EarthPT: a foundation model for Earth Observation

Michael J. Smith* **Luke Fleming** **James E. Geach**
Aspia Space Ltd., Cornwall, UK

Abstract

We introduce EarthPT – an Earth Observation (EO) pretrained transformer. EarthPT is a 700 million parameter decoding transformer foundation model trained in an autoregressive self-supervised manner and developed specifically with EO use-cases in mind. We demonstrate that EarthPT is an effective forecaster that can accurately predict future pixel-level surface reflectances across the 400-2300 nm range well into the future. For example, forecasts of the evolution of the Normalised Difference Vegetation Index (NDVI) have a typical error of approximately 0.05 (over a natural range of $-1 \rightarrow 1$) at the pixel level over a five month test set horizon, out-performing simple phase-folded models based on historical averaging. We also demonstrate that embeddings learnt by EarthPT hold semantically meaningful information and could be exploited for downstream tasks such as highly granular, dynamic land use classification. Excitingly, we note that the abundance of EO data provides us with – in theory – quadrillions of training tokens. Therefore, if we assume that EarthPT follows neural scaling laws akin to those derived for Large Language Models (LLMs), there is currently no data-imposed limit to scaling EarthPT and other similar 'Large Observation Models.'

1 Introduction

Deep learning's current 'hot topics' are foundation models in the vein of EleutherAI's GPT-NeoX, OpenAI's GPT-*N* models, DeepMind's Chinchilla, and the RWKV Foundation's eponymous model [1–5]. These remarkably simple models contain a few standard deep learning building blocks and are trained by repeatedly predicting the next item in a sequence. Surprisingly, these models' performances scale with dataset and model size via a simple power law [6, 7]. Even more astoundingly, at a certain scale of data and compute, these models display 'emergent abilities' such as apparent knowledge of arithmetic, law, geography, and history [e.g. 8]. In March 2022 a team at Google DeepMind discovered that – optimally – the size of these foundation models should be scaled in a roughly equal proportion to the size of the dataset used to train them [4]. Smith and Geach [9] demonstrated that this implies that the current constraint on state-of-the-art textual foundation model performance is dataset size, and not model size as previously thought. Although we are running out of useful high quality textual data to train foundation models, there remains an untapped abundance of high quality data in other domains [10, 11]. Smith and Geach [9] argue that astronomy is one such domain, and we argue here that remote sensing data sets, and in particular Earth Observation (EO) spatial and temporal data, can also be used as an additional non-textual data mode to aid in the training of ever larger, more generalist, and more performant foundation models.

Here we demonstrate that EO imaging data can be used to train a sizable transformer model in the spirit of large language modelling. To this end we train a Chinchilla-optimal 700M parameter decoding transformer model on 14B tokens of EO data in the form of multispectral time series for just over one hundred million individual pixels. The time series are analogous to word and sentence sequences in textual models, but in this case represent surface-level (solar) reflectance values

Tackling Climate Change with Machine Learning workshop at NeurIPS 2023, New Orleans.

*mike.smith@aspiaspace.com

measured in a number of passbands across the 400–2300 nm spectral range – i.e. the wavelengths corresponding to traditional 'optical' EO imagery.

Single pixel time series are commonly used in remote sensing to train transformer and self-attention based networks on supervised tasks [e.g. 12, 13]. However, currently few works apply these models in a self-supervised manner. Those that do are typically limited to very short – or even single step – time series inputs [e.g. 14, 15]. The closest approach in the literature to EarthPT is perhaps Tseng et al. [16]. They show that an encoding transformer model [i.e. 17] is capable of learning semantically meaningful embeddings from remote sensing time series. Their model is trained on a relatively small dataset comprised of 21.5M tokens arranged into chunks of shape `[time,channel]` $\equiv$ `[12,19]`. Tseng et al. [16] note their model's capability despite its small size. Our work has a diametrical and complementary purpose; we aim to demonstrate that a transformer model trained on EO data is capable of scaling to the extent that we have seen in the natural language domain, with similar potential for wide utilisation and impact. In particular, we demonstrate that EarthPT can accurately forecast reflectance values well into the future, thus providing a method to predict – and therefore an opportunity to mitigate – future events associated with environmental threats such as drought.

