

Task-Informed Meta-Learning for Remote Sensing

Gabriel Tseng
McGill University
Mila – Quebec AI Institute

Hannah Kerner
Arizona State University

David Rolnick
McGill University
Mila – Quebec AI Institute

Abstract

Labels in remote sensing datasets - and particularly in agricultural remote sensing datasets - can be extremely spatially imbalanced, with plentiful labels in some regions but a sparsity of labels in other regions. When developing algorithms for data-sparse regions, a natural approach is to use transfer learning from data-rich regions. While standard transfer learning approaches typically leverage only direct inputs and outputs, remote sensing data (and geospatial data more generally) are rich in metadata that can inform transfer learning algorithms, such as the spatial coordinates of data-points. We build on previous work exploring the use of meta-learning for remote sensing contexts in data-sparse regions and introduce task-informed meta-learning (TIML), an augmentation to model-agnostic meta-learning which takes advantage of task-specific metadata. We apply TIML to regression and classification tasks in remote sensing for agriculture, and find that TIML outperforms a range of benchmarks in both contexts, across a diversity of model architectures. TIML was developed for remote sensing with the goal of improving the global accuracy (and equity) of machine learning models. However, it can offer benefits to any meta-learning setup with task-specific metadata – we demonstrate this by applying TIML to the Omniglot dataset.

1. Introduction

Machine learning is useful for inferring comprehensive geospatial information from sparsely labelled remote sensing data. This is applicable to a wide range of uses, from vegetation height mapping [25] to building footprint detection [58]. In particular, learning from geospatial data is crucial to better understanding, mitigating, and responding to climate change, with applications ranging from hurricane forecasting [9] to methane detection [23] to agriculture [8].

Labels for remote sensing tasks are plentifully available in certain parts of the world (for certain tasks), but many regions are extremely data-sparse (with this data imbalance

reflecting a eurocentric and amero-centric bias as in other labeled datasets in machine learning [40]). While previous work has investigated transfer learning from data-rich areas to improve performance in data-sparse areas [38, 53], remote sensing datasets are rich in metadata that can inform transfer learning algorithms by enabling models to learn useful context between datapoints, such as the relative geographic locations of datapoints or the higher-level category of the class label [49].

We propose a new method for passing such auxiliary information to the model to improve overall performance and equitable generalization. Specifically, we build on previous work investigating the adaptation of Model-Agnostic Meta-Learning [13], in particular focussing on its utility in a remote sensing for agriculture context (where tasks are created by partitioning samples based on agro-ecological [38] or political [47, 48] boundaries).

We summarize the main contributions of this paper below:

- We introduce Task-Informed Meta-Learning (TIML), an algorithm designed to augment MAML by incorporating task metadata and removing memorized tasks.
- We show that TIML improves performance for both regression (yield estimation) and classification (crop type classification & digit classification) tasks across a diversity of neural network architectures and domains.
- We highlight TIML’s ability to learn from very few positive labels and to perform well on tasks where other transfer-learned models do poorly.

We focus on areas of remote sensing where there is potential for significant social impact; we use remote sensing data sets created specifically to inform agricultural policy [48] for which traditional methods have significantly underperformed, since data are sparse and imbalanced across crop types and geographies. However, we highlight that this method can be used beyond remote sensing; TIML is applicable to any meta-learning problem that includes task-specific metadata. We demonstrate this by applying TIML to image classification with the Omniglot dataset.

All code used to train and evaluate TIML is available at

2. Related Work

Meta-Learning Meta-learning, or *learning to learn*, consists of learning to solve a task after having seen other example tasks [45]. Recent work in this area has focused on few-shot learning, i.e., learning the new task with few training datapoints. Ravi and Larochelle [37] learn an optimizer from a set of tasks which can then be applied in a new task, while Snell et al. [42] cluster data samples in the embedding space to perform few-shot classification. We leverage model-agnostic meta-learning (MAML) [13], a few-shot meta-learning framework that uses example tasks to learn a set of initial weights that can rapidly generalize to a new task.

Task-adaptive meta-learning Task-adaptive meta-learning aims to tailor meta-models to the specific tasks they will be fine-tuned on. In particular, numerous methods adapt the model weights learned during gradient-based meta-learning using task information inferred from the available training task samples by modulating the model parameters [27, 31, 39, 46, 51, 56], varying the task learning rates [27], or adapting the optimizer [5, 41] or loss function [7]. A critical component of these methods is some representation of task i , which we denote as t_i , to inform the task adaptation. Previous approaches have attempted to infer t_i from the available training samples in a task. However, such approaches typically require many datapoints to infer task similarity and also assume that tasks are readily identifiable from training samples, which is not always true. We investigate the utility of *explicit* task metadata to represent t_i , instead of *implicitly* inferring t_i from the data. Unlike approaches that learn task information implicitly, our approach enables few-shot or even zero-shot learning, in which no labelled training data is available (Appendix B).

Learning from metadata There have been several recent studies on how metadata—or auxiliary data—can inform machine learning algorithms. A common approach is to use an additional data source to learn initial weights that are useful for a range of downstream tasks [4, 19, 55]. Alternatively, metadata may be integrated into the learning process, typically by updating the final predictions of the model to align them with information provided by the metadata [22, 28, 57]. We investigate the usefulness of metadata in few-shot learning and in particular to inform gradient-based meta-learning.

