SPECTRAL GAPS AND SPATIAL PRIORS: STUDYING HYPERSPECTRAL DOWNSTREAM ADAPTATION USING TERRAMIND

Julia A. Leonardi & Maria A. Brovelli
Department of Civil and Environmental Engineering
Politecnico di Milano
Milan, Italy
{juliaanna.leonardi, maria.brovelli}@polimi.it

Johannes Jakubik & Paolo Fraccaro
IBM Research Europe
Rüschlikon, Switzerland
{johannes.jakubik1, paolo.fraccaro}@ibm.com

ABSTRACT

Geospatial Foundation Models (GFMs) typically lack native support for Hyperspectral Imaging (HSI) due to the complexity and sheer size of high-dimensional spectral data. This study investigates the adaptability of TerraMind, a multimodal GFM, to HSI downstream tasks *without* HSI-specific pretraining by comparing two channel adaptation strategies: Naive Band Selection and physics-aware Spectral Response Function (SRF) grouping. Our results confirm the general superiority of HSI-native architectures, though TerraMind demonstrates robust adaptability to spectral tasks through simple band selection. By establishing this baseline, we underscore the necessity of developing native spectral tokenization for future multimodal GFMs.

1 INTRODUCTION

Geospatial Foundation Models (GFMs) have emerged as a paradigm shift in remote sensing, moving from task-specific architectures to generalist models pretrained on vast datasets. GFMs aim to learn scalable, transferable representations across diverse downstream tasks, which are generalizable to a global scale. However, remote sensing presents unique challenges due to high heterogeneity in sensor modalities. Although modern GFMs can incorporate diverse input types, from optical imagery to SAR, Hyperspectral Imaging (HSI), which is essential for precision agriculture, mineral exploration, and environmental monitoring, remains underrepresented. HSI offers significant spectral detail through hundreds of narrow spectral channels, but poses challenges in data complexity. Although the availability of HSI data is increasing via satellite missions like EnMAP (Kaufmann et al., 2012) and PRISMA (Guarini et al., 2018), Deep Learning (DL) applications remain task-specific (Ibañez et al., 2022; Scheibenreif et al., 2023; Xu et al., 2026). Recent works have introduced HSI-specific GFMs, such as HyperSIGMA (Wang et al., 2025) or SpectralEarth (Braham et al., 2025). However, these are largely unimodal. Current multimodal GFMs that incorporate HSI into the pretraining data (e.g., DOFA (Xiong et al., 2025)) rarely address the distinct three-dimensional feature extraction required for HSI, thereby creating a gap in effective multimodal integration. An additional challenge concerns the limited availability and diversity of downstream datasets suited for DL applications, with most benchmark HSI datasets being single scenes, not allowing for a proper training process and hindering generalization (Signoroni et al., 2019). In this work, we address whether existing multimodal GFMs, pretrained without HSI, can serve as effective baselines for HSI-specific tasks.

We fine-tune TerraMind (Jakubik et al., 2025), a multimodal GFM, on four different HSI downstream tasks. To bridge the modality gap, we compare two channel adaptation strategies: Naive Band Selection and Spectral Response Function (SRF) grouping, to align high-dimensional HSI in-

puts with the model's Sentinel-2 pretraining distribution. The experimental workflow of our research is presented in Figure 1. Our contributions are threefold. First, we present a baseline evaluation of the TerraMind GFM on four HSI-specific downstream tasks. Second, we perform a comparative analysis of channel sampling approaches for adapting HSI data to non-HSI architectures. Finally, we assess the potential of multimodal GFMs for spectral tasks by benchmarking our results against the HSI-native SpectralEarth model.

