetic drivers of solar activity. The temperature at the photosphere is ~ 4500 K. An example of the 3d vector magnetic field captured from HMI is shown in Figure 1.

Atmospheric images (atmosphere of the Sun) The Atmospheric Imaging Assembly [10, AIA] complements HMI by observing the upper layers of the Sun, ranging from the chromosphere to the outer corona, using multiple ultraviolet (UV) and extreme ultraviolet (EUV) channels. AIA operates at a cadence of 12 seconds with a spatial resolution of approximately 1.5 arcseconds, capturing dynamic atmospheric phenomena across a range of temperatures (from $\sim 10^4$ K to beyond 10^7 K). These passbands reveal radiative signatures of flares, eruptions, and coronal loops. While AIA does not directly measure magnetic fields, it provides crucial indirect evidence of the coronal response to magnetic activity rooted in the photosphere. A representative set of multi-channel AIA images is shown in Figure 1.

The raw data produced by these instruments is curated in a dataset introduced in [4], which includes preprocessing steps such as degradation correction, removal of faulty observations, and temporal co-registration.

1.2 Sun regions dataset

While full-disk data offer a comprehensive view of the Sun, the majority of solar activity relevant to forecasting tasks is concentrated in localized regions known as active regions. These regions, though occupying only a small fraction of the solar disk, are the primary sources of variability and eruptive events. To concentrate on the most relevant areas and reduce computational cost, we build our dataset from video crops centered on tracked active regions.

Spatial sampling of active regions. From 2013 to 2019, we sample 8 spatial locations per day using a probabilistic strategy based on the absolute value of the radial magnetic field component ($|B_r|$) from HMI magnetograms. This field component serves as an effective proxy for activity since high $|B_r|$ values correlate strongly with the likelihood of solar eruptions [8, 2]. The sampling probability at each pixel location x is defined as proportional to $\exp(|B_r(x)|/T)$, where T is a tunable temperature-like parameter controlling the sharpness of selection. This prioritizes magnetically active zones while preserving some randomness to avoid bias toward rare extreme events. Regions below a minimum signal-to-noise threshold are excluded, and spatial diversity is enforced via a minimum distance constraint between samples.

Temporal tracking and data extraction Once locations are selected, we track their motion across the solar disk using precomputed differential rotation maps provided by the curated dataset of [4]. This tracking compensates for the Sun's differential rotation, allowing us to follow the same region over time. For each selected location, we extract a 512×512 pixel crop every hour over a 48-hour window, resulting in a temporally consistent sequence. Each crop includes 12 channels: the three orthogonal components of the HMI magnetic field and nine co-aligned AIA EUV/UV channels.

Train/val/test datasets. In order to reduce the computational cost of our trainings, the patches are downsampled to 256×256 and select only 3 AIA channels: the 131 Å, 171 Å, and 1600 Å bands. Each resulting sample is a spatiotemporal tensor of shape $48 \times 12 \times 256 \times 256$, representing a 48-hour evolution of solar activity within an active region. We partition the dataset into training (70%), validation (15%), and test (15%) sets. The training set spans January 2014 to May 2018. The validation set comprises two disjoint intervals: January to May 2013 and July 2018 to January 2019. The test set includes data from August to November 2013 and from March to September 2019. To prevent temporal leakage and ensure unbiased evaluation, we insert one-month buffer zones between splits, avoiding overlap of active regions across different subsets. The test periods are deliberately chosen to span both solar maximum and minimum phases (see Figure 2), supporting a robust assessment of model generalization under diverse solar conditions.

2 Additional model description

2.1 Inference scheme implementation

To allow for fair comparison, all baseline schemes, including the standard autoregressive rollout and the *hierarchy-2* scheme from the FDM model [6], were used with the same temporal and computational budget as our multiscale method. All schemes are restricted to generating exactly three new frames at once, with four context frames, resulting in a fixed window size of $2K + 1 = 7$ frames ($K = 3$). This required adapting the hierarchy-2 baseline to limit the number of generated frames per call. In addition, all models are limited to generating a video which extends at maximum 9 time steps in the future, and 9 time steps in the past. Thus, each generated video spans at maximum $19 = 9 + 1 + 9$ steps (see for example our multiscale inference scheme in Fig. 4, horizon 9).

Autoregressive scheme. The autoregressive baseline progresses through time using a fixed-length context window. At each step, it generates three future frames conditioned on the most recent four, repeating this process uniformly across the sequence.

Hierarchy-2 scheme. The hierarchy-2 scheme, as described in the FDM framework [6], first generates a coarse sequence of future frames and then fills in the intermediate steps through recursive refinement. In our experiments, this scheme is adapted to output three frames per step while respecting

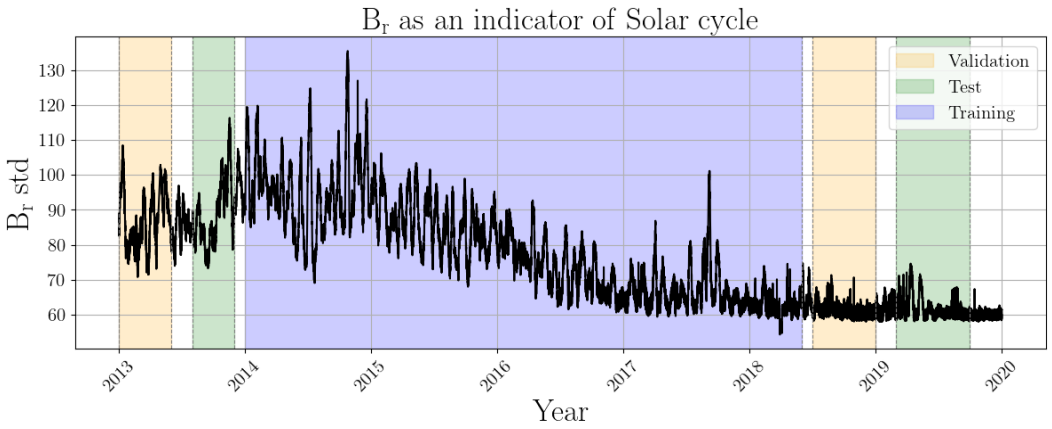


Figure 2: **Solar activity.** The standard deviation of the B_r component over the full solar disk as a proxy for solar activity, illustrating the temporal segmentation used for model development. The dataset is divided into training (blue), validation (yellow), and test (green) intervals, with each evaluation segment separated from training data by at least one full month. The split ensures disjoint active regions across sets and includes test intervals during both high and low solar activity, enabling robust model evaluation across varying solar conditions.

