Generating Physically-Consistent Satellite Imagery for Climate Visualizations

Björn Lütjens, Brandon Leshchinskiy, Océane Boulais, Farrukh Chishtie, Natalia Díaz-Rodríguez, Margaux Masson-Forsythe, Ana Mata-Payerro, Christian Requena-Mesa, Aruna Sankaranarayanan, Aaron Piña, Yarin Gal, Chedy Raïssi, Alexander Lavin, and Dava Newman

Abstract—Deep generative vision models are now able to synthesize realistic-looking satellite imagery. However, the possibility of hallucinations prevents their adoption of risk-sensitive applications, such as generating materials for communicating climate change. To demonstrate this issue, we train a generative adversarial network (GAN, pix2pixHD) to create synthetic satellite imagery of future flooding and reforestation events. We find that a pure deep learning-based model can generate photorealistic flood visualizations but hallucinate floods at locations that are not susceptible to flooding. To address this issue, we propose to condition and evaluate generative models on segmentation maps of physics-based flood models. We show that our physics-conditioned model outperforms the pure deep learning-based model and a handcrafted baseline. We evaluate the generalization capability of our method to different remote sensing data and different climate-related events (reforestation).

We publish our code and dataset which includes the data for a third case study of melting Arctic sea ice and >30,000 labeled HD image triplets—or the equivalent of 5.5 million images at 128 × 128 pixels—for segmentation guided image-to-image (im2im) translation in Earth observation. Code and data are available at github.com/blutjens/eie-earth-public.

Received 8 March 2024; revised 17 August 2024; accepted 23 September 2024. Date of publication 19 November 2024; date of current version 6 December 2024. This work was supported in part by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator through Cooperative Agreement FA8750-19-2-1000; and in part by MIT Portugal Program, National Aeronautics and Space Administration (NASA), and Google Cloud (for the duration of FDL). (Océane Boulais, Farrukh Chishtie, Natalia Díaz-Rodríguez, Margaux Masson-Forsythe, Ana Mata-Payerro, Christian Requena-Mesa, and Aruna Sankaranarayanan contributed equally to this work.) (Corresponding author: Björn Lütjens.) Björn Lütjens, Brandon Leshchinskiy, Ana Mata-Payerro, and Dava Newman are with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139 USA (e-mail: luetjens@mit.edu).

Océane Boulais is with the Scripps Institution of Oceanography, La Jolla, CA 92037 USA.
Farrukh Chishtie is with The University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and also with the Peaceful Society, Science and Innovation Foundation, Petaluma, CA 94953 USA.
Natalia Díaz-Rodríguez is with the Department of Computer Science and AI, University of Granada, 18016 Granada, Spain.
Margaux Masson-Forsythe is with the Surgical Data Science Collective, Washington, DC 20006 USA.
Christian Requena-Mesa is with the Max Planck Institute for Biogeochemistry, 07745 Jena, Germany.
Aruna Sankaranarayanan is with the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139 USA.
Aaron Piña is with U.S. Forest Service, Washington, DC 20259 USA.
Yarin Gal is with the Department of Computer Science, OX1 2JD Oxford, U.K.
Chedy Raïssi is with INRIA, 54600 Villers-les-Nancy, France.
Alexander Lavin is with the Pasteur Labs & ISI, New York City, NY 11205 USA.

This article has supplementary downloadable material available at https://doi.org/10.1109/TGRS.2024.3493763, provided by the authors.

Index Terms—Climate change, deep generative vision models, flooding, generative adversarial networks (GANs), generative AI, image-to-image (im2im) translation, physics-informed machine learning (ML), reforestation, remote sensing, synthetic satellite imagery, visualization.

I. INTRODUCTION

W ith climate change, natural disasters are becoming more intense [1]. Floods are the most frequent weather-related disaster [2] and already cost the U.S. 4.1 B USD per year [3]; this damage is projected to grow over the next decades [1].

Visualizations of climate impacts are widely used by policy and decision-makers to raise environmental awareness and facilitate dialog on long-term climate adaptation decisions [4]. Visualizations of flood risks, for example, are used in local policy-making and community discussion groups as decision-aids for flood infrastructure investments [4]. Current visualizations of flood impacts, however, are limited to color-coded flood maps [5], [6], [7] or synthetic street-view imagery [8], [9], which may not convey city-wide flood impacts in a compelling manner [10]. Our work generates synthetic satellite imagery of future coastal floods that are informed by the projections of expert-validated flood models, as illustrated in Fig. 1 and 2. As a geospatial rendering, this imagery might enable a more engaging communication of city-wide flood risks to governmental offices. We focus on deep generative vision models, such as generative adversarial

![Fig. 1. We synthesize satellite imagery that visualizes flooding. We designed the underlying generative vision model to project flooding only in locations that are consistent with a physics-based flood model. The new visualizations could facilitate intuitive and trustworthy communication of climate risks, for example, via tabletop exercises as seen on the left. Explore more results at climate-viz.github.io.](image-url)
networks (GANs) [11], as our method for visualization. Generative vision models have generated photorealistic imagery of faces [12], [13], animals [14], [15], street-level flood imagery [9], and satellite observations [16], [17], [18], [19], [20]. Synthetic satellite imagery, however, needs to be explainable [21], [22] and trustworthy [23]. Many complementary approaches exist to increase the trustworthiness of generative vision models, including interpretable networks [24], adversarial robustness [25], [26], [27], or probabilistic predictions with uncertainty [28], [29], [30]. Here, we raise a new question How can we increase trust in synthetic satellite imagery through physical-consistency? We define a synthetic image to be physically-consistent if the depictions in the image are consistent with the output of a physics-based model, as detailed in Section III-C. Our definition of physical-consistency relates to the field of physics-informed machine learning (ML) in which researchers find novel ways to embed physics domain knowledge into deep learning methods [31], [32], [33], [34], [35]. We considered various physics-informed ML methods to incorporate the physics of floods in a generative vision model as inputs [36], constrained representation [24], [37], [38], training loss [39], hard output constraints [40], [41], [42], or evaluation function [43]. We also considered embedding a generative vision model into a set of physics equations, specifically, in the differential equations of floods as learned parameters [39], [44], dynamics [45], residual [46], [47], differential operator [32], [48], or solution [39]. Finally, we decided to incorporate physics as input and evaluation functions. Specifically, we use a one-channel flood mask, that represents the projections of a physics-based flood model, and three-channel satellite imagery as inputs to a deep generative vision model and evaluate the intersection over union (IoU) of the generated image and flood input, as detailed in Section III.