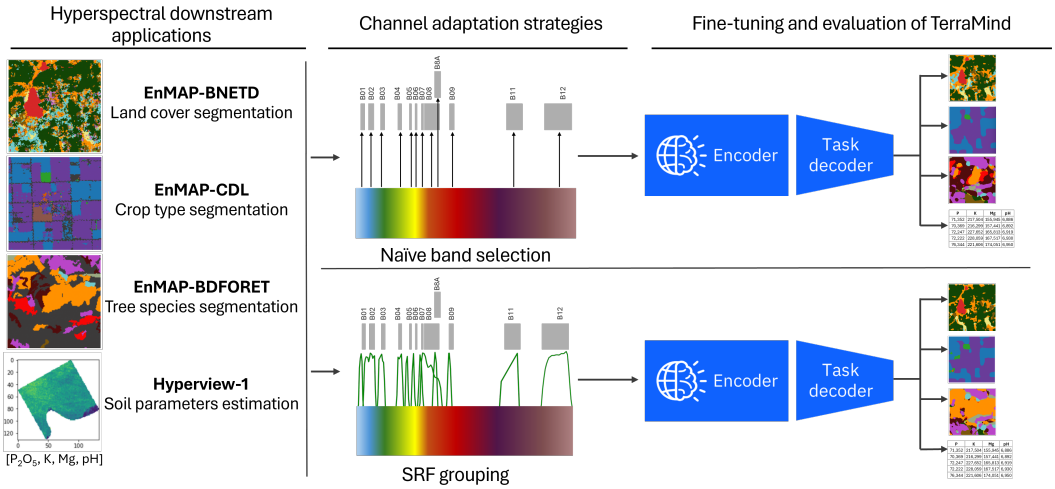

Figure 1: Experimental framework for benchmarking channel adaptation strategies (Naive Selection vs. SRF Grouping) using the TerraMind backbone on four HSI-specific applications.

2 Experimental Setup

Downstream Datasets. We evaluate performance on four HSI benchmarks selected to represent a range of spectral difficulty. We define 'spectral complexity' both by class granularity, ranging from broad land-cover mapping to fine-grained tree species identification, and empirically, based on the performance gap observed when reducing full hyperspectral inputs to 12 multispectral bands. Three datasets are from SpectralEarth (Braham et al., 2025), aligning 202-band EnMAP imagery (128×128 px chips) with ground truth masks, and the fourth is derived from the AI4EO challenge (Nalepa et al., 2024). **EnMAP-BNETD** (2,100 chips, 10 classes) is a broad land cover task, where multispectral resolution is generally sufficient for discrimination. **EnMAP-CDL** (1,600 chips, 14 classes) presents a moderately complex crop segmentation task. **EnMAP-BDForet** (2,550 chips, 12 species) is a fine-grained tree species segmentation task. Distinguishing spectrally similar subtypes requires high spectral resolution, posing a strict challenge for non-HSI GFMs. **Hyperview-1** (1,732 train/1,154 test chips, 150 bands) is a scene-level regression dataset of four soil parameters: Potassium K, Phosphorus pentoxide P_2O_5, Magnesium Mg, and acidity pH, representing the highest spectral complexity.

We report the mean Intersection over Union (mIoU) for the three segmentation tasks. For Hyperview-1, we use the challenge's normalized MSE: $\text{MSE}_{\text{norm}} = \sum_i^4 (\text{MSE}_i / \text{MSE}_i^{\text{base}})$, where $\text{MSE}_i^{\text{base}}$ is the error of a trivial mean-predictor on the training set.

TerraMind & Channel Adaptation. TerraMind is a multimodal GFM pretrained on multi-spectral, SAR, Digital Terrain Model (DEM), and Land Use/Land Cover (LULC) data, along with derived spectral index layers. The optical component of the pretraining dataset consists of Sentinel-2 data (both L2A and L1C). To leverage the pretrained weights, we project the input HSI ($X_{\text{HSI}} \in \mathbb{R}^{H \times W \times C_{\text{in}}}$) into the S2L2A spectral space. We investigate two projection methods to obtain the adapted input $\hat{X} \in \mathbb{R}^{H \times W \times 12}$.