2 Methods

This section describes the datasets that we use to train EarthPT and the hyperparameters and training routine of our chosen decoding transformer architecture.

Training imagery. ClearSky is a proprietary deep learning algorithm that accurately predicts the equivalent of European Space Agency Sentinel-2 imagery products across 10 spectral bands: Blue, Green, Red, Red Edge 1-4, NIR, and SWIR 1 and 2. The input data for ClearSky is Sentinel-1 C-band Synthetic Aperture Radar (SAR) imagery at 10 m/pixel resolution [18]. SAR imagery is impervious to cloud cover, and the ClearSky algorithm allows us to construct full multispectral imagery time series of Sentinel-2 equivalent reflectances uninterrupted by cloud. In this work we generate ClearSky inferred imagery for an area of interest in the UK defined by a 100×100 km region corresponding to the TL square of the British National Grid (BNG) reference system. We define training and validation set time series range from January 2015 to December 2022, and test set time series ranges from January 2023 to May 2023. The time series are sampled at the same cadence as the observing pattern of Sentinel-1, which for this location is five days on average.

Preprocessing. We recompose the observation arrays into a set of `float16` `NumPy` [19] arrays of shape `[index,time,channel]`, where `index` corresponds to the flattened spatial index of a $10 \times 10\,\mathrm{km}^2$ BNG tile, `time` corresponds to the date of observation, and `channel` corresponds to the individual spectral bands and the date embedding bands of the current and next observation. The date embedding is calculated via the equation $\hat{t} = (\sin(2\pi t/365), \cos(2\pi t/365))$, where t is the date of the observation in days since 1st January of the year of observation. The spectral band reflectances (originally on a 0–10,000 scale) are normalised as $\hat{v} = v/500 - 1$, which keeps them approximately in the range $[-1, 1]$. We treat each temporal observation as a separate 'token', and therefore the TL training set (a subset of the full UK data set) comprises approximately 100B tokens. Once constructed, we can efficiently access these data structures at train time via memory-mapping.

Transformer architecture. EarthPT is based on the autoregressive transformer architecture described in Radford et al. [20], with some alterations to accommodate our non-textual dataset. In place of the usual word embedding routine we use a multilayer perceptron to embed the input data so that it has the same dimensionality as the time embedding vector. To provide the model with a knowledge of the time of observation, we feed the network an additional pair of float embeddings corresponding to the date of the current and next observation. We train EarthPT in the usual autoregressive way, by repeatedly predicting the next observation in a given set of time series. We train using the Adam optimiser [21], and use the Huber loss [22]. We trained a range of model sizes from 10M to 700M trainable parameters, and we present the hyperparameters for all our models in Appendix B. The remainder of this paper focuses on our largest EarthPT model, EarthPT-700M.

In lieu of a domain-specific neural scaling law we use the Chinchilla neural scaling law as a convenient rule-of-thumb to decide our dataset size. This law suggests that a compute-optimal decoding transformer model should be trained roughly following the scaling $N \sim 20D$, where N is the

number of parameters in the model, and D is the number of tokens in the training set [4]. This corresponds to 14B tokens for our 700M parameter model. To this end we train EarthPT-700M on 8 A100 PCIe 40GB GPUs for a cumulative total of 90,000 steps of 160,000 tokens each, i.e. 560 A100-hours of computation time.

3 Results

We find that our EarthPT models share similar training behaviour with traditional LLMs; further details of training runs can be found in Appendix B. In this section we describe how EarthPT-700M performs on the task of forecasting remote sensing data.