We call our method Earth intelligence engine (EIE) and associate a novel dataset in segmentation-guided image-to-image (im2im) translation with it, as detailed in Section III-A. We show that our method outperforms a physics-unconditioned baseline model in Section IV-B and discuss generalization across space, remote sensing instruments, and other climate events (reforestation and Arctic sea ice melt) in Section IV-D.

Our work makes the following contributions.
1) A novel framework to measure physical-consistency in synthetic satellite imagery.
2) The first physically-consistent and photorealistic visualization of flood risks as satellite imagery.
3) An open-source dataset with over 30k labeled high-resolution image-triplets that can be used to study im2im translation in Earth observation.

II. RELATED WORK

We generate physically-consistent visualizations of climate change-related changes in satellite imagery by formulating a semantic image synthesis task and applying deep generative vision models to solve it.

A. Generative Vision Modeling

We formulate the generation of satellite imagery as an im2im translation problem: learn a mapping from satellite image and segmentation mask to another satellite image [12]. Deep generative vision models have been most successful at solving im2im problems [12]. GANs have been successfully used in semantic image synthesis, a subproblem within im2im, to generate photorealistic street scenery from semantic segmentation masks: DCGAN [52], Pix2pixHD [13], DRPAN [13], SPADE [53], or OASIS [54]. Similarly, probabilistic normalizing flows (NFs) [29], [55], variational autoencoders (VAEs) [56], [57], autoregressive models [58], and diffusion-based models [59] (see Appendix Fig. A.4—supplementary material) have been adapted to im2im translation. Our use-case requires a deterministic semantic image synthesis model that is capable of generating realistic high-resolution (i.e., 1024 × 1024 px) images. We decided to focus on GANs, because VAEs may often generate less realistic images [57], [60]. NFs could likely capture the distribution of possible images more accurately, but our use-case of engaging visualizations is sufficiently captured with a single deterministic high-resolution image and NFs typically require specialized model architectures that can become computationally expensive. Diffusion-based models show comparable performance to GANs in generating synthetic satellite imagery [61]. However, diffusion-models are currently computationally too expensive to train at 1024 × 1024 px resolution, can still be outperformed by GANs in similar tasks [61], and feature stochasticity that is not needed for this task. So, we decided to extend the high-resolution semantic image synthesis model, pix2pixHD [13],
TABLE I
DATA OVERVIEW. WE CREATED NINE DATASETS TO STUDY IM2IM TRANSLATION AND SEGMENTATION (SEG) FOR FLOOD, REFORESTATION, AND SEA ICE MELT EVENTS. IN TOTAL, THE DATASETS CONTAIN ≈90 GB OR 32k HD IMAGE-PAIRS OR -TRIPLETS

| Data name   | Event    | Task            | Data sources | test split | # imgs | Size  | Description                                                                 |
|-------------|----------|-----------------|--------------|------------|--------|-------|-------------------------------------------------|
| xbd2xbd     | flood    | maxar-xbd -> xbd | im2im        | random     | 3284   | 9.8 GB | Main dataset with HD image-triplets from 7 floods in different regions and times. Each image-triplet contains a pre-flood and post-flood image and a generated flood mask. |
| xbd-seg     | flood    | seg maxar-xbd   | random       | 109        | 160 MB | Ground-truth flood segmentation masks for a subset of the imagery in xbd2xbd. |
| xbd-fathom  | flood    | maxar-xbd + fathom | im2im        | inference  | 108    | 448 KB | Flood masks from a hindcast of a flood model for hurricane Harvey. Aligned with pre-flood images from the xbd2xbd validation set. |
| naip2xbd    | flood    | naip -> maxar-xbd | im2im        | held-out sensor | 2063   | 6.1 GB | Data to evaluate im2im translation between two sensors with significant internal variance. Here, mapping from NAIP aerial imagery to all Maxar images in the xbd2xbd dataset. |
| naip2hou    | flood    | naip -> maxar-hou | im2im        | held-out sensor | 5602   | 16 GB  | Ground-truth flood segmentation masks for a subset of the imagery in naip2houd. |
| hou-seg     | flood    | seg maxar-hou    | random       | 260        | 370 MB | Forests before and after reforestation, including ground-truth reforestation masks. Used to evaluate generalization across climate events. |
| forest      | reforestation | im2im | sen2 -> sen2 | random     | 1272   | 5.5 GB | Held-out images of reforestation in Guatemala. Used to evaluate generalization in space. |
| forest-gtm  | reforestation | im2im | sen2 -> sen2 | held-out region | 107   | 460 MB | Winter and Summer images of coastal areas in the Arctic, including generated sea ice masks. Not used in paper, but included to encourage visualization of the future Arctic. |
| arctic      | sea ice melt | im2im | sen2 -> sen2 | random     | 19446  | 51 GB  |                                                                                     |

III. MATERIALS AND METHODS

To take in four-channel images that include physical information and to generate satellite imagery that is both photorealistic and physically-consistent.