Naive Band Selection: We select the HSI bands centered closest to the nominal center wavelengths of the Sentinel-2 bands. Let $\Lambda_{\text{HSI}} = \{\lambda_1, \ldots, \lambda_{C_{\text{in}}}\}$ be the center wavelengths of the HSI sensor,

and $\Lambda_{\text{S2}} = \{\mu_1, \ldots, \mu_{12}\}$ be the target S2 wavelengths. For each S2 band k, we select:

$$\hat{X}_{:,:,k} = X_{\text{HSI}}\left[:, :, \underset{j}{\operatorname{argmin}} |\lambda_j - \mu_k|\right] \tag{1}$$

as the feature map. This method preserves the raw radiometric values of specific narrow bands but discards information from the rest of the spectrum.

SRF-based Spectral Resampling: To simulate a physically realistic S2 signal, we apply the Sentinel-2 SRF, mathematically following the spectral simulation approach in Avilés et al. (2025). This operation is implemented as a fixed linear transformation (weighted sum) along the spectral dimension. Let $\phi_k(\lambda)$ be the continuous SRF for Sentinel-2 band k, and λ_j be the center wavelength of HSI band j. We construct a weight matrix $\mathbf{W} \in \mathbb{R}^{C_{in} \times 12}$ where the unnormalized weights correspond to the SRF sensitivity at each HSI band center. To preserve radiometric consistency, we normalize the weights for each S2 band k to sum to unity: $w_{j,k} = \phi_k(\lambda_j)$ and $\hat{w}_{j,k} = w_{j,k} / \sum_{m=1}^{C_{in}} w_{m,k}$.

The final matrix $\widehat{\mathbf{W}} = \{\hat{w}_{j,k}\}$ is then applied to the input HSI data $X_{\text{HSI}} \in \mathbb{R}^{H \times W \times C_{in}}$ via a matrix-vector product along the spectral dimension, which aggregates information from all HSI bands falling within the S2 spectral range, providing a smoother, physics-aware representation:

$$\hat{X}_{h,w,:} = X_{\text{HSI}_{h,w,:}} \widehat{\mathbf{W}}, \quad \forall h, w. \tag{2}$$

Implementation Details. We fine-tune the `terramind_v1_base` backbone, initializing the optical modality with S2 weights. For comparability, we closely follow SpectralEarth's implementation setup. All experiments were implemented using the `terratorch` library (Gomes et al., 2025), which streamlines the adaptation of TerraMind to specific downstream tasks through its modular framework. We employ a Fully Convolutional decoder (256 channels) optimized using cross-entropy loss in the segmentation tasks. For regression, we attach a linear head (hidden dimension 256) optimized via MSE loss. All models are trained for 100 epochs using the AdamW optimizer, with early stopping (patience: 20). We utilize a Cosine Annealing scheduler for segmentation and ReduceOnPlateau for regression. To ensure statistical robustness, all experiments were repeated 10 times with different random seeds. Experiments were conducted on NVIDIA A100 SXM4 GPUs.

