* **Model architecture**
  + Unet-plus tended to have a higher F-score and minimum of precision/recall for a given F-score compared to the other models (Fig. 1 and Fig. 2). Thus chose to look at only Unet-plus models for subsequent analysis.
* **Training validation (30% holdout of digitized training data)**
  + Best model (highest F-score) had 4 inputs: RGB, TPI, Hillshade, NDVI
    - Pixel-scale accuracy was 99.5%
    - Pixel-scale recall was 94.5%
    - Pixel-scale precision was 92.3%
* **Testing validation (~50% holdout of all digitized data)**
  + Based on the training validation results, chose to run the test validation on the top model for each number of inputs
    - Also chose to run test validation for the TPI-only model to investigate the performance of using only elevation data, for example from LiDAR. The TPI-only model was the second-best model with only one input from the training validation
  + At the pixel scale, F-scores were similar for all models (0.76 – 0.78), except the RGB-only model, which had an F-score of 0.68 (Table 1). This was a result of lower recall (0.75), whereas all other models had both recall and precision of 0.81 – 0.86.
  + At the burrow scale, models with two or more input variables had similar F-scores (0.80 – 0.81). The single-input models, RGB and TPI, had F-scores of 0.66 and 0.78, respectively.
    - Likewise, predicted burrow count was highly correlated with digitized burrow counts at the tile scale (30 x 30 m) for all models, and was > 0.91 for models with two or more inputs (Fig. 3). Bias, as measured by mean percent error (MPE) was lowest for the two-input RGB + TPI model (MPE = 0.17), due mostly to improved performance at low burrow counts (Fig. 3).
* **Ground validation**
  + Across all tiles with available ground data (32 tiles) we had ground observations of 282 burrows and 108 non-burrow (but ‘burrow-like’) features.
  + The recall rate for ground observed burrows ranged from 52 – 73%.
    - In general, missed burrows tended to be smaller, shorter, less active and more vegetated (Fig. 3). The RGB and RGB+TPI model tended to miss burrows that were significantly less active than average, whereas the TPI-only model and the more complex models (3 or more inputs) tended to be more likely to miss active burrows.
    - If we looked only at burrows with observable activity (n=218), recall tended to improve by about 3 – 6 % across all models, with the top recall achieving 77% (Table 3). Precision only decreased by about 1%, resulting in an overall increase in F-score of 0.01 – 0.05 when looking only at active burrows.
  + The commission error rate (false positives) tended to be highest for badger dens and digging holes, though the number of ground observations for each was low (Table 4). Commission error for old, collapsed prairie dog burrows was highest for the TPI-only model (0.30) and higher for models that included elevation derived inputs (0.11 – 0.19) compared to the RGB-only model (0.04). In general, all models had low commission error for ant mounds, though it was slightly higher for the RGB-only model (0.06) and the RGB+TPI+Hillshade model (0.09).
  + Of the 32 tiles with ground data, 13 tiles were digitized and not used in the training
    - For the 13 digitized tiles with available ground data, digitization accuracy was (Table 5):
      * Precision = 81%
      * Recall = 74%
      * F-score = 0.77
    - For the same 13 tiles, prediction accuracy assessed with ground data was similar (Table 5), though in some cases recall was actually better with predicted data compared to the digitized data (up to 78%), though at the expense of slightly lower precision.

**Tables**

**Table 1.** Pixel-scale test validation results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **inputs** | **N inputs** | **Accuracy** | **Precision** | **Recall** | **F-score** |
| rgb | 1 | 0.98 | 0.83 | 0.75 | 0.68 |
| tpi | 1 | 0.99 | 0.84 | 0.81 | 0.77 |
| rgb\_tpi | 2 | 0.99 | 0.83 | 0.83 | 0.76 |
| rgb\_tpi\_shade | 3 | 0.99 | 0.82 | 0.84 | 0.77 |
| rgb\_tpi\_shade\_ndvi | 4 | 0.99 | 0.85 | 0.83 | 0.78 |
| rgb\_tpi\_shade\_ndvi\_dsm | 5 | 0.99 | 0.86 | 0.83 | 0.78 |