B. Climate Change Visualization Tools

Visualizations of climate change are commonly used in policy making and community discussions on climate adaptation [4], [62]. Landscape or “street-view” visualizations are used to raise environmental awareness in the general public or policy [9], [10], because they can convey the impacts of climate change, such as rising sea levels or coastal floods, in a compelling and engaging manner [10] (see Fig. 3(b)). Most landscape visualizations, however, are limited to regional information [8]. In addition, most landscape visualizations require expensive physics-based renderings and/or high-resolution digital elevation models [8]. Alternative visualization tools of coastal floods or sea-level rise are color-coded maps, such as [5], [63], and [50]. Color-coded maps convey the flood extent on a city-wide scale, but are less engaging than a photorealistic image [4]. We are generating compelling visualizations of climate change-related events as satellite imagery to aid in policy and community discussions on climate adaptation.

A. Data Overview

As part of this study, we have created nine open-source datasets in flooding, reforestation, and sea ice melt, which are summarized in Table I and detailed in Appendix A (supplementary material). The datasets are formatted as image-triplets for creating an im2im translation model or as image-pairs for creating a segmentation (seg) model. In total, the data is approximately 90 GB or 32k sets of images.

1) Main Flood Dataset: In our main dataset, xbd2xbd, we have assembled 3284 image-triplets from seven flood events in different regions and time, as detailed in Appendix A (supplementary material). Our method requires a large collection of postevent images that depict how climate events impact a landscape. However, these postevent images are usually challenging to acquire. In the case of flooding, obtaining
postflood images that display standing water is hindered by cloud-cover, time of standing flood, satellite revisit rate, atmospheric noise, and cost of high-resolution imagery. For the xbd2xbd dataset, we have sourced it by selecting flood-related data from the xBD xview2 dataset [51]. Here, each postflood image is already paired with a preflood image, which was taken by the same satellite constellation over the same region before the flood struck. The preflood and postflood data have HD 1024 × 1024 px/imag resolution with ground-sample distance ~0.5 m/px, are RGB, and were taken by the Maxar DigitalGlobe satellites.

To add the layer of physical consistency, we associate the preflood and postflood pairs with a binary low- (~30 m/px) or high-resolution (~0.5 m/px) flood mask. Because flooding hindcasts typically do not exactly match the observed flood extent (e.g., [64] for hurricane Harvey), we derive the flood mask inputs that are used to train the im2im model from the postflood images using a separate flood segmentation model. To train this flood segmentation model, we have created the xbd-seg dataset which contains 109 pairs of postflood images and corresponding hand-labeled flood segmentation masks.

2) Auxiliary Datasets: After training and evaluating the im2im model on masks of the observed flood extent, we test if it can also visualize predictions from a physics-based flood model using the xbdfathom dataset. The xbdfathom dataset pairs the preflood images in the xbd2xbd hurricane Harvey validation set with ~30 m/px flood masks from the Fathom-US hydraulic model framework hindcast, as detailed in Appendix A–A3 (supplementary material). The datasets, naip2xbd, naip2hou, and hou-seg, are used in Section IV-D2 to test if our model could visualize flooding by using NAIP aerial imagery as input, which would be available across the full U.S. East Coast. The datasets, forest and forest-gtm, are used to study reforestation visualizations in Section IV-D3 and described in Appendixes A and B (supplementary material). Finally, we created the arctic dataset in Appendixes A–C (supplementary material), which we did not use in model training, but published to facilitate extensions of our method for visualizing sea ice melt.

B. Model Architecture

The central model of our pipeline is a generative vision model that learns the physically-conditioned im2im transformation from preflood image to postflood image. We leveraged the existing implementation of pix2pixHD [13]. Pix2pixHD is a state-of-the-art semantic image synthesis model that uses multiscale generator and discriminator architectures to generate high-resolution imagery and we refer to the original paper [13] for details on the model architectures. We extended the input dimensions to 1024 × 1024 × 4 to incorporate the focus area mask. We experimented with different parameters (see Appendix C, supplementary material), but eventually decided to use the same architecture and hyperparameters as in [13]. We trained the model from scratch on each dataset. For the xbd2xbd dataset, training took 200 epochs in ~7 h on 8 × V100 Google Cloud GPUs. The resulting pipeline is modular, such that it can be repurposed for visualizing other climate impacts.

C. Trust in Flood Images Through Physical-Consistency

We define a physically-consistent model as one that fulfills laws of physics, such as conservation of momentum, mass, and energy [33]. For example, most coastal flood models consist of numerical solvers that resolve the conservation equations to generate flood extent predictions [65]. Here, we consider a flood image to be physically-consistent if it depicts the predictions of a physically-consistent model.

Specifically, we define our generated satellite imagery, \( I_G \), \( I = [0, 1]^{w \times h \times c} \) with width, \( w = 1024 \), height, \( h = 1024 \), and number of channels, \( c = 3 \), to be physically-consistent if it depicts the same flood extent as the binary flood mask, \( F \), \( F = [0, 1]^{w \times h} \). We implemented a flood segmentation model, \( m_{seg} : I \rightarrow F \), to measure the depicted flood extent in the generated image. If the IoU of the flood extent in a generated image and the flood model is better than some threshold, \( \epsilon \), the image is considered to be in the set of physically-consistent images, i.e., \( I_{phys} = \{ I_G \in I : \text{IoU}(m_{seg}(I_G), F) > \epsilon \} \). The generated image is termed photorealistic, if it is contained in the manifold of naturally possible satellite images, \( I_{photo} \subset I \). Hence, we train a GAN to learn the conditional image generation function, \( g \), that generates an image that is both, physically-consistent and photorealistic, i.e., \( g : I_{photo} \times F \rightarrow I_{photo} \cap I_{phys} \).

IV. RESULTS

In Section IV-A, we define the evaluation metrics. Section IV-B analyzes the physical-consistency and visual quality of the generated flood imagery on the xbd2xbd dataset and Section IV-C evaluates the underlying flood segmentation model. In Section IV-D1, we query the trained im2im model with flood masks from a physics-based flood model for selected locations in Houston, TX, USA. In Section IV-D2, we evaluate which steps would be necessary to generate a flood visualization layer across the full U.S. East Coast. In Sections IV-D3 and IV-D4, we extend the model to visualize reforestation, and describe a dataset for future extensions toward visualizing Arctic sea ice melting.