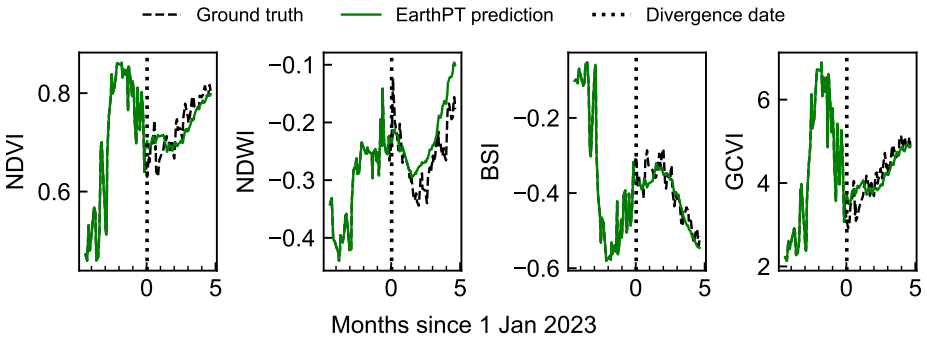

Figure 1: Predictions of some common remote sensing indicators for a randomly chosen pixel within the UK National Grid TL tile. We condition EarthPT on ClearSky time series from 1st January 2015 to 1st January 2023, with outputs after this divergence date constituting a long-term forecast to be compared to the unseen observations.

Analogously to how autoregressive language models can be used to generate text, we can use EarthPT to generate (i.e. forecast) future remote sensing data, in this case the pixel-wise surface reflectance across the optical-infrared. In Figure 1 we show EarthPT forecasts for four representative remote sensing indices: Normalised Difference Vegetation Index (NDVI), Normalised Difference Water Index (NDWI), Bare Soil Index (BSI), and Green Chloropyll Vegetation Index (GCVI). These represent time streams of a single pixel selected from the TL tile. Forecasting starts on the 1st of January 2023 and runs to May 2023. We compare the forecast to the measured values of these indices across this interval, which the model has not 'seen'. For brevity, we show a single pixel here, forecasting can be scaled across all pixels to generate predicted imagery products.

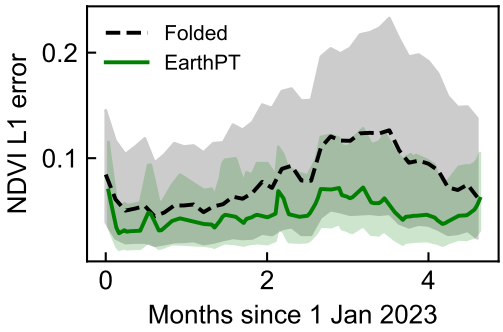

Figure 2: Median L1 error and interquartile ranges of NDVI predictions for 1M pixels in the TL63 tile. EarthPT long-term forecasts out-perform a simple phase-folded model based on historical averages out to a horizon of five months.

We can quantify performance by assessing the residual between the forecasted and measured value of the parameter of interest (e.g. NDVI) as a function of look-ahead time. Figure 2 shows the median L1 error for $\sim 10^6$ pixels in BNG tile TL63, up to five months into the future. This is compared to a prediction based on a phase-folded model which comprises an average annual time series constructed from 7 years of historical data. We find that EarthPT has a median L1 error across all time of 0.05 and the folded model has a median L1 error of 0.08, noting that NDVI has a natural range of $-1 \rightarrow 1$. We can conclude that EarthPT out-performs a phase-folded model consistently over the forecast window, delivering actionable predictions on key remote sensing indices (such as NDVI) that could be used, for example, in the prediction of drought conditions well in advance [23].