**Table 2.** Burrow-scale test validation results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **inputs** | **N inputs** | **Precision** | **Recall** | **F-score** |
| rgb | 1 | 0.70 | 0.62 | 0.66 |
| tpi | 1 | 0.79 | 0.77 | 0.78 |
| rgb\_tpi | 2 | 0.82 | 0.81 | 0.81 |
| rgb\_tpi\_shade | 3 | 0.75 | 0.86 | 0.80 |
| rgb\_tpi\_shade\_ndvi | 4 | 0.78 | 0.85 | 0.81 |
| rgb\_tpi\_shade\_ndvi\_dsm | 5 | 0.78 | 0.82 | 0.80 |

**Table 3.** Burrow-scale validation with ground data for 32 tiles with available ground data using all burrows (n=282) and using only burrows with ground-observed activity (n=218)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | ***---- all burrows ----*** | | | ***---- active burrows only ----*** | | |
| **inputs** | **N inputs** | **Precision** | **Recall** | **F-score** | **Precision** | **Recall** | **F-score** |
| rgb | 1 | 0.94 | 0.52 | 0.67 | 0.93 | 0.59 | 0.72 |
| tpi | 1 | 0.92 | 0.65 | 0.77 | 0.91 | 0.68 | 0.78 |
| rgb\_tpi | 2 | 0.94 | 0.65 | 0.77 | 0.93 | 0.71 | 0.80 |
| rgb\_tpi\_shade | 3 | 0.93 | 0.73 | 0.82 | 0.91 | 0.77 | 0.84 |
| rgb\_tpi\_shade\_ndvi | 4 | 0.94 | 0.70 | 0.80 | 0.93 | 0.74 | 0.82 |
| rgb\_tpi\_shade\_ndvi\_dsm | 5 | 0.94 | 0.68 | 0.79 | 0.92 | 0.73 | 0.82 |

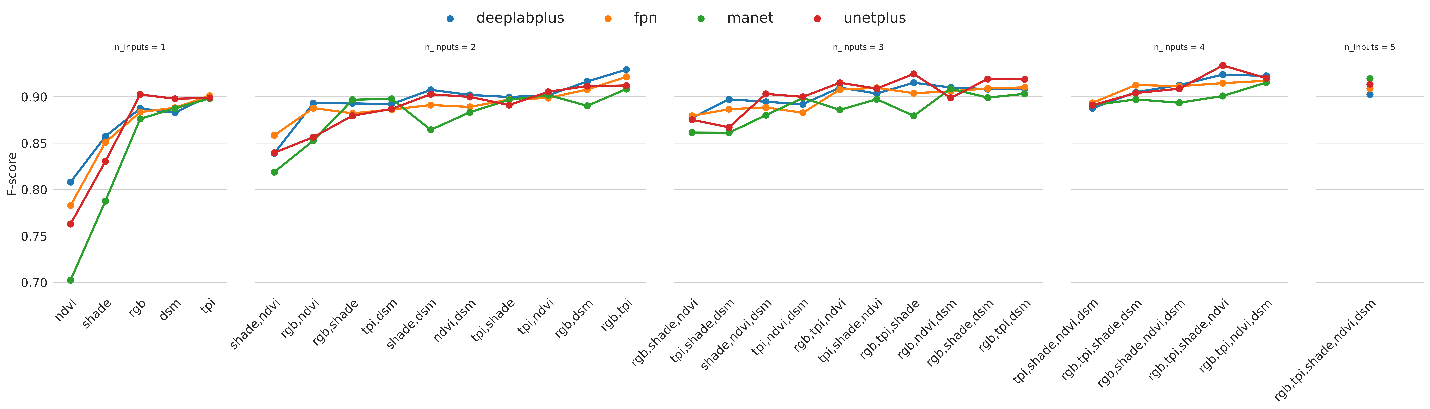
**Table 4.** Commission error (false-positive) rates for non-burrow features for each model and feature type.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature type** | **n** | **rgb** | **tpi** | **rgb, tpi** | **rgb, tpi, shade** | **rgb, tpi, shade, ndvi** | **rgb, tpi, shade, ndvi, dsm** |
| Anthill | 67 | 0.06 | 0.03 | 0.04 | 0.09 | 0.04 | 0.04 |
| Den | 8 | 0.38 | 0.50 | 0.38 | 0.38 | 0.38 | 0.38 |
| Digging | 6 | 0.17 | 0.17 | 0.17 | 0.50 | 0.33 | 0.33 |
| Old burrow | 27 | 0.04 | 0.30 | 0.15 | 0.11 | 0.15 | 0.19 |
| *Overall* | *108* | *0.06* | *0.08* | *0.06* | *0.07* | *0.06* | *0.06* |