A. Evaluation Metrics

Evaluating synthetic imagery is difficult [66], [67]. Most evaluation metrics measure photorealism or sample diversity [67], but not physical consistency [68] (e.g., SSIM [69], MMD [70], IS [71], MS [72], FID [73], [74], or LPIPS [75]).

To evaluate physical consistency we propose using the IoU between water in the generated postflood image and water in the ground-truth flood mask. This method relies on flood masks, but because there are no publicly available flood segmentation models for high-resolution visual satellite imagery, we trained our own model (see Section IV-C). This segmentation model created flood masks of the generated and ground-truth postflood image which allowed us to measure the overlap of water in between both (best IoU is 1, lowest is 0).

To evaluate photorealism, we used the commonly used perceptual similarity metric learned perceptual image patch similarity (LPIPS) [75]. LPIPS computes the feature vectors (of an ImageNet-pretrained AlexNet CNN architecture) of
the generated and ground-truth postflood image and returns the mean-squared error between the two feature vectors (best LPIPS is 0, worst is 1).

Because the joint optimization over two metrics poses a nontrivial hyperparameter optimization problem, we propose to combine the evaluation of physical consistency (IoU) and photorealism (LPIPS) in a new metric (FVPS), called flood visualization plausibility score (FVPS). The FVPS is the harmonic mean over the submetrics, IoU and \((1 - \text{LPIPS})\), that are both \([0, 1]-\)bounded. Due to the properties of the harmonic mean, the FVPS is 0 if any of the submetrics is 0; the best FVPS is 1. In other words, the FVPS is only 1 if the imagery is both photorealistic and physically-consistent. A small number, \(\epsilon\), is added for numerical stability

\[
\text{FVPS} = \frac{2}{\text{IoU} + \frac{\epsilon}{1 - \text{LPIPS}}}.
\]

(1)

B. Physical-Consistency and Photorealism

We train and evaluate our physics-informed GAN on the xbd2xbd dataset, as detailed in Section III. The dataset contains over 3k HD image-triplets from floods in multiple regions and times. We used all regions during training and evaluated on randomly held-out images of two regions with most visible floods. We evaluate all models on inference with high-res. (~0.5 m/px) and coarse-grained low-res. (~30 m/px) flood masks to imitate flood model predictions at different resolutions. Fig. 4 visualizes samples from our model using high-res. flood maps. We compare our model against a GAN with the same model architecture, but an ablated flood mask (baseline GAN) in Section IV-B1, a photoshopped baseline with perfect flood extent in Section IV-B2, and a VAE-based model in Section IV-B3. Overall, we find that our model is on-par with the VAE and outperforms the GAN and photoshopped baseline in terms of physical-consistency or photorealism, respectively.

1) GAN Without Physics Information Generates Photorealistic, But Non-Physically-Consistent Imagery: The baseline GAN uses the default pix2pixHD [13] with the same architecture and hyperparameters as the physics-informed GAN, but only uses the preflood image and not the flood mask as input. The baseline GAN visualized floods at locations where there was no flood according to the flood mask, for example, in Fig. 5(e)-top versus (b)-top. In practice, it would be dangerous to operationalize a visualization with false-positive flood predictions. And, the inaccurately modeled flood extent illustrates the importance of measuring physical-consistency, as defined in Section III-C. In our case, we measure physical-consistency via the IoU of the flood masks and, across the high-res. validation set (in Appendix A, supplementary material), the baseline GAN has a lower IoU (0.226) in comparison to the physics-informed GAN (0.502). Despite the photorealism of the baseline GAN (LPIPS = 0.293 versus 0.265), the physical-inconsistency renders the model nontrustworthy as confirmed by the low FVPS (0.275).

2) Handcrafted Baseline Model Generates Physically-Consistent But Not Photorealistic Imagery: Similar to common flood visualization tools [50], the handcrafted model overlays the flood mask input as a hand-picked flood brown (#998d6f) onto the preflood image, as shown in Fig. 5(g)-top. On the low-res. validation set, this model should have a perfect IoU (1) by construction, but our GAN-based flood segmentation model struggles with hard boundaries and measures an acceptable IoU (0.361). Qualitatively, the handcrafted baseline looks pixelated with flood masks at the resolution of a physical model [5] of 30 m/px and does not account for varying flood color between events (see Fig. 4). The high LPIPS (0.415) indicates quantitatively that the handcrafted visualizations are less photorealistic than the physics-informed GAN (0.283). Despite the lower photorealism, we believe that a handcrafted baseline is a reasonable alternative to generative vision models considering the benefit-cost ratio. In practice, the handcrafted baseline’s main benefit would be the perfect IoU and the GAN’s would be the interpolation of low res. flood masks onto the imagery’s resolution, which can become especially relevant in densely populated areas.

3) Proposed Physics-Informed GAN Generates Physically-Consistent and Photorealistic Imagery: To create the physics-informed GAN, we trained a pix2pixHD [13] on the xbd2xbd data, as detailed in Section III and Appendix C (supplementary material). This model successfully learned how to synthesize photorealistic postflood images from a preflood image and a flood mask, as shown in Fig. 4-left. The model improves over all other models either in terms of IoU, LPIPS, or FVPS (see Table II). The learned image transformation “in-paints” the flood mask in the correct flood colors and displays an average flood height that does not cover structures (e.g., buildings and trees), as shown in 64 randomly sampled test images in Appendix Fig. A.3 (supplementary material). Occasionally, city-scenes show scratch patterns, e.g., Appendix Fig. A.3 (top-left, supplementary material). This could be explained by the unmodeled variance in off-nadir angle, sun inclination, GPS calibration, color calibration, atmospheric noise, dynamic objects (cars), or flood impacts, which is partially addressed in our generalization experiments in Section IV-D2. We also train a VAE-based model with GAN components on the same in-/outputs than the physics-informed GAN. Specifically, we use a VAEGAN called BicycleGAN [57], which has the potential to create ensemble forecasts over the unmodeled flood impacts, such as the probability of destroyed buildings. The generated VAEGAN images look smeared Fig. 5(f)-top.
and the LPIPS (0.449) is worse than for the physics-informed GAN (0.265) on the high res. validation set.