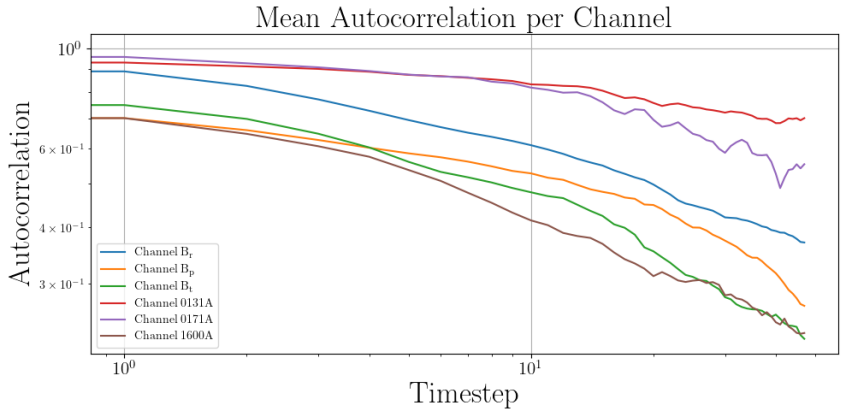


Figure 3: **Long-memory on solar data.** Autocorrelation as a function of time for several representative channels in our dataset. All channels exhibit a slow and smooth decay of autocorrelation over time, confirming the presence of persistent long-range dependencies. This motivates the use of our multiscale inference scheme, which uses exponentially spaced temporal templates to efficiently capture long-range dependencies.

the 19-frame context limit set above for fair comparisons. Despite this adjustment, the method retains its hierarchical trajectory structure, providing an alternative to the autoregressive rollouts.

Multiscale scheme (ours). Our proposed multiscale inference scheme relies on multiscale templates to capture long-range dependencies efficiently. This design exploits the long-memory characteristics of physical systems like solar dynamics, enabling the model to integrate coarse and fine temporal context. The reasoning behind our multiscale template stems from the long-memory property inherent to such physical processes, as illustrated in Figure 3. This phenomenon has been extensively studied in the context of scale-invariant processes [14, 1, 12, 15], particularly through the use of wavelet-based analysis [19]. Once the inference strategy is fixed (see Fig. 4, with a horizon of 9), the corresponding templates, and masks defining the conditioning and generated data, are used during training.

Positional encodings. To enable our model to generate videos with varying time steps, we incorporate temporal information at multiple stages of the denoiser. First, time indices of the frames

are added as input channels. Second, in the ViT denoiser, attention layers incorporate relative time indices using a RoPE [20] positional encoding. In addition to frame time positional encodings, we also add pixel-wise latitude and longitude as additional channels to all models, facilitating accurate predictions near the limb of the Sun.

2.2 Baseline architectures

In addition to aligning inference schemes, we ensure that all model architectures are compared under identical training conditions. Each model is trained for 40 epochs on the same dataset, with consistent preprocessing, with similar hyperparameters (see below). In particular, all models have roughly 60M parameters.

Axial ViT (AViT). The Axial Vision Transformer (AViT) [13] employs axial attention to model spatiotemporal dependencies efficiently in high-dimensional sequences. In our paper, an AViT is paired with the standard autoregressive inference scheme, it takes four frames as input (the conditioning) and predicts the three following frames. As a fully deterministic model, it serves as a strong baseline for assessing the value of stochastic modeling in solar forecasting.

Auto-regressive diffusion (AR-diffusion). The AR-diffusion model [7] applies diffusion sampling in a step-by-step autoregressive fashion. At each iteration of the sampling process, the model generates the next state conditioned on 4 previous steps.

2.3 Training hyperparameters

We use the AdamW optimizer with a learning rate of 1×10^{-4}, cosine learning rate scheduling, and a batch size of 64. Input patches are of shape $1 \times 8 \times 8$, and the model consists of 16 transformer blocks with 4 attention heads each. All models are trained on 40 epochs. Each epoch consists of 2000 batches with a batch size of 64, covering most of the training dataset. All models were trained using a single node with 8 NVIDIA H100 GPUs.

3 Multiscale inference schemes with longer horizons

In the experiments we chose multiscale templates of size $2K + 1 = 7$ and a longest template $\mathbb{T} = \{-9, -3, -1, 0, 1, 3, 9\}$, with a horizon of 9 steps in both past and future directions. This inference scheme, shown in Fig. 4, was used in all experiments.

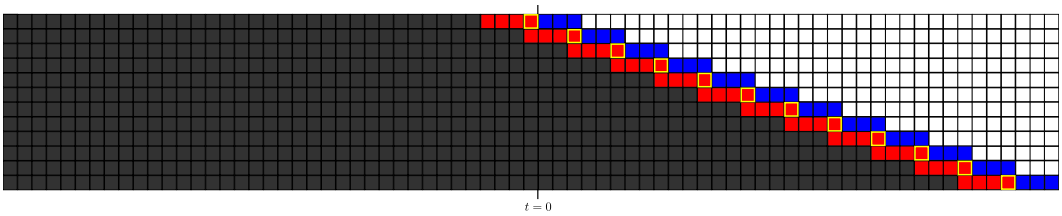
To better capture long-range dependencies, one can consider longer templates and more complex multiscale inference schemes. Fig. 4 shows multiscale inference schemes based on a longest template $\mathbb{T} = \{-18, -4, -1, 0, 1, 4, 18\}$ with horizon 18, and on a longest template $\mathbb{T} = \{-36, -6, -1, 0, 1, 6, 36\}$ with horizon 36. Algorithm [**?**] presents a general inference scheme.