Few-shot learning in geospatial contexts Few-shot learning has been extensively explored in geospatial ma-

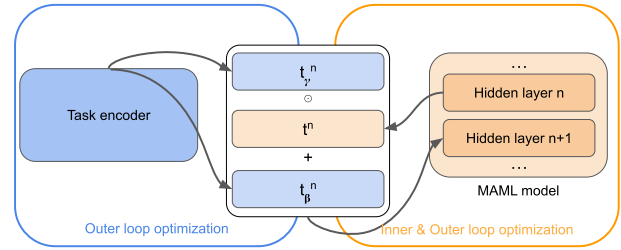


Figure 1. An illustration of the encoder, and the modulation of the MAML learner’s hidden vectors using the encoder’s output. We highlight the differing optimization regimes for the encoder and the MAML learner – the encoder’s output remains static through the MAML learner’s inner loop optimization.

chine learning, particularly by attempting to transfer knowledge from data-rich to data-sparse regions. This has been achieved in a variety of ways, ranging from transfer learning [18, 53] to multi-task learning [10, 20] to meta-learning (including MAML) [38, 47, 48]. We consider how models may be adapted to an unseen target task, which can prevent performance degradation for unseen tasks observed by prior studies in this domain [38]. In addition, there is plentiful metadata available in geospatial machine learning tasks [55]; we investigate how such metadata can be used to improve performance.

3. Task-Informed Meta-Learning

Model-Agnostic Meta-Learning (MAML) learns model weights θ that are close to optimal for each of a variety of different tasks, allowing the optimal weights for a specific task to be reached with little data and/or few gradient steps. These weights θ are updated by fine-tuning them on a training task (inner loop training), yielding updated weights θ' . A gradient for θ is then computed with respect to the loss of the updated model, $L_{\theta'}$ which is used to update θ (outer loop training).

Our approach, Task-Informed Meta-Learning (TIML) (Algorithm 1), builds on MAML by leveraging explicit task-level metadata. We introduce a task encoder to modulate the weights of the meta-learner. We also introduce *forgetfulness*, a technique to ensure that already memorized tasks do not impede learning.

Task encoder TIML modulates parameters in the meta-learner based on embeddings calculated using task information. We encode the task-specific information into a set of vectors, two for each hidden layer to be modulated in the meta-model, denoted t_γ^k and t_β^k for the k th layer. We use

Algorithm 1 Task-Informed Meta-Learning

- 1: **Require:** $p(\mathcal{T})$: Distribution over tasks
 - 2: **Require:** α, β : step size hyperparameters
 - 3: randomly initialize meta model parameters θ_m , task encoder parameters θ_e
 - 4: **while** not done **do**
 - 5: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$ with task information t_i
 - 6: **for all** \mathcal{T}_i, t_i **do**
 - 7: Generate task embeddings $\mu_i = f(t_i; \theta_e)$
 - 8: Evaluate $\nabla_{\theta_m} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_m}, \mu_i)$
 - 9: Compute adapted meta parameters with gradient descent: $\theta'_{m_i} \leftarrow \theta_m - \alpha \nabla_{\theta_m} \mathcal{L}_{\mathcal{T}_i}(f_{\theta_m}, \mu_i)$
 - 10: **end for**
 - 11: Update $\theta_m \leftarrow \theta_m - \beta \nabla_{\theta_m} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{m_i}}, \mu_i)$
 - 12: Update $\theta_e \leftarrow \theta_e - \beta \nabla_{\theta_e} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_{m_i}}, \mu_i)$
 - 13: **end while**
-

feature-wise linear modulation (FiLM [35]) to modulate the hidden vector outputs of the meta-model using these task embeddings. That is, given a hidden vector output \mathbf{h} , we compute the modulated hidden vector output using a linear transformation and pass the modulated output to the next layer in the network: $\mathbf{h}_{out}^k = (\mathbf{t}_{\gamma}^k \odot \mathbf{h}^k) + \mathbf{t}_{\beta}^k$, where \odot is the Hadamard product. The task embeddings are updated in the outer loop during MAML training, so they remain constant for all datapoints when the meta-learner is being fine-tuned for a specific task (inner loop training). We illustrate this in Figure 1.

We use a task encoder to learn the embeddings. This encoder consists of linear layers with GeLU activation [16], group normalization [54] and dropout [44]. The task information is encoded into a hidden task vector. Independent linear layers are then used to generate an embedding for each hidden vector in the classifier to be modulated. We keep the task encoder hyperparameters constant for all experiments (details in Appendix A), which demonstrates the insensitivity of this method to hyperparameter settings.

Forgetfulness We find that when the distribution of tasks is imbalanced, the model is likely to memorize over-represented tasks to the detriment of its ability to learn more difficult or rarer tasks [11]. Unlike more complex methods designed to deal with this problem [6, 11, 17, 27], explicit task information allows us to introduce a simple method to prevent memorization of certain tasks: removing training tasks the model has memorized. We define memorization as having exceeded a performance threshold for a task over a continuous set of epochs. We call this method “forgetfulness.” Reducing the training set size before training has been previously explored [15, 30] to reduce training

time; forgetfulness does this dynamically to improve performance.