C. Flood Segmentation Model

Our approach requires a flood segmentation model to generate ground-truth flood masks and evaluate the generated postflood images, but there does not exist any open-source model that segments floods in high-resolution (<1 m/px) satellite imagery. Hence, we labeled our own dataset of 109 HD image-pairs, xbd-seg, and trained a flood segmentation model. Our model is based on pix2pix [12] with a custom loss function and validated with fivefold cross-validation due to the small data size. The model is trained and evaluated at ~0.5 m/px and flood masks at ~30 m/px are derived via nearest neighbor downsampling of the model outputs. The segmentation model at ~0.5 m/px has a mean IoU of 0.343 which matches the expected performance of pix2pix [12] and is sufficient for our task. While one might expect an IoU closer to 1.0, many of our ground-truth flood masks have <5% of positive labels which skew the IoU toward zero if not perfectly predicted (e.g., Appendix Fig. D.8-bottom-left, supplementary material). Our segmentation model is detailed in Appendix B (supplementary material) and the xbd-seg data in Appendix A–A2 (supplementary material).

D. Generalization Performance

Our project aims to create satellite imagery that visualizes climate phenomena across the globe. So far, however, we have only evaluated our model on one climate phenomenon (floods), one remote sensing instrument (Maxar satellite imagery), a few selected locations, and observation-derived flood masks. To extend our model, we first test if the im2im model can be applied to flood masks from a physics-based flood model.

1) Inference With Physics-Based Flood Mask Inputs: We query the physics-informed GAN—trained on xbd2xbd—for the postflood imagery from hurricane Harvey and flood masks in xbd2xbd and xbdfathom. To qualitatively analyze the results, we first plot the predictions (third column) when using the low-res. flood mask inputs (second column) from the xbd2xbd validation dataset, in Appendix Fig. D.7 (supplementary material). Then, we use the same model and preflood image (first column), but use the flood masks from the Fathom Global hydraulic model (sixth column) in xbdfathom as input and plot the predictions in the fifth column.

For both datasets, the generated imagery seems to contain brown-colored floods in areas matching the flood mask inputs. Especially, the images in the second row illustrate the capacity to visually match the input flood masks in the third column.
of the im2im model to visualize a different flood extent depending on the input mask. The average IoU across all images in xbd2fathom dataset is 0.398 which is on a similar scale as the IoU on the low-res. xbd2xbd data (0.365) in Section II. Limitations of the im2im model are visible in forests occasionally being over-painted with a flood brown or houses that appear smeared and are discussed in Appendix D (supplementary material). Overall, these results indicate that the model can generalize from being trained with observation-derived flood masks to being tested with physics-simulated flood masks, assuming that the distribution of pre-flood images is held constant.

2) Generalization Across Location and Remote Sensing Instruments: In order to create a visualization layer of flood hazards across the U.S. East Coast, we would need, among others: flood mask inputs across the U.S. East Coast, pre-flood images across the U.S. East Coast, and a model that has been validated for these data streams. We also note that our work only evaluates the im2im model on a tile-by-tile basis and stitching multiple tiles into a seamless large-scale tif entails a set of challenges that are beyond the scope of this work [17], [76].

The flood mask inputs are relatively easy to access, for example, by using an NWS SLOSH National Storm Surge Hazard Map for a hurricane category X storm [5] as discussed in Appendix E (supplementary material). Because this layer is only available at 30 m/px, we tested our model’s performance on low-res. flood masks in Section IV-B.

The flood mask inputs at ≈30 m/px resolution could be accessed for research purposes, for example, from the Fathom Global flood inundation model which was run for multiple climate scenarios [77]. Alternatively, the National Weather Service publishes a 30 m/px National Storm Surge Hazard Map for hurricanes of different categories [5]. As Section IV-D1 showed that 30 m/px physics-based flood masks can be used as inputs, we consecutively evaluate if different pre-flood images can be used as inputs.

The pre-flood images in xbd2xbd are from Maxar, which is not freely available across the full U.S. East Coast. Hence, we downloaded pre-flood images from the open-access U.S.-wide mosaic of 1.0 m/px visual aerial imagery from the 2019 National Agriculture Imagery Program (NAIP) [78]. For the new dataset, naip2xbd, we pair the pre-flood NAIP images with the post-flood Maxar images and flood masks from xbd2xbd. During the pairing process, we orthorectified the Maxar images to match the NAIP images.

The im2im task from NAIP to Maxar imagery in naip2xbd is significantly more challenging than the Maxar to Maxar task in xbd2xbd, because the learned image-transformation needs to account for the sources of variance in differing remote sensing instruments. These are, for example, resolution, atmospheric noise, color calibration, inclination angle, and other factors. We first applied the physics-informed GAN from xbd2xbd without fine-tuning on the new naip2xbd data, but saw that it generated unintelligible imagery in Appendix Fig. A1-top (supplementary material). After retraining the same model from scratch on the naip2xbd data, the image quality was still relatively poor (not shown). We believe that the variance within the xbd2xbd post-flood images data was too large to learn a mapping from NAIP to Maxar imagery with only 2063 HD images in naip2xbd.

To reduce the learning task complexity, we created the naip2hou dataset, for which we sourced post-flood image tiles from a single open-access Maxar satellite pass over West Houston, TX, USA, on August 31, 2017, post-hurricane Harvey [79]. To rerun our pipeline, we labeled an additional 260 flood segmentation masks (hou-seg, taking ~20 h), retrained the flood segmentation model, and generated flood masks for the naip2hou dataset, as shown in Appendix Fig. D8 (supplementary material). The naip2hou dataset contains 5602 HD image-triplets. Then, we retrained the physics-informed GAN from scratch on the new dataset for 15 epochs in ~5 h on 1 × V100 GPU. We use a random 80–20 train-val split.