4 Additional model evaluation on solar dynamics

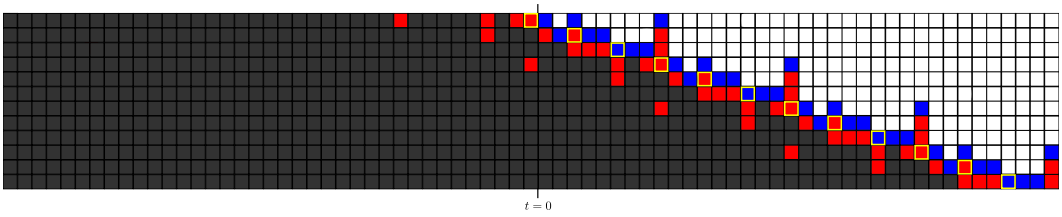
4.1 Physical parameters evaluation

We evaluate the physical validity of predictions using SHARP parameters (Table 1) commonly used in solar physics to characterize active regions and their flare potential [2, 8, 9]. Just like the power spectrum, these statistics are computed on a single state.

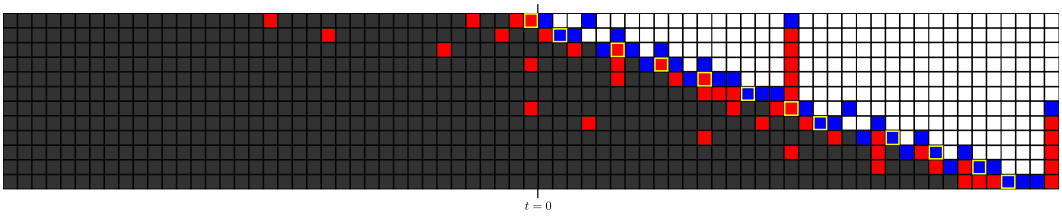
SHARPs computation. We first map our vector magnetic field into the Cylindrical Equal-Area (CEA) system of reference. This is done because it is important to be in a uniform coordinate grid (equal area per pixel), while in native HMI CCD coordinates pixel scale varies across the field of view due to projection effects, which makes it invalid to integrate or compare pixel by pixel across the region. CEA corrects for that by projecting the data into a grid where each pixel covers the same surface area on the solar sphere. While SHARP includes sixteen parameters in total, we focus on three representative and intuitive quantities: the total unsigned magnetic flux, the horizontal gradient of the total magnetic field, and the horizontal gradient of the vertical magnetic field. This selection allows us to analyze all three components of the vector magnetic field while maintaining interpretability and physical relevance.



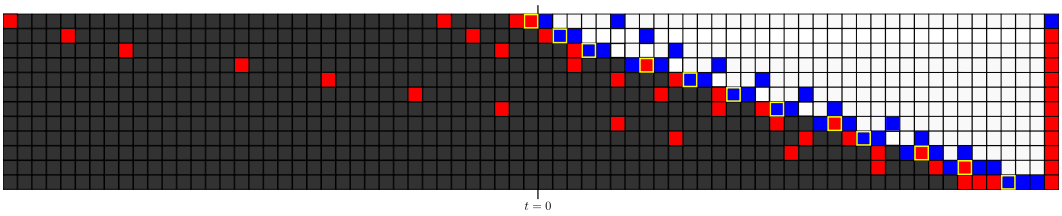
Autoregressive



Multiscale with horizon 9 (used in experiments)



Multiscale with horizon 18



Multiscale with horizon 36

Figure 4: **Multiscale inference schemes with various horizons.** The horizon denotes the furthest time step the model can predict in a single shot under a given inference scheme. For example, a horizon of 3 corresponds to standard autoregressive prediction. Increasing the horizon allows the model to be conditioned more frequently on the past, including distant steps, enabling more stable long-range forecasting at the same inference cost.

Total unsigned flux (UsFlux). The total unsigned flux is computed from the radial component of the vector magnetic field, $|B_z|$, and represents the total absolute magnetic flux through the area A of the active region:

$$\Phi = \int |B_z| \, dA \simeq \sum |B_z| \, dA. \tag{1}$$

This quantity measures the amount of magnetic energy stored in the region and serves as a proxy for its magnetic complexity. The unsigned flux is usually directly proportional to the flaring probability of the region [11].

Horizontal Gradient of Total Field (MeanGBT). The horizontal gradient of the total magnetic field is defined as:

$$\overline{|\nabla B_{\text{tot}}|} = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}, \quad \text{with} \quad B = \sqrt{B_x^2 + B_y^2 + B_z^2}, \tag{2}$$

Algorithm 1 Multiscale inference scheme for generic future horizon H, and time steps generated at once K.

Input: Integer future horizon H, generation size K (time steps), template set $\mathcal{T} = \{\tau^{(1)}, \ldots, \tau^{(N)}\}$ ordered by increasing horizon
Output: Action list $A = [(n, \text{shift}, \text{mask})_1, \ldots, (n, \text{shift}, \text{mask})_M]$ specifying template index, time shift, and conditioning mask at each generation step
Initialize: completed$[t] \leftarrow$ False for $t = 1, \ldots, H$
Initialize: $A \leftarrow []$
while $\exists t \in \{1, \ldots, H\}$ with completed$[t] =$ False **do**
  best $\leftarrow$ None, best_score $\leftarrow \infty$
  for $n = N$ **to** 1 **do**
    for shift $= 0$ **to** H **do**
      if $\max(\tau^{(n)}) +$ shift $> H$ **then**
        continue
      end if
      $I \leftarrow \{t + \text{shift} : t \in \tau^{(n)}\}$
      # Check that the shifted template is conditioned on exactly $K + 1$ steps
      overlap $\leftarrow |\{t \in I : \text{completed}[t] = \text{True}\}|$
      if overlap $= K + 1$ **then**
        $I_{\text{future}} \leftarrow \{t \in I : t > 0\}$
        $\mathcal{G} \leftarrow \{t \in \{1, \ldots, H\} : \text{completed}[t] = \text{True}\}$
        $L \leftarrow |\mathcal{G} \setminus I_{\text{future}}|$ *# Steps already generated not covered by the template*
        $a \leftarrow (\max(\tau^{(n)}) + \text{shift} = H)$
        mask $\leftarrow [\text{completed}[t]]_{t \in I}$ *# Store which indices are conditioned*
        # Favors a template which anchors on the very last future step
        if $a =$ True **then**
          best $\leftarrow (n, \text{shift}, \text{mask})$
          break both loops
        else
          # Favors maximal self-conditioning for temporal coherence
          if $L <$ best_score **then**
            best $\leftarrow (n, \text{shift}, \text{mask})$
            best_score $\leftarrow L$
          end if
        end if
      end if
    end for
  end for
  if best $\neq$ None **then**
    Unpack $(n_{\text{best}}, \text{shift}_{\text{best}}, \text{mask}_{\text{best}}) \leftarrow$ best
    $I_{\text{best}} \leftarrow \{t + \text{shift}_{\text{best}} : t \in \tau^{(n_{\text{best}})}, t > 0\}$
    Mark completed$[t] \leftarrow$ True for all $t \in I_{\text{best}}$
    Append $(n_{\text{best}}, \text{shift}_{\text{best}}, \text{mask}_{\text{best}})$ to A
  else
    break
  end if
end while
return A