4. Datasets & Experimental Setup

We evaluated TIML on a range of data sets with heterogeneous tasks and limited available data but potentially useful task metadata. Specifically, we consider crop type prediction (a classification task) and crop yield estimation (a regression task) to demonstrate the suitability of TIML in both contexts. While TIML was developed with remote sensing in mind, it can be applied more generally; we demonstrate this by evaluating the performance of TIML on the Omniglot data set.

4.1. Crop Type Classification

Up-to-date crop maps are critical to understanding the agricultural impacts of weather events and climate change [43]. Crop type classification – used to produce these maps – involves predicting whether or not a given instance contains a crop of interest. Specifically, given a remote sensing-derived pixel time series for a specific latitude and longitude and a crop of interest, the goal is to output a binary value describing whether the crop of interest is being grown at that location.

We use the CropHarvest data set [48]. This data set consists of 90,480 globally distributed datapoints, of which 30,899 (34.2%) have multi-class agricultural labels and the remainder have binary “crop” or “non-crop” labels. Each datapoint is accompanied by a remotely sensed pixel time series from: Sentinel-2 L1C optical and Sentinel 1 synthetic aperture radar satellite observations, ERA5 climatology data (precipitation and temperature), and slope and elevation from a Digital Elevation Model. The time series includes one year of data at monthly timesteps.

As in the CropHarvest benchmarks, we constructed meta-learning tasks spatially using bounding boxes for countries drawn by Natural Earth [33]. Tasks consist of binary classification of pixels as either crop vs. non-crop or a specific crop type vs. rest. This yielded 525 tasks, which were randomly split into training and validation tasks. We withheld the three CropHarvest evaluation tasks (described in Section 4.1) from the initial training. For each evaluation task, we fine-tuned the model on that task’s training data before evaluating on that task’s test data.

Task Metadata Task metadata is encoded in a 13-dimensional vector. Three dimensions encode spatial information, consisting of spherical latitude and longitude coordinates transformed to Cartesian coordinates (thus ensuring transformed values at the extreme longitudes are close to each other) using $[\cos(\text{lat}) \times \cos(\text{lon}), \cos(\text{lat}) \times \sin(\text{lon}), \sin(\text{lat})]$. The remaining 10 dimensions communicate the type of task the model is being asked to learn.

	Model	Kenya	Brazil	Togo	Mean
AUC ROC	Random Forest	0.578 ± 0.006	0.941 ± 0.004	0.892 ± 0.001	0.803
	No pre-training	0.329 ± 0.011	0.898 ± 0.010	0.861 ± 0.002	0.700
	Crop pre-training	0.694 ± 0.001	0.820 ± 0.002	0.894 ± 0.000	0.801
	MAML	0.729 ± 0.001	0.831 ± 0.005	0.878 ± 0.001	0.813
	MMAML [51]	0.690 ± 0.023	0.854 ± 0.037	0.878 ± 0.005	0.807
	TIML	0.794 ± 0.003	0.988 ± 0.001	0.890 ± 0.000	0.890
	no forgetfulness	0.779 ± 0.003	0.877 ± 0.003	0.893 ± 0.001	0.850
	no encoder	0.712 ± 0.001	0.977 ± 0.002	0.895 ± 0.000	0.862
	no task info or encoder	0.690 ± 0.001	0.977 ± 0.002	0.876 ± 0.001	0.848
	F1 score	Random Forest	0.559 ± 0.003	0.000 ± 0.000	0.756 ± 0.002
No pre-training		0.782 ± 0.000	0.764 ± 0.012	0.720 ± 0.005	0.734
Crop pre-training		0.819 ± 0.001	0.619 ± 0.005	0.713 ± 0.002	0.613
MAML		0.828 ± 0.001	0.496 ± 0.001	0.662 ± 0.001	0.652
MMAML [51]		0.794 ± 0.006	0.720 ± 0.044	0.733 ± 0.007	0.749
TIML		0.838 ± 0.000	0.835 ± 0.012	0.732 ± 0.002	0.802
no forgetfulness		0.840 ± 0.000	0.537 ± 0.002	0.764 ± 0.002	0.724
no encoder		0.840 ± 0.000	0.473 ± 0.002	0.691 ± 0.001	0.691
no task info or encoder		0.837 ± 0.001	0.473 ± 0.001	0.645 ± 0.002	0.652

Table 1. Results for the **crop type classification** evaluation tasks. All results are averaged from 10 runs and reported with the accompanying standard error. We report the area under the receiver operating characteristic curve (AUC ROC) and the F1 score using a threshold of 0.5 to classify a prediction as the positive or negative class. We highlight the **first** and **second** best metrics for each task. TIML achieves the highest F1 score of any model on the Brazil task and the best AUC ROC and F1 scores when averaged across the 3 tasks. We highlight the improvement of TIML relative to other transfer/meta-learning methods, showing its ability to leverage task metadata when learning.

This consists of a one-hot encoding of crop categories from the UN Food and Agriculture Organization (FAO) indicative crop classification [1], with an added class for non-crop. For crop vs. non-crop tasks, positive examples are given the value $\frac{1}{n}$ across all the $n = 9$ crop type categories.