4 Future Work

Foundation models are notoriously flexible, and so one can envision myriad downstream tasks. In the field of geospatial data analysis, we can consider how EarthPT could be deployed for land cover classification. To illustrate, we generate representation embeddings by extracting the outputs of the penultimate neuronal layer and obtain the embedding of a pixel's time series by simply taking the mean of all of its output embeddings (one embedding is output at each time step). Each embedding has a dimensionality of 1280, but we can visualise them by projecting onto a two-dimensional manifold. We use principle component analysis (PCA) as our projection technique [24]. Figure 3 shows a selection of emergent remote sensing indices (introduced above) for a set of embeddings of time series across 2022. By colour-coding the projected embedding space we see that it has a physically meaningful organisation, with coherent structure of, for example, the time-averaged NDVI, BSI, RGB, etc. If we were to cluster and calibrate the embedding space with semantically meaningful labels (e.g. crop type, growth stage, event) this could be used to create a dynamic and highly granular land cover classifier. Furthermore, we anticipate that fine-tuning with EarthPT-learnt embeddings will be beneficial for a range of downstream tasks [see for example 17, 25]. One could imagine training EarthPT to produce a single embedding space for all EO (and other) multi-modal data types [26]. This would be remarkably powerful tool for interpreting remote sensing data, where we foresee diverse applications in a range of sectors, from agriculture to insurance and beyond.

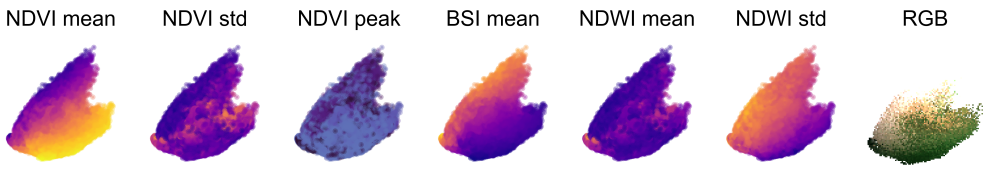

Figure 3: EarthPT embeddings for the two million pixel time series located on the TL63 and TL64 BNG tiles. We colour each scatter plot with a different set of emergent remote sensing index values. 'RGB' is the colour of a pixel in that part of the embedding space at the height of the summer of 2022. 'Mean' is the mean of a given index across the 2022 calendar year, and 'std' is the standard deviation of the index across the year. 'NDVI peak' is the time of the year corresponding to maximum NDVI; darker values are in the winter, and lighter values are in the summer. Note the coherent structure in the projected embedding space.

While useful as a rule-of-thumb, the Chinchilla scaling laws may not be suitable for EO datasets, and so follow-up work will derive a specific scaling law for our ClearSky dataset. This in turn will give us a solid theoretical grounding for further scaling of EarthPT, allowing us to train a significantly larger model. For example, with our ClearSky model for the UK we have access to 4.3T (trillion) tokens that could be used to train EarthPT, and when considering larger geographic coverage we theoretically have access to over a quadrillion tokens. Compute cost aside, we could safely train a 50T parameter model on this data, assuming that our model scaling roughly follows the Chinchilla scaling law. This 50T parameter model would be around three orders of magnitude larger than the current largest optimally-trained models [4, 27]. Consequently, unlike traditional LLMs, EarthPT and other similar 'Large Observation Models' are far from their theoretical data limit [9, 28].

5 Conclusions

Inspired by the recent explosion of interest in LLMs, we present an Earth Observation foundation model trained on time series taken from our ClearSky generative algorithm. Our EarthPT Large Observation Model is capable of forecasting surface level optical reflectance (and therefore a wide range of common remote sensing indices) at the pixel level, months into the future. EarthPT can also produce semantically meaningful embeddings for an input time series, and we show that these capture useful information that could be exploited for land cover classification, amongst other downstream tasks. We are developing these applications and improving and extending EarthPT as part of ongoing R&D. Excitingly, the number of tokens available for training is of order 10^{15}, so we are not currently data constrained. If neural scaling laws hold, then improving EarthPT (and similar Large Observation Models) is a solved problem: it is a simple matter of scaling data and compute.

Acknowledgements

This project is part-funded by the UK Government through the UK Shared Prosperity Fund. Cornwall Council has been chosen by Government as a Lead Authority for the fund and is responsible for monitoring the progress of projects funded through the UK Shared Prosperity Fund in Cornwall and the Isles of Scilly.

Data and code availability

Please contact Aspia Space directly for data and model access at contact@aspiaspace.com.