Table 1: **Predictions performances measured with physical parameters.** We compare different inference schemes (Autoregressive, Hierarchy-2 [6], Ours – Multiscale). For each, we evaluate at three different time windows (1-4 hours, 4-16 hours, 16-32 hours) using multiple metrics: the normalized mean absolute error of representative solar physics quantities from [3] – the unsigned flux (UsFlux), the Mean Horizontal Gradient of the Total Field (MeanGBT) and of the Vertical Field (MeanGBZ)

				Relative error		
Model	Denoiser	Scheme	SHARP	1:4	4:16	16:32
DiT	ViT	Autoreg.	UsFlux	0.25	0.40	0.50
DiT	ViT	Hierarchy-2 [6]	UsFlux	**0.19**	**0.38**	0.52
DiT	ViT	Multiscale (ours)	UsFlux	0.23	**0.38**	**0.48**
DiT	ViT	Autoreg.	MeanGBT	0.18	0.30	0.37
DiT	ViT	Hierarchy-2 [6]	MeanGBT	**0.12**	0.28	0.38
DiT	ViT	Multiscale (ours)	MeanGBT	0.14	**0.27**	**0.33**
DiT	ViT	Autoreg.	MeanGBZ	0.15	0.25	0.31
DiT	ViT	Hierarchy-2 [6]	MeanGBZ	**0.088**	0.22	0.31
DiT	ViT	Multiscale (ours)	MeanGBZ	0.10	**0.21**	**0.27**

and N denoting the number of pixels over which the sum is computed. This parameter quantifies how rapidly the total magnetic field strength changes across the horizontal plane. High values of this gradient indicate complex magnetic structures with strong shear or twist, which are typically associated with the onset of solar flares [11].

Horizontal Gradient of Vertical Field (MeanGBZ). The horizontal gradient of the radial (vertical) component B_z is given by:

$$\overline{|\nabla B_z|} = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}, \tag{3}$$

with N the number of pixels over which the sum is computed. This metric reflects how quickly the radial field changes in the horizontal direction. High values of $\overline{|\nabla B_z|}$ often appear near polarity inversion lines, where the magnetic field polarity reverses. These regions are important because they are commonly associated with magnetic reconnection events that can trigger solar flares [18].

Tab. 1 shows the Normalized Mean Absolute Error (NMAE) between predicted and observed values for each SHARP parameter. This metric is defined as the mean absolute error normalized by the mean absolute value of the observed parameter. Specifically, for a given metric M, the NMAE is defined as:

$$\text{NMAE}(M) = \frac{1}{N} \sum_{i=1}^{N} \frac{|M_i - M_i^{\text{obs}}|}{|M_i^{\text{obs}}|},$$

where M_i is computed on a predicted state, and M_i^{obs} is computed on the observed state.

4.2 Predicted trajectories

Additional examples of predicted trajectories are shown for different inference schemes (Fig. 5,6,7) and for different models, with their respective preferred inference schemes (Fig. 8,9,10).

5 Additional experiments on a fluid dynamics dataset

To showcase the generality of our multiscale inference scheme, we applied it to the Navier-Stokes data from PDEArena [5]. To make the dynamical system partially observable, we downsampled it to 32×32 (from 128×128) and considered only the density field (discarding velocity). The results on this new example of partially observable process are presented in Tab. 2 below using the same metric as in the main paper (see MAE on the power-spectrum in Tab. 1).