3 Empirical Findings

Table 1 summarizes the results of our channel adaptation experiments, along with the HSI-native baselines reported in Braham et al. (2025). Qualitative results of the experiments with a short analysis can be found in the Appendix A. We observe two consistent patterns across the four datasets: **First**, the Naive Band Selection strategy consistently yields superior performance compared to the physics-aware SRF Grouping. In segmentation tasks, Naive Selection offers a gain of approximately +0.4% to +3.4% mIoU, while in the Hyperview-1 regression task, it achieves a significantly higher leaderboard rank (#6) compared to SRF (#25). This significant performance difference between the two band selection methods can suggest that TerraMind's pretraining has induced a strong sensitivity to specific spectral anchors corresponding to the Sentinel-2 center wavelengths. Naive selection preserves the raw radiometric distribution of these central wavelengths, making it a better simulation of S2 data for TerraMind. In contrast, SRF grouping creates a weighted-average signal along the spectral dimension. This acts as a 1D spectral low-pass filter, smoothing out sharp, narrow-band absorption features that are critical for discrimination. While physically more representative of the sensor properties, they represented a shift in TerraMinds learned representations. **Second**, comparing TerraMind to the HSI-native SpectralEarth baseline reveals a clear correlation between the performance gap and the spectral complexity of the task. On EnMAP-BNETD, the gap is small ($\sim$3% mIoU), indicating that the task is "spectrally easy" and the model's pretrained spatial features are sufficient to compensate for the reduced spectral resolution (12 vs. 202 bands). However, on the "Moderate" CDL and "Hard" BDForet datasets, which require distinguishing between spectrally similar agricultural classes or tree species, the performance gap widens to $\sim$8% and $\sim$11%, respectively. This confirms that for fine-grained classification, the 12-band approximation is insufficient to capture the nuanced spectral signatures present in the full HSI input, regardless of the GFM's spatial

Table 1: Comparative performance on downstream tasks (mean $\pm$ std over 10 runs). Metrics are mIoU for segmentation (higher is better) and Normalized MSE for regression (lower is better). **Bold** indicates the best performing model; underline indicates the best adaptation strategy for TerraMind.

Model	Adaptation	Segmentation Tasks (mIoU ↑)			Regression (nMSE ↓)
		EnMAP-BNETD *(Easy)*	**EnMAP-CDL** *(Moderate)*	**EnMAP-BDForet** *(Hard)*	**Hyperview-1** *(V. Hard)*
TerraMind	Naive Selection	$\underline{0.465 \pm 0.002}$	$\underline{0.693 \pm 0.006}$	$\underline{0.657 \pm 0.007}$	$\underline{0.813}$ (Rank #6)
	SRF Grouping	0.461 ± 0.003	0.679 ± 0.006	0.623 ± 0.006	0.831 (Rank #25)
SpectralEarth *(Upper Bound)*	None (Full HSI)	$\mathbf{0.495 \pm 0.001}$	$\mathbf{0.774 \pm 0.003}$	$\mathbf{0.766 \pm 0.005}$	**0.810** (Rank #5)

priors. Notably, the lower absolute mIoU on the 'easy' task results from its dense annotations, which penalize spatial errors more strictly than the sparsely annotated harder tasks (see Appendix A).

Surprisingly, on the Hyperview-1 soil parameter estimation task, defined as the most spectrally challenging dataset, the Naive Selection approach with TerraMind performs competitively with the specialized SpectralEarth (0.813 vs 0.810 normalized MSE). This result indicates that the spatial representations learned by TerraMind can compensate for reduced spectral resolution, but it could also point to significant spectral redundancy in the task itself. Research in soil spectroscopy demonstrates that nutrients like P, K, and Mg can be detected indirectly via correlations with primary constituents like Organic Matter (OM) and clay minerals (Bogrekci & Lee, 2005; Demattê et al., 2017). These constituents exhibit broad spectral responses that align closely with Sentinel-2's bands. Specifically, OM in the visible/red-edge and clay minerals in the SWIR (centered near 2200 nm). Consequently, the Naive Selection effectively captures these proxy signals while filtering out the noise inherent in the full hyperspectral continuum, allowing the TerraMind to perform effectively despite discarding $> 90\%$ of the bands. Finally, the very small performance gap between TerraMind and SpectralEarth suggests a limit to performance on this dataset inherent to the task difficulty or label noise. This limit cannot be easily overcome with increased spectral resolution for generalist foundation models, implying that such problems demand more specialized modeling approaches.

4 Discussion & Conclusion

This study investigated adapting multimodal GFMs not pretrained on hyperspectral data to hyperspectral downstream tasks. We did so by examining two techniques of HSI channel sampling. The results of our preliminary research on the integration of HSI into a multimodal GFM aim to answer the research question: Can a non-HSI GFM serve as an effective baseline for HSI-specific tasks?