The resulting model for naip2hou has acceptable performance, illustrated in Appendix Fig. A1-top (supplementary material). It does not outperform a handcrafted baseline model in physical-consistency (not shown), but outperforms it in photorealism (LPIPS = 0.369 versus 0.465). This indicates that im2im translation across remote sensing instruments is feasible and a visualization of flood hazards along the U.S. East Coast could be realized in follow-up work.

With xbd2xbd and naip2hou, we created a dataset of a combined 8886 clean image-triplets that we are releasing as the flood-section of our open-source dataset to study segmentation-guided im2im translation in Earth observation. Furthermore, albeit small and geographically biased, the xbd2seg and hou-seg data with a combined 369 HD image-pairs are the first open-source datasets for flood segmentation on visual high-resolution (<1 m/px) satellite imagery, to the extent of the author’s knowledge, and will be made available as part of the dataset.

3) Generalization Across Different Climate Phenomena—Visualizing Reforestation: Visualizing negative climate impacts such as flooding might evoke anger, fear, or guilt in some viewers. These emotions can encourage pro-environmental behavior [80], [81], [82], but also cause inaction from feeling overwhelmed [83] and hope is needed to maintain environmental engagement [84]. Here, we extend the EIE to visualize the impact of positive hopeful actions, specifically, reforestation. This visualization has already and can be used to encourage policymakers, carbon finance investors [85], [86], or landowners to increase reforestation efforts.

To synthesize satellite imagery of reforested land, the EIE uses an image of a deforested area along with a binary mask of where trees will be planted as input, as illustrated in Appendix Fig. A2 (supplementary material). To train and evaluate the model we collected the forest dataset, which contains image-triplets of an RGB pre-reforestation satellite image, a binary reforestation area mask (1 = reforestation), and an RGB post-reforestation satellite image. We assembled a total of 1026 train and 246 test high-resolution (50 cm/px) 1024 × 1024 image-triplets via Google Earth Pro (Map data: Google, Maxar Technologies, CNS/Airbus). The dataset spans four different countries: Uruguay, Sierra Leone, Peru, and
TABLE III

|                      | LPIPS ↓ random split | LPIPS ↓ spatial split |
|----------------------|----------------------|-----------------------|
| GAN (ours)           | 0.503                | 0.574                 |
| Green mask (RGB=33,64,61) | 0.794                | 0.848                 |
| Green mask (RGB=78,116,85) | 0.845                | 0.957                 |

Mexico, as detailed in Appendixes A and B (supplementary material).

We trained the generative vision model that performed best on floods, pix2pixHD, using several augmentation techniques and the default [13] hyperparameters, as detailed in Appendixes A–C (supplementary material). We evaluated the model on a random validation split on forest and a spatial split testing on 107 held-out images, forest-gtm, from Guatemala. We evaluate LPIPS and compare the GAN to a baseline model that applies uniformly colored masks.

Fig. 5(d)-bottom shows how our model generates photorealistic visualizations of reforestation projects. The generated imagery looks more realistic than handicrafted baseline models (e) and (f), where the reforested area pixels are set to a mean forest color. Our quantitative analysis in Table III confirms that our model outperforms the baselines in both, a random and spatial split. We plot an additional random selection of generated images with their respective inputs in Appendix Fig. D.9 (supplementary material).

4) Visualizing Arctic Sea Ice Melt: The retreat of the Summer Arctic sea ice extent is one of the most important and imminent consequences of climate change [1]. However, visualizations of melting Arctic sea ice are limited to physics-based renderings, such as [87]. There also does not exist publicly-available daily high-resolution (less than 500 m) visual satellite imagery due to satellite revisit rate and cloud cover. To enable the extension of the EIE for visualizing Arctic sea ice melt, we publish the arctic dataset of ≈20k image triplets of Winter image, Summer image, and ice segmentation mask, as detailed in Appendixes A–C (supplementary material).

V. CONCLUSION

We proposed a new pipeline to create synthetic and physically-consistent satellite images of future climate events using deep generative vision models. In this section, we discuss limitations and future work.

1) Limitations: First of all, satellite imagery is currently an internationally trusted source for analysis in deforestation, development, or military domains [88], [89]. With the increased capability of data-generating models, more work is needed in the identification of and education around misinformation and ethical and trustworthy AI [21], [26]. We point out that our satellite imagery is synthetic, and should only be used as a scientific communication aid to better explain our results to decision makers or the general public [4], and we take the first steps toward guaranteeing trustworthiness in synthetic satellite imagery.

We had originally envisioned an operational coastal flood visualization across the U.S. East Coast, but discovered that the variance within and between remote sensing instruments is too large to develop an operational visualization as part of a single research paper. Our dataset is not small, as 3–5k HD image-triplets would equal 192–320k triplets at 128 × 128 px resolution. However, our dataset contains spatial and temporal biases including a bias for U.S. areas and vegetation-filled areas. The latter likely contributes to our model rendering human-built structures, such as streets and out-of-distribution skyscrapers in Appendix Fig. A.3-top-left (supplementary material), as smeared. Apart from the data limitation, smeared features are still a current concern in state-of-the-art GAN architectures [54] and generative vision models continue to struggle with the sharp lines in remote sensing imagery in Appendix Fig. A.4 (supplementary material). We overcame part of our data limitations by focusing our study on Houston, TX, USA and using several augmentation techniques, detailed in Appendix C (supplementary material), but this work would likely still benefit from more diverse postflood images from, e.g., the Maxar or NOAA data archives, or incorporating pretrained geospatial embeddings [61], [90].