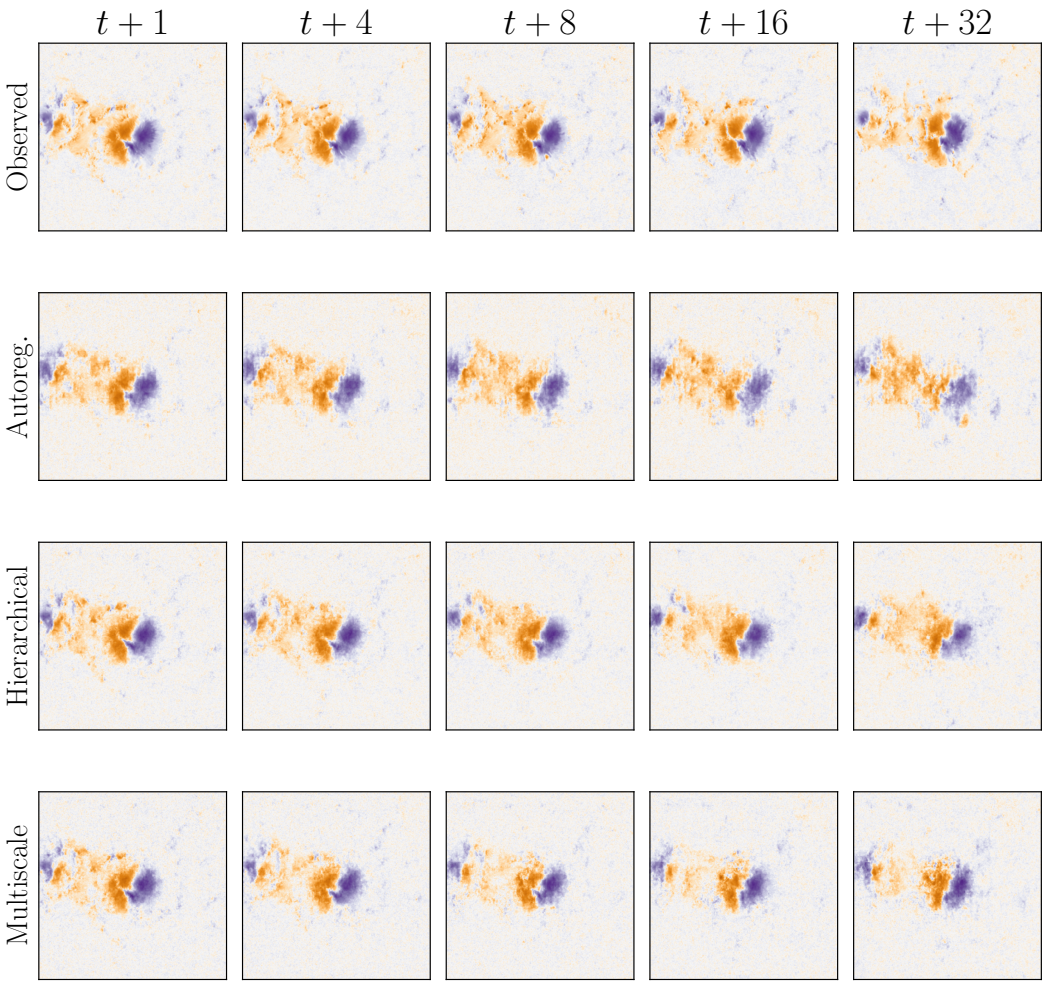


Figure 5: **Example of predictions (1/3), for different inference schemes.** From top to bottom: observed data, autoregressive, hierarchy-2 [6], multiscale (ours).

Table 2: **Synthetic example: partially observable fluid dynamics.** We compare different inference schemes (Autoregressive and Multiscale – ours) on the synthetic example, evaluated over three time windows: 1–4, 4–16, and 16–36 steps. The metric is the Mean Absolute Error (MAE) of the power spectrum.

	MAE Power Spectrum		
Inference scheme	1:4	4:16	16:36
Autoregressive	0.39	0.43	0.35
Multiscale (ours)	**0.38**	**0.36**	**0.31**

We observe that our multiscale inference scheme outperforms the autoregressive baseline, particularly in the long-term interval (16–36), where it achieves a significantly lower MAE on the power spectrum. This demonstrates that our approach is not specific to solar dynamics prediction but can be applied successfully to other partially observable systems.

References

[1] Patrice Abry, Patrick Flandrin, Murad S Taqqu, and Darryl Veitch. Wavelets for the analysis, estimation, and synthesis of scaling data. *Self-similar network traffic and performance evaluation*,

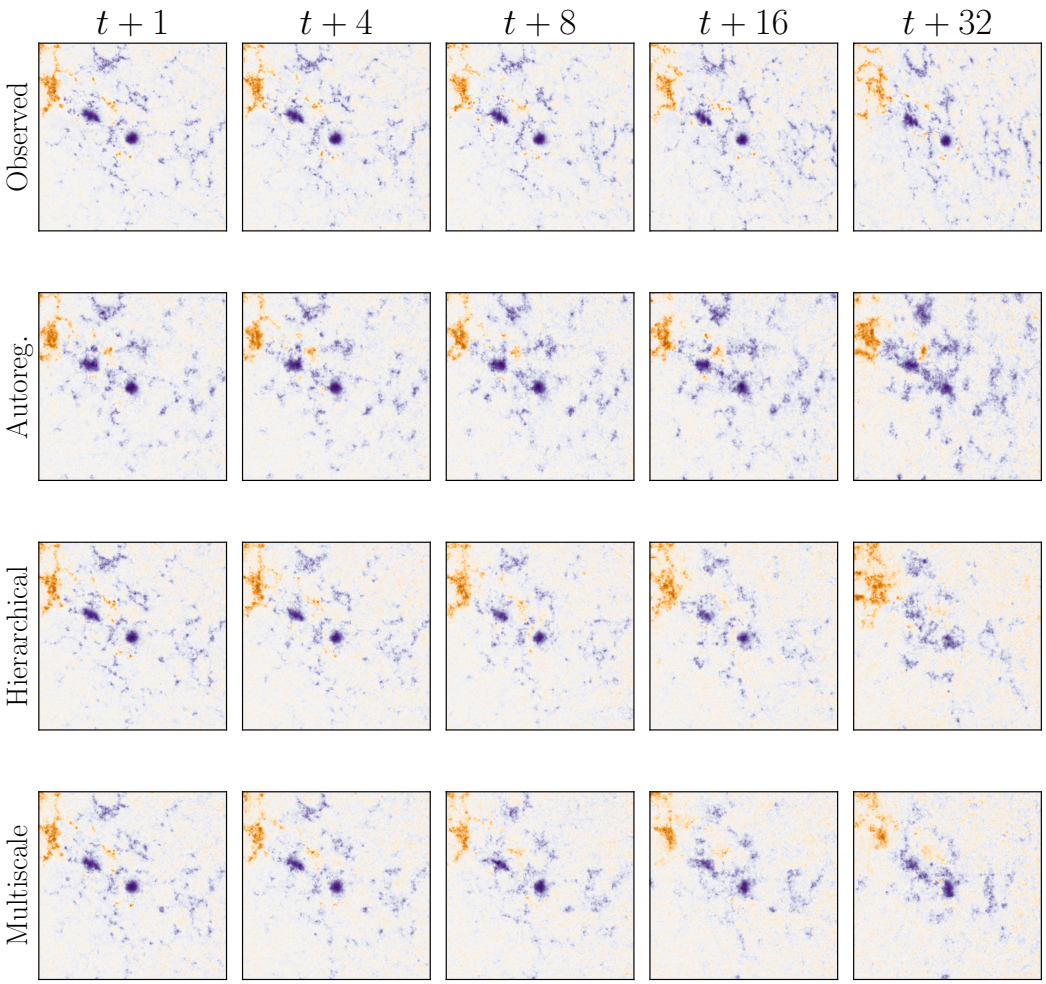


Figure 6: **Example of predictions (2/3), for different inference schemes.** From top to bottom: observed data, autoregressive, hierarchy-2 [6], multiscale (ours).

pages 39–88, 2000.

[2] M. G. Bobra and S. Couvidat. Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-learning Algorithm. *The Astrophysical Journal*, 798(2):135, January 2015.

[3] M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. D. Leka. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs - Space-Weather HMI Active Region Patches. *Solar Physics*, 289:3549–3578, September 2014.

[4] Richard Galvez, David F Fouhey, Meng Jin, Alexandre Szenicer, Andrés Muñoz-Jaramillo, Mark CM Cheung, Paul J Wright, Monica G Bobra, Yang Liu, James Mason, et al. A machine-learning data set prepared from the nasa solar dynamics observatory mission. *The Astrophysical Journal Supplement Series*, 242(1):7, 2019.