Our findings suggest TerraMind produces a competitive baseline when the downstream task prioritizes spatial semantics over spectral precision. On EnMAP-BNETD, the model reached within 3% of the HSI-native SpectralEarth model, leveraging pre-learned spatial representations to compensate for the massive spectral reduction ($202 \rightarrow 12$ bands). However, for tasks requiring fine-grained spectral discrimination, such as EnMAP-BDForet, the "spectral gap" cannot be overcome via simple subsampling. This highlights that while existing multimodal GFMs are powerful generalist tools, they cannot replace specialized HSI architectures for tasks where precise spectral information guides prediction. Counter-intuitively, Naive Band Selection consistently outperformed physics-based SRF Grouping. We attribute this to how TerraMind anchors its representations to specific S2L2A bands, suggesting this may be a model-specific result that necessitates validation through similar experiments on other multimodal GFMs. Alternatively, the performance drop suggests that SRF sampling acts as a spectral low-pass filter (Avilés et al., 2025), smoothing out precise spectral features and thereby reducing their informativeness for class discrimination.

These preliminary results underscore the need to move beyond simple adaptation toward full integration. Future research will extend benchmarks to other spectrally intensive applications and sensor modalities (e.g., PRISMA, EMIT). Methodologically, we plan to integrate native HSI support within TerraMind via a hyperspectral tokenizer to ingest full-spectrum data. Finally, to fully isolate the contribution of learned spatial priors, future work will incorporate stronger baselines and empirically validate the model's representation geometry.

ACKNOWLEDGEMENTS

This study was funded by the Italian Space Agency, ASI, under convention n. 2024-28-HH.0 - CUP n. F43D24000110001.

We acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).

Gemini 3 was used to improve the manuscript's language.

REFERENCES

Ruben Gonzalez Avilés, Linus Scheibenreif, Nassim Ait Ali Braham, Benedikt Blumenstiel, Thomas Brunschwiler, Ranjini Guruprasad, Damian Borth, Conrad Albrecht, Paolo Fraccaro, Devyani Lambhate, and Johannes Jakubik. Hyperspectral vision transformers for greenhouse gas estimations from space. *arXiv preprint arXiv:2504.16851*, 2025. doi: 10.48550/arXiv.2504.16851.

I. Bogrekci and W. S. Lee. Spectral soil signatures and sensing phosphorus. *Biosystems Engineering*, 92(4):527–533, 2005. ISSN 1537-5110. doi: 10.1016/j.biosystemseng.2005.09.001.

Nassim Ait Ali Braham, Conrad M. Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. Spectralearth: Training hyperspectral foundation models at scale. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:16780–16797, 2025. doi: 10.1109/JSTARS.2025.3581451.

José A. M. Demattê, L. Ramirez-Lopez, K. P. P. Marques, and A. A. Rodella. Chemometric soil analysis on the determination of specific bands for the detection of magnesium and potassium by spectroscopy. *Geoderma*, 288:8–22, 2017. ISSN 0016-7061. doi: 10.1016/j.geoderma.2016.11.013.

Carlos Gomes, Benedikt Blumenstiel, Joao Lucas de Sousa Almeida, Pedro Henrique de Oliveira, Paolo Fraccaro, Francesc Marti Escofet, Daniela Szwarcman, Naomi Simumba, Romeo Kienzler, and Bianca Zadrozny. Terratorch: The geospatial foundation models toolkit. *arXiv preprint arXiv:2503.20563*, 2025. doi: 10.48550/arXiv.2503.20563.

R. Guarini, R. Loizzo, C. Facchinetti, F. Longo, B. Ponticelli, M. Faraci, M. Dami, M. Cosi, L. Amoruso, V. De Pasquale, N. Taggio, F. Santoro, P. Colandrea, E. Miotti, and W. Di Nicolantonio. Prisma hyperspectral mission products. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 179–182, 2018. doi: 10.1109/IGARSS.2018.8517785.