4.1.1. Evaluation

The CropHarvest data set includes 3 evaluation tasks that test the ability of a pre-trained model to learn from few in-distribution datapoints in a variety of agroecologies:

Togo crop vs. non-crop: The goal of this task is to classify datapoints as crop or non-crop in Togo. The training set consists of 1,319 datapoints and the test set consists of 306 datapoints – 106 (35%) positive and 200 (65%) negative – sampled from random locations within the country.

The two other evaluation tasks consist of classifying a specific crop. Thus, “rest” below includes all other crop and non-crop classes. For both tasks, entire polygons delineating a field (as opposed to single pixels within a field) were collected, allowing evaluation across the polygons. However, during training, only the polygon centroids were used.

Kenya maize vs. rest: The training set consists of 1,345 (266 positive and 1,079 negative) samples. The test set consists of 45 polygons with 575 (64%) positive and 323 (36%) negative pixels.

Brazil coffee vs. rest: The training set consists of 794 (21 positive and 773 negative) samples. The test set consists of 66 polygons with 174,026 (25%) positive and 508,533 (75%) negative pixels.

4.1.2. Experiments

We evaluated TIML by training it on the CropHarvest data set and fine-tuning it on the evaluation tasks, as was done for the benchmark results released with the data set in [48]. TIML can be applied to any neural network architecture. We use the same base classifier and hyperparameters as in [48]: an LSTM model followed by a linear classifier.

Ablations We performed 3 ablations to quantify the contribution from the different components of TIML:

- **No forgetfulness:** TIML trained without forgetfulness; no tasks are removed in the training loop.
- **No encoder:** TIML with no encoder. The task information is instead appended to every raw input timestep and passed directly to the classifier.
- **No task information or encoder:** No task information passed to the model at all. This model is effectively a normal MAML model, trained with forgetfulness.

Baselines We compared TIML to 5 baselines. As with TIML, we fine-tuned these models on each benchmark task’s training data and then evaluated them on the task’s test data:

- **MMAML** [51]: Multimodal Modal-Agnostic Meta-Learning, which infers task-clusters from the fine-tuning data and uses this to condition the MAML model.
- **MAML**: A Model-Agnostic Meta-Learning classifier without the task information.
- **Crop pre-training**: A classifier pre-trained to classify all data as crop or non-crop (without task metadata), then fine-tuned on the test task training data.
- **No pre-training**: A randomly initialized classifier, which is not pre-trained on the global CropHarvest data set but instead is trained directly on the test task training data.

In addition, we trained a **Random Forest** baseline implemented using scikit-learn [34] (further implementation details are available in Appendix C).

4.2. Yield Estimation

Accurate and timely yield estimates are a key input to food security forecasts [8] and are necessary to better understand how food production can be sustainably managed [26]. This is especially critical as climate-related hazards (heat and drought) affect 75% of global harvested area [21] and have reduced global average yields of maize, soybeans and wheat by 11.6%, 12.4% and 9.2%, respectively [29, 36]. Yield estimation is a regression task which consists of predicting the amount of crop harvested per unit of land in a given area, given remote sensing data of that area. We estimate soybean yield in the highest soybean-producing states in the United States.

We recreated the yield prediction dataset originally collected by [57]. This dataset consists of county-level soybean yields for the 11 U.S. states accounting for over 75% of national soybean production from 2009 to 2015 (shared under the U.S. Public Domain), and remote sensing data (specifically MODIS [50, 52] products, which are shared through the LP DAAC¹). Since counties cover large areas, inputting the raw satellite data to the model would create extremely high-dimensional inputs. You et al. [57] therefore assumed *permutation invariance*; that the positions of farmland pixels in a county do not affect yield, since they only indicate the positions of cropland. This allows all cropland pixels (selected using the MODIS land cover map [14]) in a county to be mapped to a histogram of pixel values, significantly reducing the dimensionality of the input. We constructed meta-learning tasks by defining tasks as individual counties, with task (X, y) pairs consisting of histograms and yields for different years.

¹All LP DAAC current data and products acquired through the LP DAAC have no restrictions on reuse, sale, or redistribution.

Task Metadata As in Section 4.1, we included the Cartesian-coordinate location of each task’s county. We additionally included a one-hot encoding of which U.S. state the county is in.

Evaluation We used temporal validation to evaluate model performance: for each year in $\{2011, 2012, 2013, 2014, 2015\}$, we trained a model using all the data prior to that year, and evaluated the performance of the model for that year.

4.2.1. Experiments

We applied TIML to the network architectures originally used by [57] – an LSTM and a CNN-based regressor. In addition to the remote sensing input, the Deep Gaussian Process baseline model (described below) receives as input the year of each training point. We therefore appended the year to each timestep of the input to the TIML LSTMs, so the model has comparable inputs to the Deep Gaussian Process. The CNN models receive only the remote sensing data as input.