[5] Jayesh K Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling. *arXiv preprint arXiv:2209.15616*, 2022.

[6] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.

[7] Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. *arXiv preprint arXiv:2309.01745*, 2023.

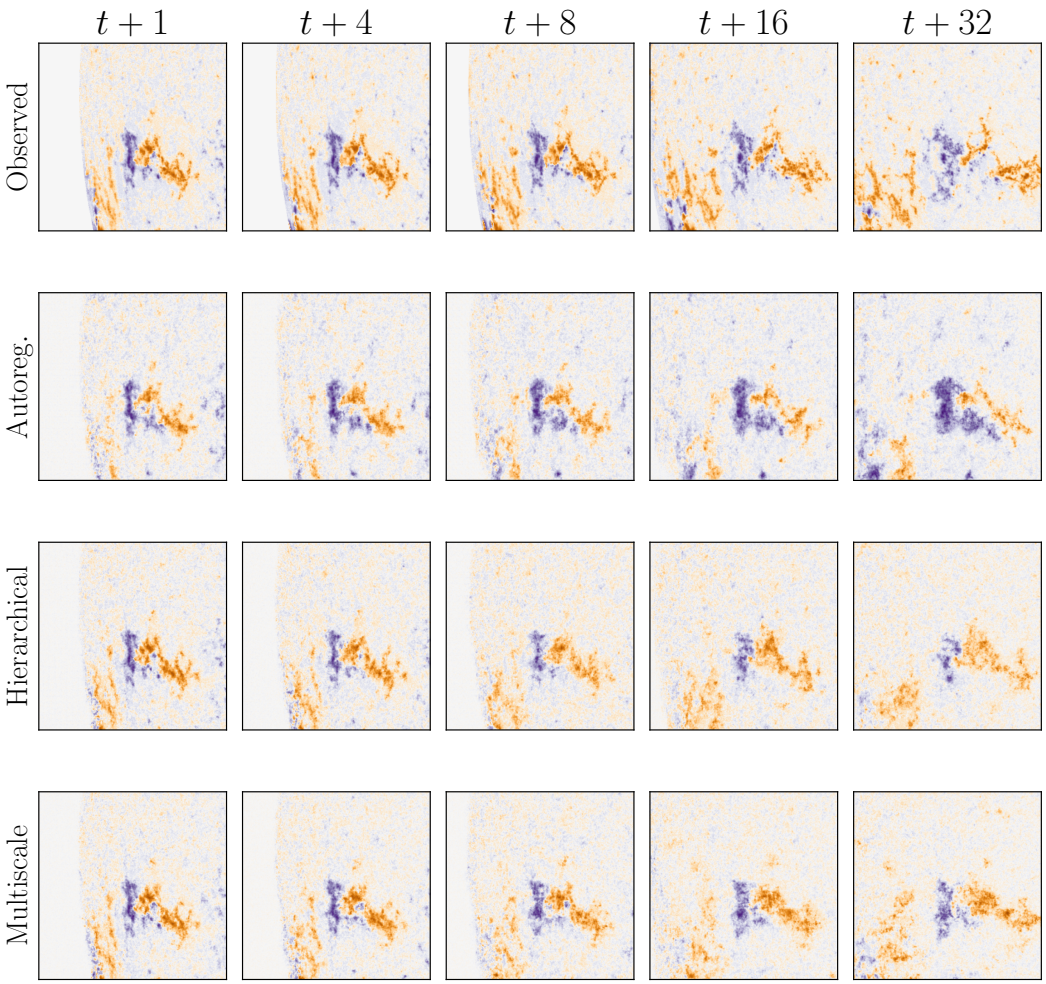


Figure 7: **Example of predictions (3/3), for different inference schemes.** From top to bottom: observed data, autoregressive, hierarchy-2 [6], multiscale (ours).

[8] K. D. Leka and G. Barnes. Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. I. Data, General Approach, and Sample Results. , 595(2):1277–1295, October 2003.

[9] K. D. Leka and G. Barnes. Photospheric Magnetic Field Properties of Flaring vs. Flare-Quiet Active Regions. IV: A Statistically Significant Sample. *The Astrophysical Journal*, 656:1173–1186, 2007.

[10] James R. Lemen, Alan M. Title, David J. Akin, Paul F. Boerner, Catherine Chou, Jerry F. Drake, Dexter W. Duncan, Christopher G. Edwards, Frank M. Friedlaender, Gary F. Heyman, Neal E. Hurlburt, Noah L. Katz, Gary D. Kushner, Michael Levay, Russell W. Lindgren, Dnyanesh P. Mathur, Edward L. McFeaters, Sarah Mitchell, Roger A. Rehse, Carolus J. Schrijver, Larry A. Springer, Robert A. Stern, Theodore D. Tarbell, Jean-Pierre Wuelser, C. Jacob Wolfson, Carl Yanari, Jay A. Bookbinder, Peter N. Cheimets, David Caldwell, Edward E. Deluca, Richard Gates, Leon Golub, Sang Park, William A. Podgorski, Rock I. Bush, Philip H. Scherrer, Mark A. Gummin, Peter Smith, Gary Auker, Paul Jerram, Peter Pool, Regina Soufli, David L. Windt, Sarah Beardsley, Matthew Clapp, James Lang, and Nicholas Waltham. The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):17–40, January 2012.