Damian Ibañez, Ruben Fernandez-Beltran, Filiberto Pla, and Naoto Yokoya. Masked auto-encoding spectral–spatial transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. ISSN 1558-0644. doi: 10.1109/TGRS.2022.3217892.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. Terramind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7383–7394, October 2025.

Hermann Kaufmann, S Förster, Hendrik Wulf, K Segl, Luis Guanter, M Bochow, U Heiden, A Müller, W Heldens, T Schneiderhan, et al. Science plan of the environmental mapping and analysis program (enmap). Enmap technical report, GFZ Data Services, 2012.

Jakub Nalepa, Lukasz Tulczyjew, Bertrand Le Saux, Nicolas Longépé, Bogdan Ruszczak, Agata M. Wijata, Krzysztof Smykala, Michal Myller, Michal Kawulok, Ridvan Salih Kuzu, Frauke Albrecht, Caroline Arnold, Mohammad Alasawedah, Suzanne Angeli, Delphine Nobileau, Achille

Ballabeni, Alessandro Lotti, Alfredo Locarini, Dario Modenini, Paolo Tortora, and Michal Gumiela. Estimating soil parameters from hyperspectral images: A benchmark dataset and the outcome of the hyperview challenge. *IEEE Geoscience and Remote Sensing Magazine*, 12(3):35–63, 2024. ISSN 2168-6831. doi: 10.1109/MGRS.2024.3394040.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2166–2176, 2023. doi: 10.1109/CVPRW59228.2023.00210.

Alberto Signoroni, Mattia Savardi, Annalisa Baronio, and Sergio Benini. Deep learning meets hyperspectral image analysis: A multidisciplinary review. *Journal of Imaging*, 5(5):52, 2019. ISSN 2313-433X. doi: 10.3390/jimaging5050052.

Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. Hypersigma: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025. doi: 10.1109/TPAMI.2025.3557581.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356v3*, 2025. doi: 10.48550/arXiv.2403.15356.

Yichu Xu, Di Wang, Hongzan Jiao, Lefei Zhang, and Liangpei Zhang. Mambamoe: Mixture-of-spectral-spatial-experts state space model for hyperspectral image classification. *Information Fusion*, 127:103811, 2026. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.103811.

A APPENDIX

Figures 2 through 4 display sample predictions from the fine-tuned TerraMind model on the segmentation datasets. The predictions confirm the findings in Section 3. First, performance degrades on spectrally complex tasks, with significant class confusion evident in Figure 4, where distinct tree species are mixed up due to insufficient spectral detail. The qualitative difference between adaptation strategies is also evident. In the first example of Figure 2, the SRF Grouping method fails to detect the 'Forest Gallery' class and produces over-smoothed features that do not match the fragmented Ground Truth. In contrast, Naive Band Selection better preserves these fine spatial structures (Figure 2, last row) and maintains better separation between pine classes (Figure 4).

We attribute the general coarseness of the spatial boundaries to the use of a standard 16×16 patch size, as supported by findings in Braham et al. (2025).

Finally, we note that the sparse annotations in CDL and BDForet result in the model predicting valid classes within 'ignore index' (background) regions. This sparsity masks spatial errors on the harder tasks, inflating their metrics compared to the more densely annotated BNETD, where every boundary mistake is penalized.