Baselines We compared TIML to 2 baselines: the Deep Gaussian Process models (proposed by [57] with the yield estimation data set) and standard MAML. To train a Deep Gaussian Process, a deep learning model is first trained to estimate yield given the remote sensing data set described above. The final hidden vector $h(x)$ of the model (for each input) is used as input to a Gaussian process $y(x) = f(x) + h(x)^T$ where $f(x) \sim \mathcal{GP}(0, k(x, x'))$. The kernel function k is conditioned on both the location of the datapoint (defined by its latitude and longitude) and the year of the datapoint. We included baselines with and without a Gaussian process (i.e., using the outputs of the deep learning models directly instead of passing the final hidden vectors to a Gaussian process). We note that this implementation of Deep Gaussian Processes by [57] differs from [12]. Finally, we highlight that the MAML LSTM model also receives the year appended as input, as is the case for the TIML LSTM model.

The MODIS data sets have been updated from version 5 to 6 since the original Deep Gaussian Process models were run. We therefore retrained the models to obtain our baseline results. We used the same hyperparameters as [57], with the addition of early stopping during training. We included the original results from [57] for comparison.

4.3. Grouped-Omniglot

The Omniglot data set [24] is a one-shot learning data set consisting of 1,623 handwritten characters drawn from 50 alphabets. We constructed tasks by considering only characters from a single alphabet together – a task therefore consists of one-shot classification of *characters drawn from the*

Model	2011	2012	2013	2014	2015	Mean
LSTM	5.62 ± 0.10	6.60 ± 0.29	5.57 ± 0.21	6.63 ± 0.13	6.69 ± 0.31	6.22
+ GP	5.32 ± 0.10	5.83 ± 0.18	5.70 ± 0.19	5.61 ± 0.12	5.24 ± 0.14	5.54
+ MAML	26.90 ± 0.01	30.97 ± 0.01	29.57 ± 0.01	30.84 ± 0.01	32.02 ± 0.01	30.06
+ TIML	5.16 ± 0.03	5.77 ± 0.05	5.39 ± 0.02	5.24 ± 0.04	4.89 ± 0.04	5.29
CNN	6.08 ± 0.77	6.94 ± 1.83	6.42 ± 1.23	4.80 ± 0.83	5.57 ± 0.38	5.96
+ GP	5.55 ± 0.14	6.18 ± 0.49	6.44 ± 0.67	4.87 ± 0.31	6.02 ± 0.26	5.81
+ MAML	12.93 ± 0.05	8.28 ± 0.07	7.98 ± 0.04	12.05 ± 0.05	7.69 ± 0.06	9.79
+ TIML	5.23 ± 0.02	6.59 ± 0.02	5.34 ± 0.01	4.93 ± 0.02	6.35 ± 0.01	5.69
[57]						
LSTM + GP	5.77	6.23	5.96	5.70	5.49	5.83
CNN + GP	5.70	5.68	5.83	4.89	5.67	5.55

Table 2. The RMSE of county-level model performance for the **yield estimation** task. We use temporal validation to evaluate the model. Specifically, for each year, models are trained with data up to that year and evaluated with that year’s data. All models are calculated from an average of 10 runs, with the standard error reported. We highlight the **first** and **second** best metrics for each task. For completeness, we include the results reported by [57], but highlight that these results were obtained on the MODIS 5.1 data set (whilst all other models were trained on the MODIS 6.0 data set) and are the result of 2 runs, compared to 10 runs for all other models. TIML improves on the Deep Gaussian Process models for both architectures, even though MAML performs significantly worse than other models. This suggests that in some cases, the task information is necessary for meta-learning to work.

same alphabet. We highlight that this is a much more challenging setup than the typical setup of mixing all characters together, since more similar characters will need to be differentiated. We used a 5-way 1-shot regime to train and evaluate the model.

Task Metadata As task metadata, we used a one-hot encoding the alphabet from which a task is drawn. This metadata is less detailed than the crop classification and yield estimation tasks and is intended to demonstrate the utility of TIML even in scenarios with minimal metadata.

Evaluation For evaluation, we selected 5 characters per alphabet and held them out from the training set. We evaluated the models by fine-tuning them on a single example (per alphabet-set), and measuring the model accuracy across all remaining examples per character.

4.3.1. Experiments

We evaluated TIML using the CNN architecture proposed for Omniglot and trained the model for 60,000 steps as in Finn et al. [13]. As **baselines**, we compared TIML to MAML and MMAML [51], which we trained using the same model and training procedure. We emphasize that the Omniglot data set is one of the data sets originally used by MAML [13] and MMAML [51].

5. Results

In this section, we describe the results of TIML, its ablations and the benchmark models for each data set.

5.1. Crop Type Classification

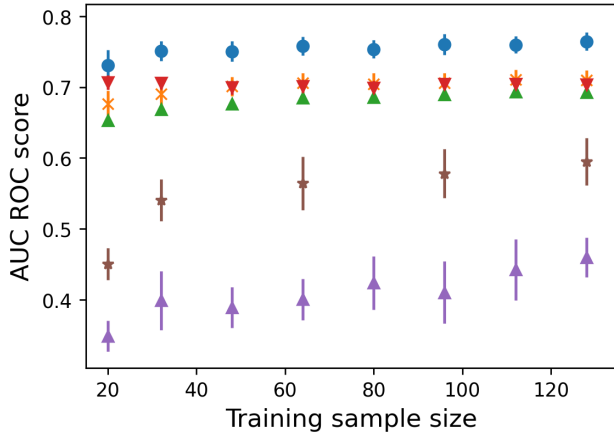
Table 1 shows the model results for TIML, its ablations and all baseline models when trained on the CropHarvest data set. Like Tseng et al. [48], we report the AUC ROC score and the F1 score calculated using a threshold of 0.5. Overall, TIML is the best performing algorithm on the CropHarvest data set, achieving the highest mean F1 and AUC ROC scores. TIML is consistently the best performing algorithm on every task.