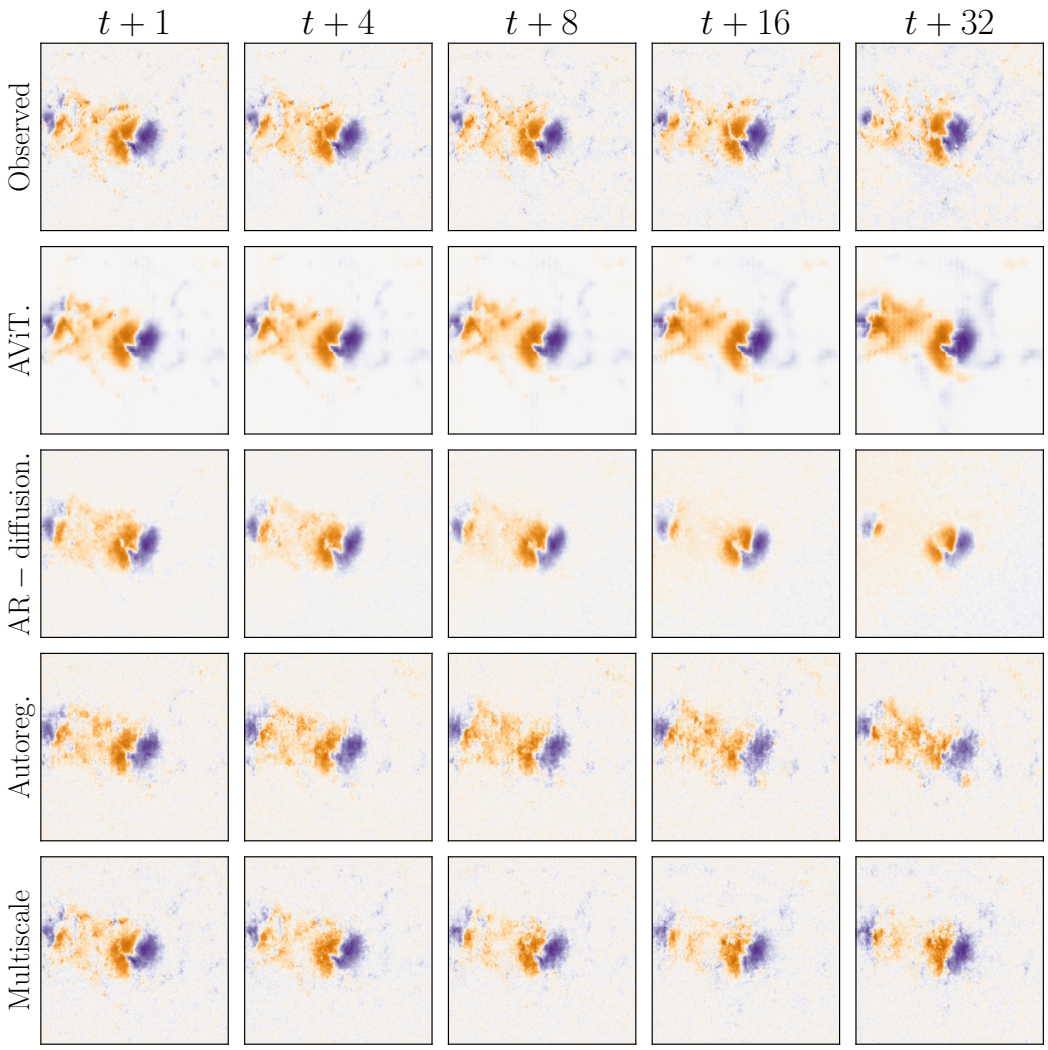


Figure 8: **Example of predictions (1/3), for different models.** From top to bottom: observed data, AViT [13] model, autoregressive-diffusion model [7], our model with an autoregressive inference scheme, our model with a multiscale inference scheme.

[11] Ting Li, Anqin Chen, Yijun Hou, Astrid M. Veronig, Shuhong Yang, and Jun Zhang. Magnetic flux and magnetic nonpotentiality of active regions in eruptive and confined solar flares. *The Astrophysical Journal Letters*, 917(2):L29, aug 2021.

[12] Benoit B Mandelbrot. *Multifractals and 1/f noise: Wild self-affinity in physics (1963–1976)*. Springer, 2013.

[13] Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. Multiple physics pretraining for spatiotemporal surrogate models. *Advances in Neural Information Processing Systems*, 37:119301–119335, 2024.

[14] EJ McCoy and AT Walden. Wavelet analysis and synthesis of stationary long-memory processes. *Journal of computational and Graphical statistics*, 5(1):26–56, 1996.

[15] Rudy Morel, Gaspar Rochette, Roberto Leonarduzzi, Jean-Philippe Bouchaud, and Stéphane Mallat. Scale dependencies and self-similar models with wavelet scattering spectra. *Applied and Computational Harmonic Analysis*, 75:101724, 2025.