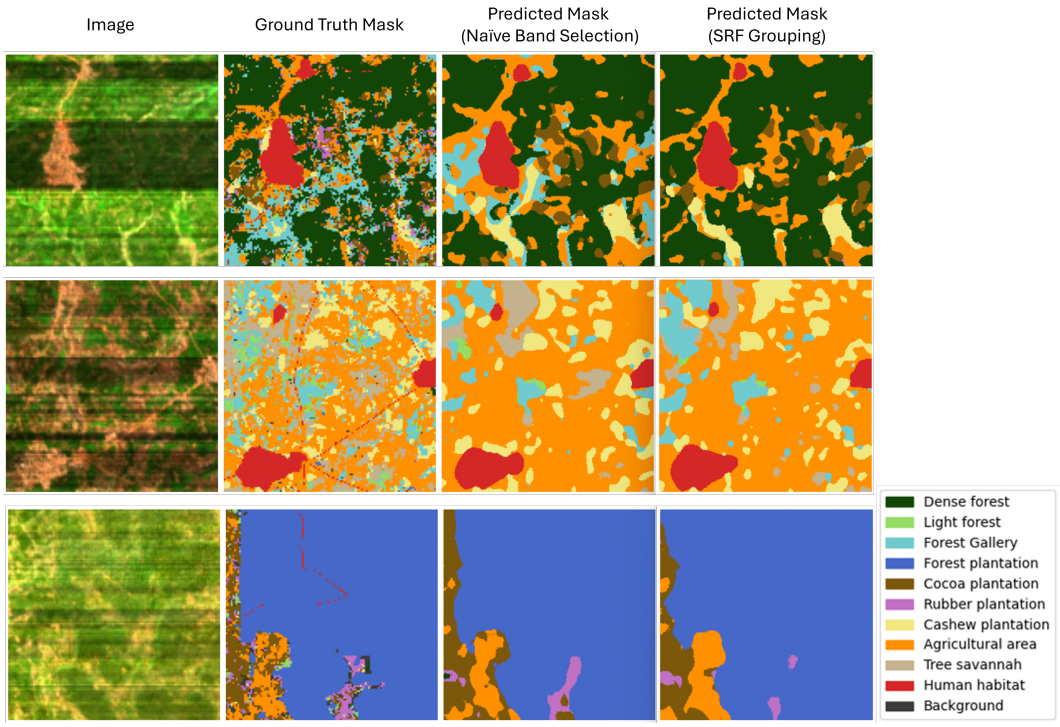

Figure 2: Prediction examples of TerraMind on the *EnMAP-BNETD* dataset. Compared to SRF Grouping (right), Naive Band Selection (center) better preserves fine spatial structures.