TIML excels at learning from small data set sizes. It is the only transfer/meta-learning model that outperforms a randomly-initialized model in the challenging Brazil task, where there are only 26 positive datapoints. We plot the performance of the models as a function of training set size in Figure 2 for the Kenya and Togo evaluation tasks (the Brazil task is already in the small-data set size regime). In both the Kenya and Togo tasks, TIML achieves the highest or near-highest ROC AUC scores for all subset sizes, and its advantage over other algorithms increases for smaller sample sizes.

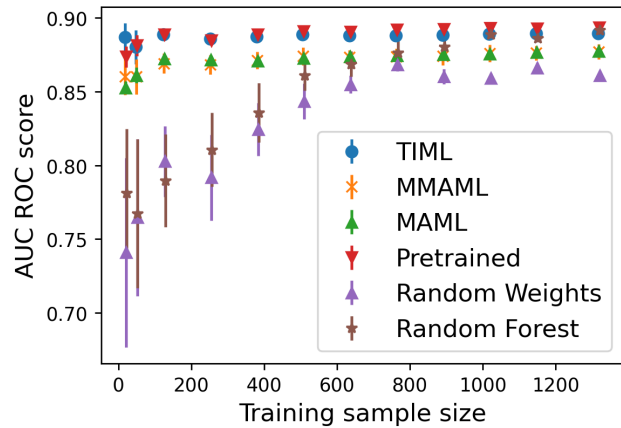
5.2. Yield Estimation

We report the results for the yield estimation data set in Table 2. Like You et al. [57], we report the RMSE score averaged across all counties and use temporal validation to evaluate the models. The LSTM models used in TIML and MAML receive the year as input (to match the data provided to the Deep Gaussian Process), but the CNN models do not.

For both the LSTM and CNN architectures, TIML is the most performant model. This is the case even though the



(a) Kenya: Maize vs. Rest



(b) Togo: Crop vs. Non Crop

Figure 2. Results of TIML and the benchmark models when trained on a subset of the evaluation training data for the Crop Type Classification Task. We plot results for (Figure 2a) the Kenya Maize vs. rest evaluation task and (Figure 2b) Togo Crop vs. Non Crop evaluation task. Results are averaged from 10 runs and reported with standard error bars. Subsets are balanced so that they contain an equal number of positive and negative samples. For all training set sizes, TIML is the best performing model in the Kenya task and at or near best performance in the Togo task. We highlight that the advantage of TIML over other algorithms in general increases for smaller training sets.

Deep Gaussian Process is much more memory-intensive, since it requires all predictions and hidden vectors (for the training and test data) to be computed together for the Gaussian process modelling step; this may be infeasible for larger data sets. TIML requires substantially less memory since it considers each county independently. It is also worth noting that while TIML achieves the best result of all models, MAML performs significantly worse than all other models. This suggests that in some contexts, the task metadata is necessary for meta-learning to work.

5.3. Grouped Omniglot

The results on grouped-omniglot for the 5-way 1-shot task are shown in Table 3. We re-emphasize the difficulty of this setup (which groups similar characters together) compared to usual [13] approach of mixing all the characters, reflected in significantly poorer performance for the MAML baseline relative to this usual regime. The task metadata in this case is minimal, consisting of a one-hot encoding representing the alphabet a character-set was drawn from. Nonetheless, we find that TIML improves on both MAML and MMAML, indicating its effectiveness even with relatively little metadata.

6. Discussion

In this section, we discuss the key components of the TIML algorithm, and their contribution to the overall results. Specifically, we discuss experimental evidence for the importance of (i) metadata encodings, (ii) forgetfulness and (iii) task-explicit vs. task-inferred modulations.

	Accuracy (%)
MAML	80.93 ± 1.06
MMAML	81.53 ± 1.03
TIML	82.89 ± 0.98

Table 3. Grouped-omniglot results, averaged from 3 random seeds with standard error, with the best results highlighted. TIML improves on both MAML and MMAML.

6.1. Importance of metadata encoding

The success of TIML is due not just to the presence of task metadata but also to the way that metadata is encoded and passed to the meta-learner. In the “no encoder” ablation conducted on the **crop type classification** task, we provide task metadata directly to the learner by concatenating it with the input data. This approach does improve somewhat upon standard MAML, indicating the helpfulness of task metadata in learning, but it performs significantly worse than the full TIML algorithm.

6.2. Importance of forgetfulness

On the **crop type classification** task, we observe that using standard MAML or pre-training using global crop data actually results in lower performance on the Brazil task compared to an LSTM initialized with random weights. We hypothesize this may be due to the difference in distribution of the Brazil task data relative to the other tasks the models are trained on.