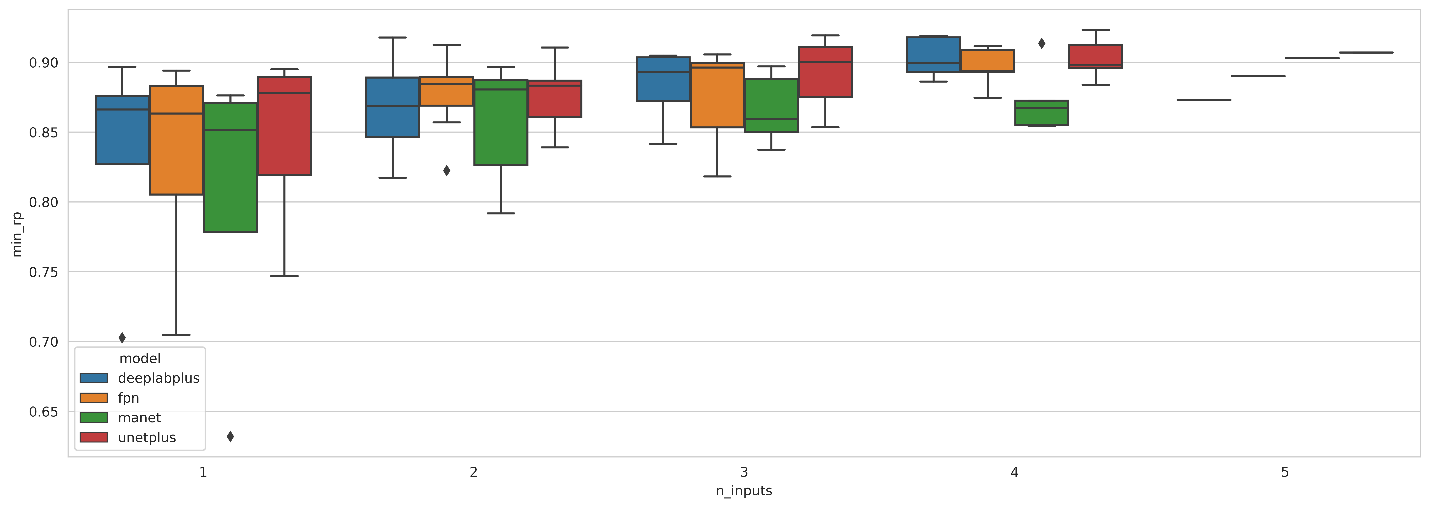
**Table 5.** Burrow-scale validation with ground data for 13 tiles that had both ground data and digitized data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **inputs** | **N inputs** | **Precision** | **Recall** | **F-score** |
| Digitized | NA | 0.81 | 0.74 | 0.77 |
| rgb | 1 | 0.74 | 0.64 | 0.69 |
| tpi | 1 | 0.79 | 0.76 | 0.77 |
| rgb\_tpi | 2 | 0.78 | 0.70 | 0.74 |
| rgb\_tpi\_shade | 3 | 0.72 | 0.78 | 0.75 |
| rgb\_tpi\_shade\_ndvi | 4 | 0.75 | 0.77 | 0.76 |
| rgb\_tpi\_shade\_ndvi\_dsm | 5 | 0.70 | 0.76 | 0.73 |