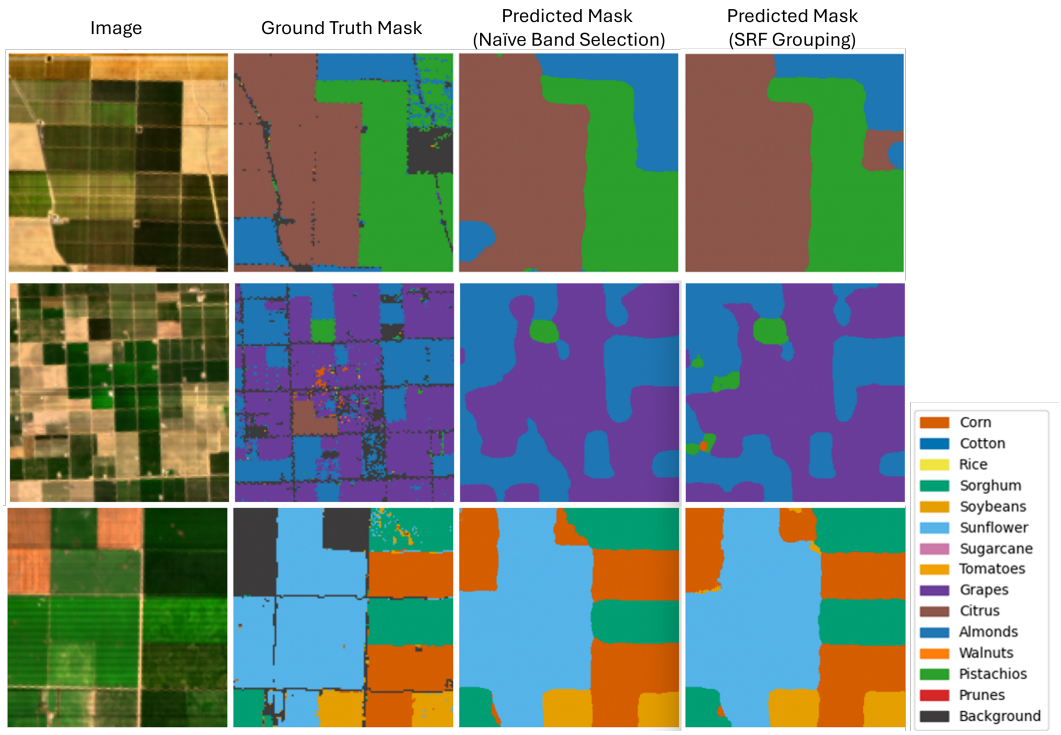

Figure 3: Prediction examples of TerraMind on the *EnMAP-CDL* dataset using both band sampling techniques. Note that the background class was set as the ignore index during training.

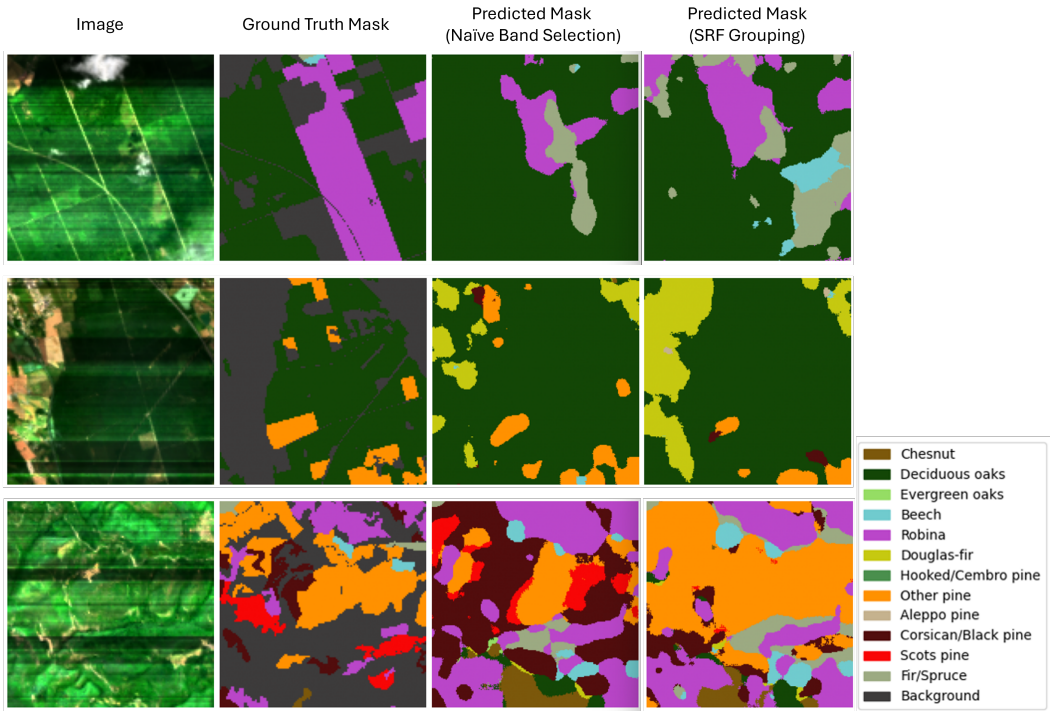

Figure 4: Prediction examples on *EnMAP-BDFORET*. The spectral similarity of tree species leads to visible class confusion in the SRF predictions (right), whereas Naive Band Selection (center) achieves better discrimination between pine subtypes. Note the extensive background regions (grey), which were set as the ignore index during training.