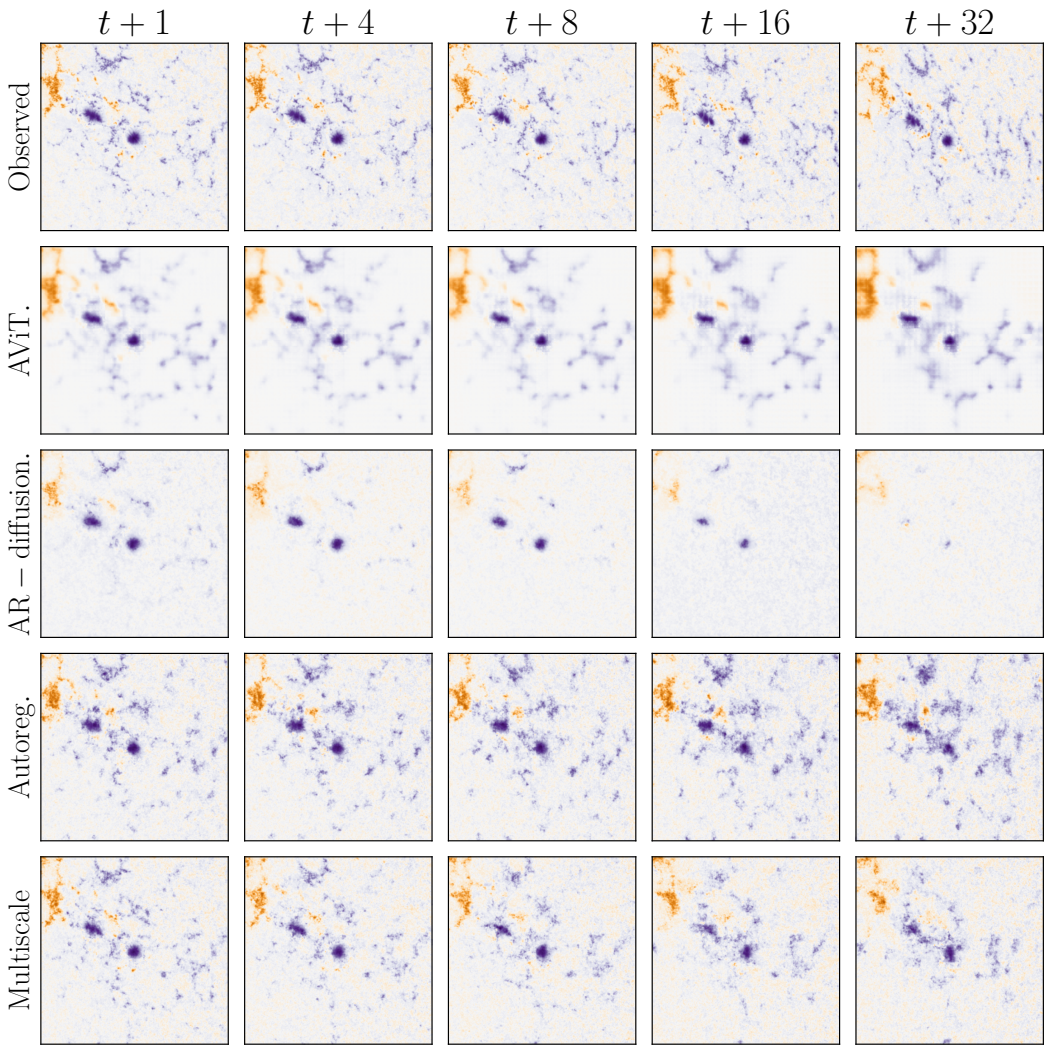


Figure 9: **Example of predictions (2/3), for different models.** From top to bottom: observed data, AViT [13] model, autoregressive-diffusion model [7], our model with an autoregressive inference scheme, our model with a multiscale inference scheme.

[16] W. Dean Pesnell, B. J. Thompson, and P. C. Chamberlin. The Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):3–15, January 2012.

[17] P. H. Scherrer, J. Schou, R. I. Bush, A. G. Kosovichev, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, J. Zhao, A. M. Title, C. J. Schrijver, T. D. Tarbell, and S. Tomczyk. The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):207–227, January 2012.

[18] I. N. Sharykin, I. V. Zimovets, and I. I. Myshyakov. Flare energy release at the magnetic field polarity inversion line during the m1.2 solar flare of 2015 march 15. ii. investigation of photospheric electric current and magnetic field variations using hmi 135 s vector magnetograms. *The Astrophysical Journal*, 893(2):159, apr 2020.

[19] Mallat Stephane. A wavelet tour of signal processing, 1999.

[20] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *arXiv e-prints*, page arXiv:2104.09864, April 2021.

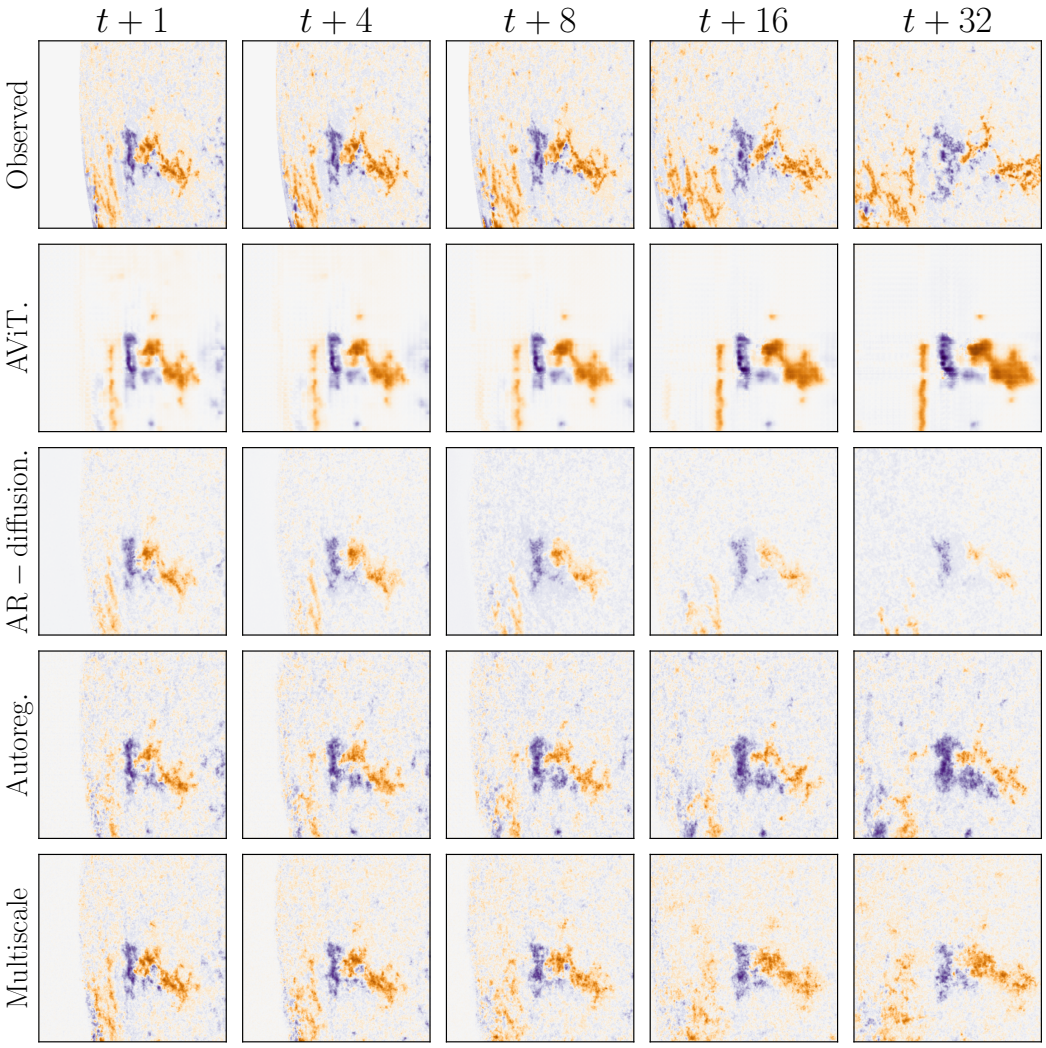


Figure 10: **Example of predictions (3/3), for different models.** From top to bottom: observed data, AViT [13] model, autoregressive-diffusion model [7], our model with an autoregressive inference scheme, our model with a multiscale inference scheme.