2) Future Work: We envision a tool that can visualize local climate impacts and adaptation techniques at the global scale. By changing the input data, future work can visualize the impacts of other well-modeled, visible, and climate-attributable events, including Arctic sea ice melt, wildfires, or droughts. Nonbinary climate impacts, such as inundation height, or drought strength could be generated by replacing the binary flood mask with continuous model predictions. Opportunities are abundant for further work in visualizing our changing Earth. This work opens exciting possibilities in generating realistic and physically-consistent imagery with the potential impact of improving climate mitigation and adaptation.

DATA AVAILABILITY STATEMENT

The code for this study is available at github.com/blutjens/eie-earth-public. The data has been published at huggingface.co/datasets/blutjens/eie-earth-intelligence-engine.

ACKNOWLEDGMENT

The authors are very thankful to Leo Silverberg and Margaret Maynard-Reid for creating the demo at https://climate-viz.github.io. They thank Ritwik Gupta for the continuous help in using the xBD dataset, Richard Strange for the help with cloud computing, Prof. Marco Tedesco for advice on the Arctic sea ice, Guy Schumann for flood modeling, Mark Veillette and Cait Crawford for technical direction, and James Parr, Leah Lovgren, Sara Jennings, and Jodie Hughes for the organization of Frontier Development Lab (FDL), USA, and enabling these connections. They greatly appreciate the advice on decision-/policymaking in coastal climate adaptation by Derek Loftis, Sagy Cohen, Capt. John Radovan, Maya Nasr, and Janot Mendler de Suarez. Furthermore, they greatly appreciate the feedback and direction from Esther Wolff, Hannah Munguia-Flores, Peter Morales, Nicholas Mehrle, Prof. Bistra Dilkina, Freddie Kalaitzis, Graham Mackintosh, Michael van...
Pohle, Gail M. Skofronick-Jackson, Tsengdar Lee, Madhulika Guhathakurta, Julien Cornebise, Maria Molina, Massy Mascolo, Scott Penberthy, John Karecz, Jack Kaye, Mich Lin, Campbell Watson, the FDL research community, and the anonymous reviewers. They also gratefully acknowledge support from Earthshot Labs for the reforestation visualization and greatly appreciate the feedback and help from Patrick Leung. They thank Phil Cherner and the members of the Harvard Visualization Research and Teaching Laboratory for co-creating the tabletop visualization platform in Fig. 1, specifically Rus Gant, Kachina Studer, and Michael Quan. They declare no conflict of interest.

This research was supported by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Author Contributions: Conceptualization—Björn Lütjens and Brandon Leshchinskiy; methodology—Björn Lütjens, Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodríguez, Océane Boulais, and Alexander Lavin; software—Björn Lütjens and Margaux Masson-Forsythe; validation—Björn Lütjens, Aaron Piña, Yarin Gal, Chedy Raïssi, and Dava Newman; formal analysis—Björn Lütjens, Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodríguez, Océane Boulais, Alexander Lavin, and Margaux Masson-Forsythe; investigation—Björn Lütjens and Margaux Masson-Forsythe; data curation and floods—Björn Lütjens, Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodríguez, Océane Boulais, Alexander Lavin, Aruna Sankaranarayanan, and Margaux Masson-Forsythe; writing (original draft preparation)—Björn Lütjens; writing (review and editing)—Björn Lütjens, Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodríguez, Océane Boulais, Alexander Lavin, Aruna Sankaranarayanan, and Margaux Masson-Forsythe; visualization—Björn Lütjens; supervision—Björn Lütjens and Dava Newman; project administration—Björn Lütjens and Brandon Leshchinskiy; funding acquisition—Björn Lütjens, Brandon Leshchinskiy, and Dava Newman. They have read and agreed to the published version of the manuscript.

Björn Lütjens is leading the work and contributed to all sections. Brandon Leshchinskiy started the work. Brandon Leshchinskiy, Christian Requena-Mesa, Farrukh Chishtie, Natalia Díaz-Rodríguez, Océane Boulais, Ana Mata-Payerro, and Alexander Lavin contributed to the flood experiments in Sections III-C–III-C. Aruna Sankaranarayanan contributed to the ice experiment. Margaux Masson-Forsythe contributed to the reforestation experiment. Aaron Piña, Yarin Gal, Chedy Raïssi, and Dava Newman advised the work.

REFERENCES

[1] V. Masson-Delmotte et al., Eds. Global warming of 1.5°C. An IPCC Special Report on The Impacts of Global Warming of 1.5°C Above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. IPCC, 2018. [Online]. Available: https://archive.ipcc.ch/report/sr15/pdf/sr15_ipcc.pd

[2] The Human Cost of Weather-related Disasters 1995-2015, Centre for Res. Epidemiology Disasters (CRED) UN Office for Disaster Risk Reduction UNISDR, Geneva, Switzerland, 1995.

[3] NOAA Nat. Centers for Environ. Inf. (NCEI), (2023). U.S. Billion-Dollar Weather and Climate Disasters (2023). [Online]. Available: https://www.ncdc.noaa.gov/billions/

[4] S. R. Sheppard, Visualizing Climate Change A Guide to Visual Communication of Climate Change and Developing Local Solutions. New York, NY, USA: Taylor & Francis, Mar. 2012.

[5] NOAA Nat. Weather Service Nat. Hurricane Center Storm Surge Predict. Unit. (2020). National Storm Surge Hazard Maps, Texas To Maine, Category 5. [Online]. Available: https://noaa.maps.arcgis.com/apps/MapSeries/index.html?appid=d9ed7904ibe4441a9c4d7b277935fad&entry=1

[6] Climate Central. (Apr. 2023). Coastal Risk Screening Tool. [Online]. Available: https://coastal.climatecentral.org/

[7] NOAA. (Apr. 2023). Sea Level Rise Viewer. [Online]. Available: https://coast.noaa.gov/slr/

[8] B. Strauss. (2015). Sargasso Seas: Sea Level Rise Analysis. [Online]. Available: https://sealevel.climatecentral.org/

[9] V. Schmidt et al., “Visualizing the consequences of climate change using cycle-consistent adversarial networks,” in Proc. Int. Conf. Learn. Represent. (ICLR) Workshop Tackling Climate Change with AI, 2019, pp. 1–12.