TIML, by contrast, performs significantly better than a

randomly initialized LSTM, and indeed much better than all other methods. We see that forgetfulness plays a key role here, as training TIML without forgetfulness results in similar performance to the randomly initialized LSTM. However, it is not just forgetfulness that is key here, as training TIML with forgetfulness but without the metadata encoder causes the F1 score to drop precipitously. We therefore hypothesize that task information provides useful context around which tasks are being kept and forgotten during training, allowing TIML to learn from more difficult tasks in the “forgetful” regime without forgetting easier tasks it has already learned. Training TIML with forgetfulness significantly boosts performance in the Brazil task without substantially impacting performance on the other tasks, and yields significantly higher mean F1 and AUC ROC scores when measured across all tasks.

6.3. Task-explicit vs. task-inferred modulation

On the **grouped-omniglot** and **crop type classification** tasks, we compare the effect of passing explicit task information via TIML and of inferring this task information via MMAML [51]. Overall, we find that when task metadata is present, it can lead to a significant improvement in performance compared to inferring task clusters.

We again highlight that this improvement specifically comes when the task information is added using TIML (Section 6.1). To our knowledge, TIML is the first approach that aims to leverage explicit metadata in gradient-based meta-learning algorithms.

6.4. Limitations

Meta-learning - and therefore TIML - relies on large labelled pre-training dataset exists from which tasks can be constructed. In geospatial contexts, this typically means semantically similar labels distributed over large areas (e.g. global crop type labels, as is the case for CropHarvest). In addition, TIML requires modifications to model layers via modulation - while we demonstrate that this is applicable across different model architectures, this requirement introduces complexity in the implementation of TIML.

7. Conclusion

We introduce task-informed meta-learning (TIML), a method for conditioning meta-learning models with explicit task metadata. The metadata is encoded into a set of vectors which are used to modulate the weights learned by a MAML learner prior to task-specific fine-tuning. TIML also includes a new technique called “forgetfulness,” which we show can improve performance when there are many similar tasks to learn from. We evaluated TIML for both classification and regression tasks using a variety of neural network architectures (recurrent and convolutional networks), demonstrating its utility in a vari-

ety of regimes—including those with very few data points and those for which standard MAML fails completely. In addition, TIML outperforms naïve methods for incorporating task-metadata and traditional clustering-based task-adaptive methods, achieving state-of-the-art performance on crop-classification and yield estimation tasks. While usefulness for societal impact and geographic equity motivated us to focus in particular on remote sensing for agriculture tasks, we showed that TIML is not specific to agriculture and can also be useful in other meta-learning problems with task-level metadata.

References

- [1] Programme, concepts and definitions. In *World Programme for the Census of Agriculture*. FAO, 2020. 4
- [2] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your MAML. In *International Conference on Learning Representations (ICML)*, 2019. 11
- [3] Sébastien M R Arnold, Praateek Mahajan, Debajyoti Datta, Ian Bunner, and Konstantinos Saitas Zarkias. learn2learn: A library for Meta-Learning research. 2020. 11
- [4] Kumar Ayush, Burak Uzket, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021. 2
- [5] Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. Meta-learning with adaptive hyperparameters. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [6] Sungyong Baik, Seokil Hong, and Kyoung Mu Lee. Learning to forget for meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3
- [7] Sungyong Baik, Janghoon Choi, Heewon Kim, Dohee Cho, Jaesik Min, and Kyoung Mu Lee. Meta-learning with task-adaptive loss function for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [8] Inbal Becker-Reshef, Christina Jade Justice, Brian Barker, Michael Laurence Humber, Felix Rembold, Rogerio Bonifacio, Mario Zappacosta, Mike Budde, Tamuka Magadzire, Chris Shitote, Jonathan Pound, Alessandro Constantino, Catherine Nakalembe, Kenneth Mwangi, Shinichi Sobue, Terence Newby, Alyssa Whitcraft, Ian Jarvis, and James Verdin. Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM crop monitor for early warning. *Remote Sensing of Environment*, 2020. 1, 5
- [9] Léonard Boussioux, Cynthia Zeng, Dimitris Bertsimas, and Théo J Guenais. Hurricane forecasting: A novel multimodal machine learning framework. In *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*, 2021. 1
- [10] Tony Chang, Brandon P Rasmussen, Brett G Dickson, and Luke J Zachmann. Chimera: A multi-task recurrent convolutional neural network for forest classification and structural estimation. *Remote Sensing*, 2019. 2