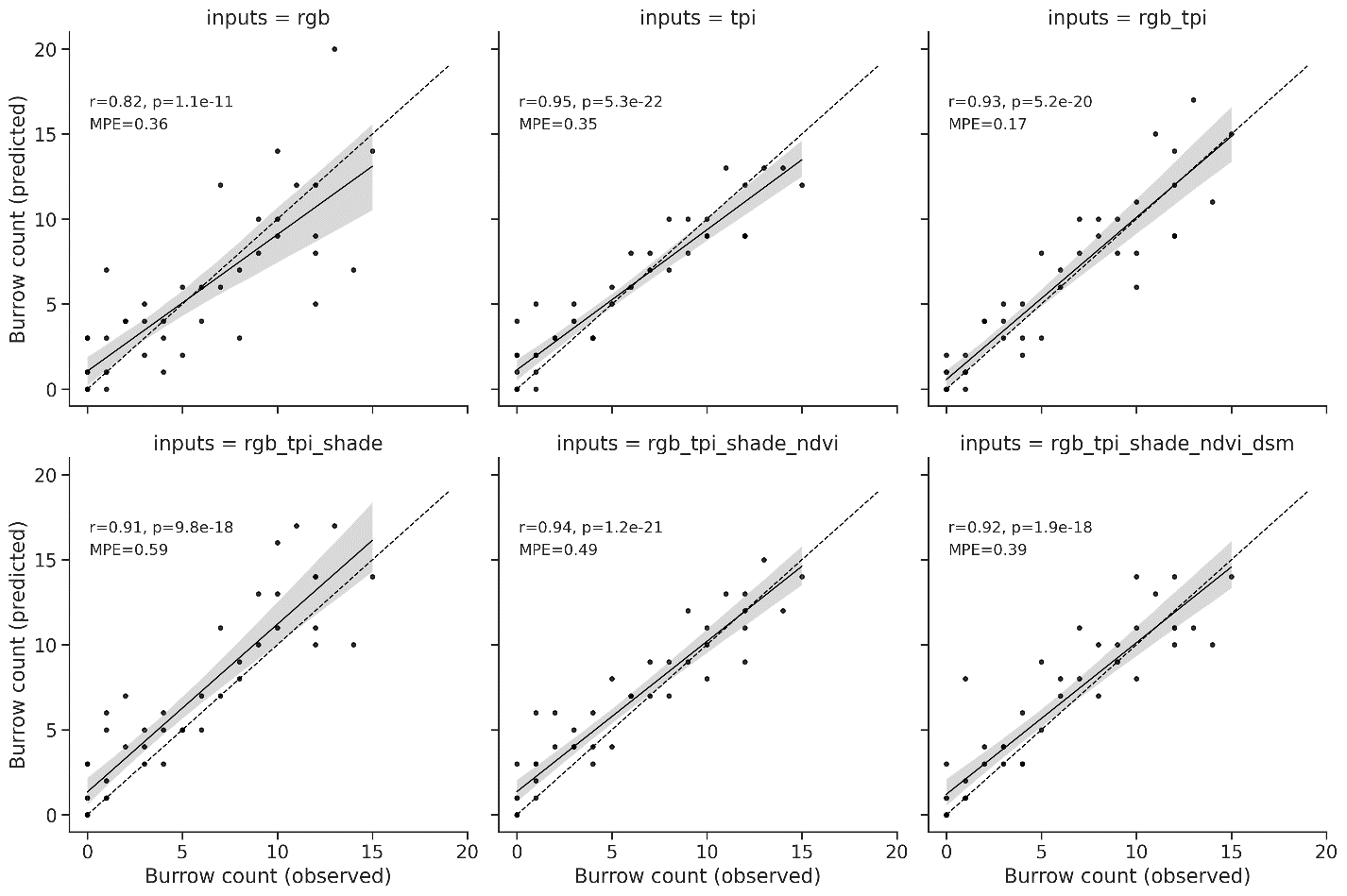
**Figures**



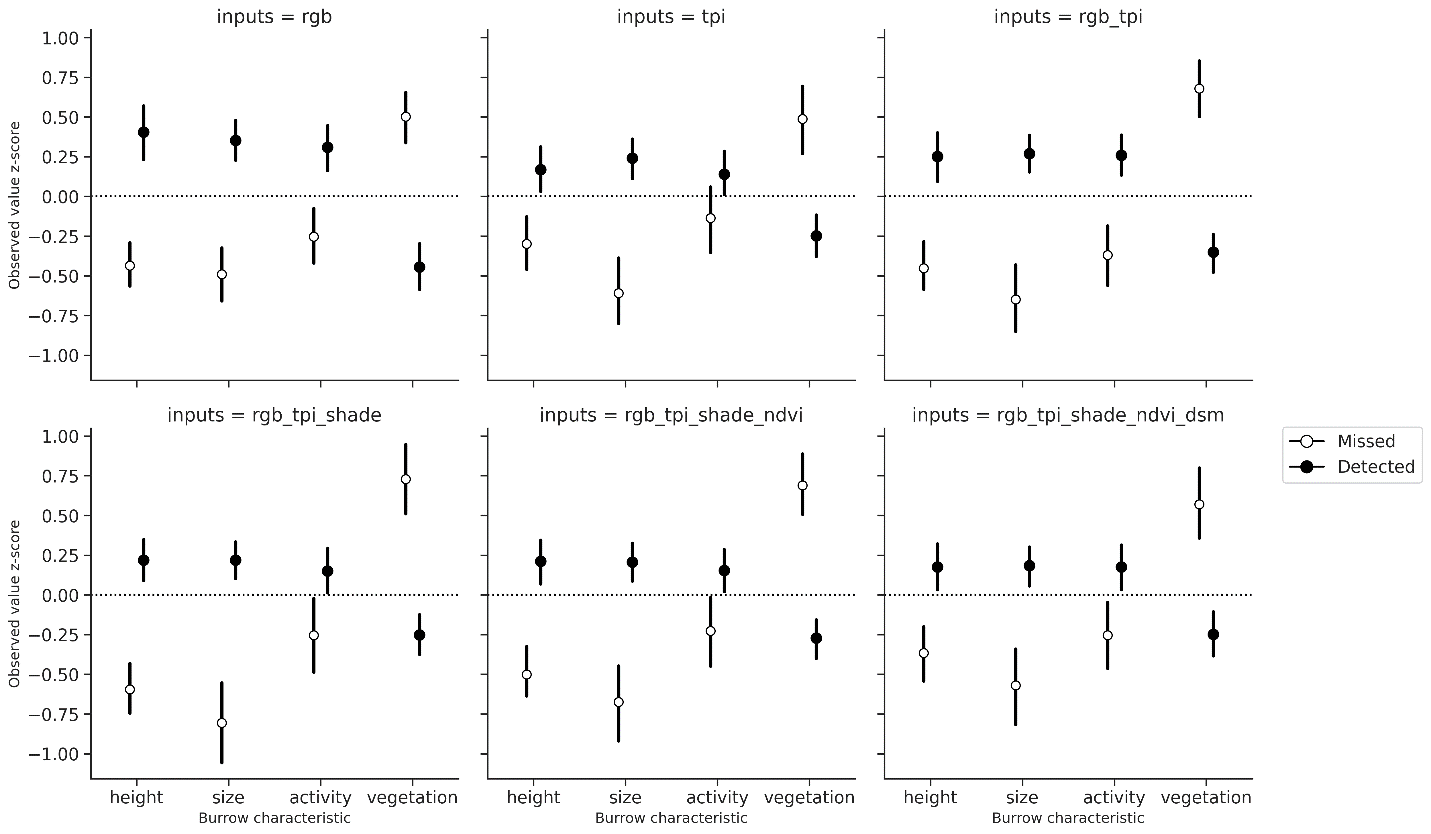
**Fig. 1.** F-score for each model architecture and combination of model inputs



**Fig. 2.** Distribution of the minimum of Recall/Precision for each model architecture across each of the number of model inputs.



**Fig. 3.** Correlation between predicted and observed burrow counts at the scale of individual testing tiles (30 x 30 m). ‘r’ is the Pearson correlation coefficient, ‘p’ is the alpha (significance) of the correlation and ‘MPE’ is the mean percent error of predictions, an indication of bias.



**Fig. 4.**