[10] S. R. J. Sheppard, “Landscapes visualisation and climate change: The potential for influencing perceptions and behaviour,” Environ. Sci. Policy, vol. 9, no. 6, pp. 637–651, Dec. 2005.

[11] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 1–11.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5967–5976.

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional GANs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8798–8807. [Online]. Available: https://github.com/NVIDIA/pix2pixHD

[14] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2242–2251.

[15] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” 2018, arXiv:1809.11096.

[16] C. Requena-Mesa, M. Reichstein, M. Mahecha, B. Kraft, and J. Denzler, “Predicting landscapes from environmental conditions using generative networks,” in Proc. German Conf. Pattern Recognit. Cham, Switzerland: Springer, 2019, pp. 203–217.

[17] A. Frühstück, I. Alhashim, and P. Wonka, “TileGAN: Synthesis of large-scale non-homogeneous textures,” ACM Trans. Graph., vol. 38, no. 4, pp. 1–11, Jul. 2019, doi: 10.1145/3306346.3322993.

[18] T. Mohandas, A. Kulkarni, D. Northrup, E. Mwebaze, and H. Alemohammad, “Generating synthetic multispectral satellite imagery from sentinel-2,” in Proc. NeuroIPS 2020 Workshop AI for Earth Sci., 2020, pp. 1–10.

[19] P. Singh and N. Komodakis, “Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks,” in Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), Jul. 2018, pp. 1772–1775.

[20] N. Audebert, B. Le Saux, and S. Lefevre, “Generative adversarial networks for realistic synthesis of hyperspectral samples,” in Proc. IGARSS - IEEE Int. Geosci. Remote Sens. Symp., Jul. 2018, pp. 4359–4362.

[21] A. B. Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Inf. Fusion, vol. 38, pp. 82–115, Jun. 2020.

[22] S. Ali et al., “Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence,” Inf. Fusion, vol. 99, Nov. 2023, Art. no. 101805.
M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local Nash equilibrium,” in Proc. NIPS, 2017, pp. 6626–6637.

S. Zhou, A. Luccioni, G. Cosne, M. S. Bernstein, and Y. Bengio, “Establishing an evaluation metric to quantify climate change image realism,” Mach. Learn., Sci. Technol., vol. 1, no. 2, Dec. 2020, Art. no. 025005.

R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 586–595.

X. Li, R. Feng, X. Guan, H. Shen, and L. Zhang, “Remote sensing image mosaicking: Achievements and challenges,” IEEE Geosci. Remote Sens. Mag., vol. 7, no. 4, pp. 8–22, Dec. 2019.

O. E. J. Wing, P. D. Bates, and N. Quinn, “A 30 m global flood inundation model for any climate scenario,” ESS Open Arch., Oct. 2023. [Online]. Available: https://naip-usdaonline.hub.arcgis.com/

USDA-FSA-APFO Aerial Photography Field Office. (2019). National Geospatial Data Asset National Agriculture Imagery Program (NAIP) Imagery. [Online]. Available: http://gis.apfo.usda.gov/arcgis/services/NAIP

Maxar DigitalGlobe. (Aug. 31, 2017). Open Data Program, Hurricane Harvey. [Online]. Available: https://www.maxar.com/open-data/

E. F. Thomas, C. McGarty, and K. I. Mavor, “Transforming ‘apathy into movement’ the role of prosocial emotions in motivating action for social change,” Personality Social Psychol. Rev., vol. 13, no. 4, pp. 310–333, 2009.

P. Slovic, M. L. Finucane, E. Peters, and D. G. MacGregor, “Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality,” Risk Anal. Int. J., vol. 24, no. 2, pp. 311–322, Apr. 2004.

S. Bamberg and G. Möser, “Twenty years after hines, hungerford, and tomera: A new meta-analysis of psycho-social determinants of pro-environmental behaviour,” J. Environ. Psychol., vol. 27, no. 1, pp. 14–25, Mar. 2007.

S. O’Neill and S. Nicholson-Cole, “‘Fear won’t do it’: Promoting positive engagement with climate change through visual and iconic representations,” Sci. Commun., vol. 30, no. 3, pp. 355–379, 2009.

M. Ojala, “Hope and climate change: The importance of hope for environmental engagement among young people,” Environ. Educ. Res., vol. 18, no. 5, pp. 625–642, Oct. 2012.

G. Reiersen et al., “Reforestree: A dataset for estimating tropical forest carbon stock with deep learning and aerial imagery,” in Proc. AAAI Conf. Artif. Intell., Jun. 2022, vol. 36, no. 11, pp. 12119–12125.

B. Lütjens, L. Liebenwein, and K. Kramer, “Machine learning-based estimation of forest carbon stocks to increase transparency of forest preservation efforts,” in Proc. NeurIPS 2019 Workshop Tackling Climate Change Mach. Learn., 2019, pp. 1–3.

NASA/Goddard Space Flight Center Scientific Visualizat. Studio. (2020). Annual Arctic Sea Ice Minimum 1979-2020 With Area Graph. Accessed: Mar. 2021. [Online]. Available: https://climate.nasa.gov/climate_resources/155/video-annual-arctic-sea-ice-minimum-1979-2020-with-area-graph/

M. C. Hansen et al., “High-resolution global maps of 21st-century forest cover change,” Science, vol. 342, no. 6160, pp. 850–853, Nov. 2013.

K. Anderson, B. Ryan, W. Sonntag, A. Kavvada, and L. Friedl, “Earth observation in service of the 2030 agenda for sustainable development,” Geo-Spatial Inf. Sci., vol. 20, no. 2, pp. 77–96, Apr. 2017.

K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, “SatCLIP: Global, general-purpose location embeddings with satellite imagery,” 2023, arXiv:2311.17179.