References

[1] S. Black et al. "GPT-NeoX-20B: An Open-Source Autoregressive Language Model". In: *arXiv* (2022). DOI: 10.48550/arXiv.2204.06745. eprint: 2204.06745.

[2] T. Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[3] OpenAI. "GPT-4 Technical Report". In: *OpenAI Whitepaper* (2023). URL: https://openai.com/research/gpt-4.

[4] J. Hoffmann et al. "Training Compute-Optimal Large Language Models". In: *arXiv* (2022). DOI: 10.48550/arXiv.2203.15556. eprint: 2203.15556.

[5] B. Peng et al. "RWKV: Reinventing RNNs for the Transformer Era". In: *arXiv* (2023). DOI: 10.48550/arXiv.2305.13048. eprint: 2305.13048.

[6] C. Cortes et al. "Learning Curves: Asymptotic Values and Rate of Convergence". In: *Advances in Neural Information Processing Systems* 6 (1993). URL: https://proceedings.neurips.cc/paper/1993/hash/1aa48fc4880bb0c9b8a3bf979d3b917e-Abstract.html.

[7] J. Kaplan et al. "Scaling Laws for Neural Language Models". In: *arXiv* (2020). DOI: 10.48550/arXiv.2001.08361. eprint: 2001.08361.

[8] J. Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *arXiv* (2022). DOI: 10.48550/arXiv.2201.11903. eprint: 2201.11903.

[9] M. J. Smith and J. E. Geach. "Astronomia ex machina: a history, primer and outlook on neural networks in astronomy". In: *R. Soc. Open Sci.* 10.5 (2023), p. 221454. ISSN: 2054-5703. DOI: 10.1098/rsos.221454.

[10] R. Friel. *Chinchilla's Wild Implications*. 2022. URL: https://www.alignmentforum.org/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications.

[11] P. Villalobos et al. "Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning". In: *arXiv* (2022). DOI: 10.48550/arXiv.2211.04325. eprint: 2211.04325.

[12] V. S. F. Garnot et al. "Satellite Image Time Series Classification With Pixel-Set Encoders and Temporal Self-Attention". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 13–19. DOI: 10.1109/CVPR42600.2020.01234.

[13] M. Russwurm and M. Körner. "Self-attention for raw optical Satellite Time Series Classification". In: *ISPRS J. Photogramm. Remote Sens.* 169 (2020), pp. 421–435. ISSN: 0924-2716. DOI: 10.1016/j.isprsjprs.2020.06.006.

[14] Y. Cong et al. "SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery". In: *arXiv* (2022). DOI: 10.48550/arXiv.2207.08051. eprint: 2207.08051.

[15] C. J. Reed et al. "Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning". In: *arXiv* (2022). DOI: 10.48550/arXiv.2212.14532. eprint: 2212.14532.

[16] G. Tseng et al. "Lightweight, Pre-trained Transformers for Remote Sensing Timeseries". In: *arXiv* (2023). DOI: 10.48550/arXiv.2304.14065. eprint: 2304.14065.

[17] J. Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

[18] P. S. Agram et al. "An Efficient Global Scale Sentinel-1 Radar Backscatter and Interferometric Processing System". In: *Remote Sens.* 14.15 (2022), p. 3524. ISSN: 2072-4292. DOI: 10.3390/rs14153524.

[19] C. R. Harris et al. "Array programming with NumPy". In: *Nature* 585 (2020), pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2.

[20] A. Radford et al. "Language Models are Unsupervised Multitask Learners". In: *OpenAI Whitepaper* (2019). URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[21] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015. URL: http://arxiv.org/abs/1412.6980.

[22] P. J. Huber. "Robust Estimation of a Location Parameter". In: *Ann. Math. Stat.* 35.1 (1964), pp. 73–101. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703732.

[23] E. E. Salakpi et al. "Forecasting vegetation condition with a Bayesian auto-regressive distributed lags (BARDL) model". In: *Nat. Hazards Earth Syst. Sci.* 22.8 (2022), pp. 2703–2723. ISSN: 1561-8633. DOI: 10.5194/nhess-22-2703-2022.