- [11] Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3
- [12] Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013. 5
- [13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)*, 2017. 1, 2, 6, 7, 11
- [14] Mark A. Friedl, Damien Sulla-Menashe, Bin Tan, Annemarie Schneider, Navin Ramankutty, Adam Sibley, and Xiaoman Huang. MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, 2010. 5
- [15] Rui Han, Chi Harold Liu, Shilin Li, Lydia Y. Chen, Guoren Wang, Jian Tang, and Jieping Ye. Slimml: Removing non-critical input data in large-scale iterative machine learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 3
- [16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 3
- [17] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3
- [18] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 2016. 2
- [19] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2
- [20] Hannah Kerner, Gabriel Tseng, Inbal Becker-Reshef, Catherine Nakalembe, Brian Barker, Blake Munshell, Madhava Paliyam, and Mehdi Hosseini. Rapid response crop maps in data sparse regions. In *ACM SIGKDD Conference on Data Mining and Knowledge Discovery Workshops*, 2020. 2
- [21] Wonsik Kim, Toshichika Iizumi, and Motoki Nishimori. Global patterns of crop production losses associated with droughts from 1983 to 2009. *Journal of Applied Meteorology and Climatology*, 2019. 5
- [22] Dan M Kluger, Sherrie Wang, and David B Lobell. Two shifts for crop mapping: Leveraging aggregate crop statistics to improve satellite-based maps in new regions. *Remote Sensing of Environment*, 2021. 2
- [23] Satish Kumar, Carlos Torres, Oytun Ulutan, Alana Ayasse, Dar Roberts, and B.S. Manjunath. Deep remote sensing methods for methane detection in overhead hyperspectral imagery. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1
- [24] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. 5
- [25] Nico Lang, Konrad Schindler, and Jan Dirk Wegner. Country-wide high-resolution vegetation height mapping with sentinel-2. *Remote Sensing of Environment*, 2019. 1
- [26] Tyler J. Lark, Seth A. Spawn, Matthew Bougie, and Holly K. Gibbs. Cropland expansion in the United States produces marginal yields at high costs to wildlife. *Nature Communications*, 2020. 5
- [27] Hae Beom Lee, Hayeon Lee, Donghyun Na, Saehoon Kim, Minseop Park, Eunho Yang, and Sung Ju Hwang. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 3
- [28] Oisín Mac Aodha, Elijah Cole, and Pietro Perona. Presence-Only Geographical Priors for Fine-Grained Image Classification. In *International Conference on Computer Vision (ICCV)*, 2019. 2
- [29] Michael Matiu, Donna P Ankerst, and Annette Menzel. Interactions between temperature and drought in global and regional crop yield variability during 1961-2014. *PLoS one*, 12 (5):e0178339, 2017. 5
- [30] L. Ohno-Machado, H. S. Fraser, and A. Ohn. Improving machine learning performance by removing redundant cases in medical data sets. *Proceedings. AMIA Symposium*, 1998. 3
- [31] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems (NeurIPS)*, 2018. 2
- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 11
- [33] Tom Patterson and Nathaniel Vaughn Kelso. Natural Earth. <https://www.naturalearthdata.com/>. 3
- [34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 2011. 5
- [35] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3, 12, 13
- [36] Hans-Otto Pörtner, Debra C Roberts, H Adams, C Adler, P Aldunce, E Ali, R Ara Begum, R Betts, R Bezner Kerr, R Biesbroek, et al. Climate change 2022: Impacts, adaptation and vulnerability. *IPCC Sixth Assessment Report*, 2022. 5
- [37] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017. 2
- [38] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification.

- In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1, 2
- [39] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *International Conference on Learning Representations (ICLR)*, 2018. 2
- [40] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017. 1
- [41] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. On modulating the gradient for meta-learning. In *European Conference on Computer Vision*. Springer, 2020. 2
- [42] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2
- [43] Xiao-Peng Song, Matthew C. Hansen, Peter Potapov, Bernard Adusei, Jeffrey Pickering, Marcos Adami, Andre Lima, Viviana Zalles, Stephen V. Stehman, Carlos M. Di Bella, Maria C. Conde, Esteban J. Copati, Lucas B. Fernandes, Andres Hernandez-Serna, Samuel M. Jantz, Amy H. Pickens, Svetlana Turubanova, and Alexandra Tyukavina. Massive soybean expansion in South America since 2000 and implications for conservation. *Nature Sustainability*, 2021. 3
- [44] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 2014. 3
- [45] Sebastian Thrun and Lorien Pratt. *Learning to Learn*. Springer Science & Business Media, 1998. 2
- [46] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 2
- [47] Gabriel Tseng, Hannah Kerner, Catherine Nakalembe, and Inbal Becker-Reshef. Learning to predict crop type from heterogeneous sparse labels using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 1, 2
- [48] Gabriel Tseng, Ivan Zvonkov, Catherine Nakalembe, and Hannah Kerner. CropHarvest: a global satellite dataset for crop type classification. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021. 1, 2, 3, 4, 6, 11
- [49] Mehmet Ozgur Turkoglu, Stefano D’Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 2021. 1
- [50] Eric Vermote. MODIS/terra surface reflectance 8-day 13 global 500m SIN grid v006, 2015. 5
- [51] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J. Lim. Multimodal model-agnostic meta-learning via task-aware modulation. In *Neural Information Processing Systems (NeurIPS)*, 2019. 2, 4, 5, 6, 8
- [52] Zhengming Wan, Simon Hook, and Glynn Hulley. MODIS/aqua land surface temperature/emissivity 8-day 13 global 1km SIN grid v006, 2015. 5
- [53] Anna X. Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS)*, 2018. 1, 2
- [54] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 3
- [55] Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness. In *International Conference on Learning Representations (ICLR)*, 2020. 2
- [56] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. Hierarchically structured meta-learning. In *International Conference on Machine Learning (ICML)*, 2019. 2
- [57] Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 2, 5, 6
- [58] Qing Zhu, Cheng Liao, Han Hu, Xiaoming Mei, and Haifeng Li. Map-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 1