[24] K. Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. ISSN: 1941-5982. DOI: 10.1080/14786440109462720.

[25] P. Graff et al. "ADS". In: *Monthly Notices of the Royal Astronomical Society, Volume 441, Issue 2, p.1741-1759* 441.2 (2014), p. 1741. ISSN: 0035-8711. DOI: 10.1093/mnras/stu642.

[26] S. Reed et al. "A Generalist Agent". In: *Transactions on Machine Learning Research* (2022). ISSN: 2835-8856. URL: https://openreview.net/forum?id=1ikK0kHjvj.

[27] H. Touvron et al. "Llama 2: Open Foundation and Fine-Tuned Chat Models". In: *arXiv* (2023). DOI: 10.48550/arXiv.2307.09288. eprint: 2307.09288.

[28] H. W. Leung and J. Bovy. "Towards an astronomical foundation model for stars with a Transformer-based model". In: *arXiv* (2023). DOI: 10.48550/arXiv.2308.10944. eprint: 2308.10944.

[29] A. Lacoste et al. "Quantifying the Carbon Emissions of Machine Learning". In: *arXiv* (2019). DOI: 10.48550/arXiv.1910.09700. eprint: 1910.09700.

A Carbon emissions

The training of deep learning models requires considerable energy, contributing to carbon emissions. We trained EarthPT-700M on 8 A100 PCIe 40GB GPUs. A cumulative total of 560 A100-hours of computation was performed, corresponding to a GPU energy expenditure of $145\,\mathrm{kWh}$. Assuming a carbon efficiency of $0.193\,\mathrm{kg\,CO_2eq./kWh}$ for the UK, we estimate our emissions as $\sim 30\,\mathrm{kg\,CO_2eq.}$ We conducted these estimations via the excellent machine learning impact calculator presented in Lacoste et al. [29].

B Hyperparameters and scaling tests

Hyperparameters used to train all our EarthPT models are shown in Table 1, and our training run loss curves for all our model sizes are shown in Figure 4. We can see in Figure 4 that the larger models are still learning at the end of training, and so we expect that training larger models on more data would improve performance.

Table 1: Hyperparameters used to train our EarthPT models. Following Hoffmann et al. [4], we decay the learning rate by a factor of 10 over a horizon of a length $1.1\times$ the total training steps.

	EarthPT model size			
Hyperparameter	700M	300M	100M	10M
Number of layers	36	26	20	10
Number of heads	20	16	10	10
Embedding dimension	1280	1024	640	320
Block size	256	256	256	256
Batch size	0.164M	0.164M	0.164M	0.164M
Total training steps	90,000	90,000	90,000	90,000
Max learning rate	2E-5	2E-5	2E-5	2E-5

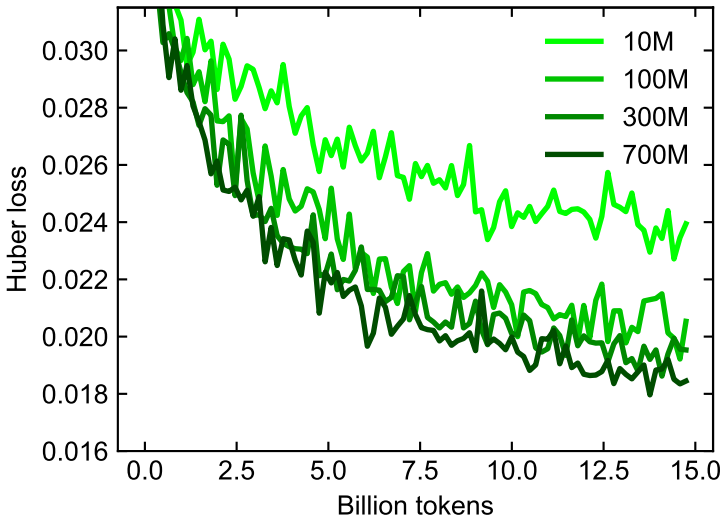

Figure 4: Loss curves for our various EarthPT training runs.