Eragrostis tef:

Predicting the potential extent of *Eragrostis tef* under future environmental conditions using distribution modelling

Jack Smith s2269872

Introduction:

Eragrostis tef, also known as teff, is a species of lovegrass native to the Horn of Africa. It is known to have been cultivated in Ethiopia and Eritrea for over five thousand years, where the teff seeds are harvested for use as the primary ingredient of injera (a fermented flat pancake-like bread). Teff is a regional staple, feeding over 100 million people (Kassier, 2002). However, cultivation in other regions around the globe has been increasing over the last 25 years, initially in United States and South Africa, but today numerous other countries are cultivating teff, including: Australia, India, Germany, The Netherlands, Belgium and Spain. A National Research Council (1996) report also predicted that teff has the potential to be successfully grown in many other countries across the globe (including: Mexico, Bolivia, Peru, Ecuador, Pakistan and Nepal) based on the potential teff has displayed in adapting to a broad range of climates.

Successful cultivation of *E. tef* across the globe beyond the specie's native range, has the potential to exceed the cultivative success achieved in Ethiopia or Eritrea, due in part to the traditional agricultural methods used in these regions, and the lack of funding that has hitherto been committed to researching better agronomical practices for this species (Ketema, 1993; Ghebrehiwot et al., 2008). It has been suggested that the main factors effecting yield loss of *E. tef* in Ethiopia are drought, water logging, heat and frost (Tefera, 2002). Others have more recently linked the success of *E. tef* farming in the Netherlands to day length (Delden et al., 2012). Most literature seems to concur that the distribution of teff today is incredibly limited, when compared to the species' range of bioclimatic requirements (Ghebrehiwot et al., 2008). This could suggest that records of the present distribution of the species are likely an inadequate predictor of the species' potential distribution.

In this investigation we aim to model the distribution potential of *E. tef* in the future under both rcp45 and rcp85 climatic predictions. With the null hypothesis that the distribution will be unchanged from that seem today, based on GBIF occurrence data.

Methods:

Occurrence data for *Eragrostis tef* were downloaded from the GBIF.org (2019), filters were set to include only human observations and preserved specimens, that included geographic coordinates. This yielded 470 records from 45 published datasets.

These records were then imported into ArcGis, where the coordinates were plotted onto a world map in a WGS84 projection (Figure 1). This revealed three data points that were in the sea (off the coasts of New Caledonia, Durban and Cape Town) these were removed from the dataset. All other points were left in.

Environmental data was collected from the WorldClim (version 1.4) online database (worldclim.org) (Hijmans et al., 2005). Full sets of Bioclimatic variables were taken for both present and future condition, at 5 arc-minute resolutions in both cases. The Hadley Global Environment Model 2 - Earth System (HadGEM2-ES) (Bellouin et al., 2007) was used as the source of future condition predictions, under the representative concentration pathways 4.5 and 8.5 (rcp45 and rcp85) forecasts, again for all bioclimatic variables.

This data was input into R, where it was analysed using the sp, rgdal, raster, biomod2 and usdm packages. The extent of the occurrence data (WGS84 coordinates) was converted into a shapefile, to train the distribution model on a select raster area, focussed on where the *E. tef* observations occur (Figure 2). However, given that the occurrences of *E. tef* retrieved from GBIF were widely (yet irregularly) distributed, the area of interests (AOI) only cropped a small portion of the WGS84 global view (Figure 3).

The 19 bioclimatic variables were refined, to avoid environmental autocorrelation, using a sequence of: literature (see introduction) to suggest the most likely variables in influencing the success of *E. tef*; followed by a Pearson's pairwise correlation statistic, with a 0.7/-0.7 threshold of variable association/dependence; and finally a variance inflation factor (VIF) statistic, with a dependence threshold of 10. The shortlisted bioclimatic variables selected in this manner were: Bio2, Bio5, Bio8, Bio14, Bio15 and Bio19. The bioclimatic data used in these statistics was that of present climatic conditions, across the *E. tef* AOI. For the Pearson statistic the r^2 correlations between every pair of variables were generated in R before being exported to Microsoft Excel for manipulation. The VIF test for multicollinearity was implemented in R using the usdm package.

Species distribution models (SDM) were generated using Maxent (Appendix 1), where the *E. tef* occurrence distribution points (GBIF) were input along with present environmental data (WorldClim) for the six bioclimatic variables of interest (Bio2, Bio5, Bio8, Bio14, Bio15 and Bio19). The parameters used in the program are documented in Appendix 1. Maxent *E. tef* distribution models were collected for both the future rcp45 and rcp85 predictions based on these climatic variables. Models were trained on five replicate runs.

Model Output Results:

Close inspection of the *E. tef* data distribution on a (WGS84 projection) world map (Figure 1) reveals three points that were in the sea, which were removed. There are also clear concentration in South Africa, the Horn of Africa and Central Europe. Of these locations the Horn of Africa displays the fewest points, despite being the within the plant's native range. Many of the points on the map appear in isolation, such as one in England, Scotland, Wales, Japan, Argentina, Senegal and several other isolated locations, many of which are also housed close to their respective capital cities, and are therefore more likely specimens housed in collections than observations of wild occurrences.

Figure 1: Global WGS84 projection of 470 occurrences of *E. tef*.
Constructed in ArcMaps, using GBIF occurrence data.

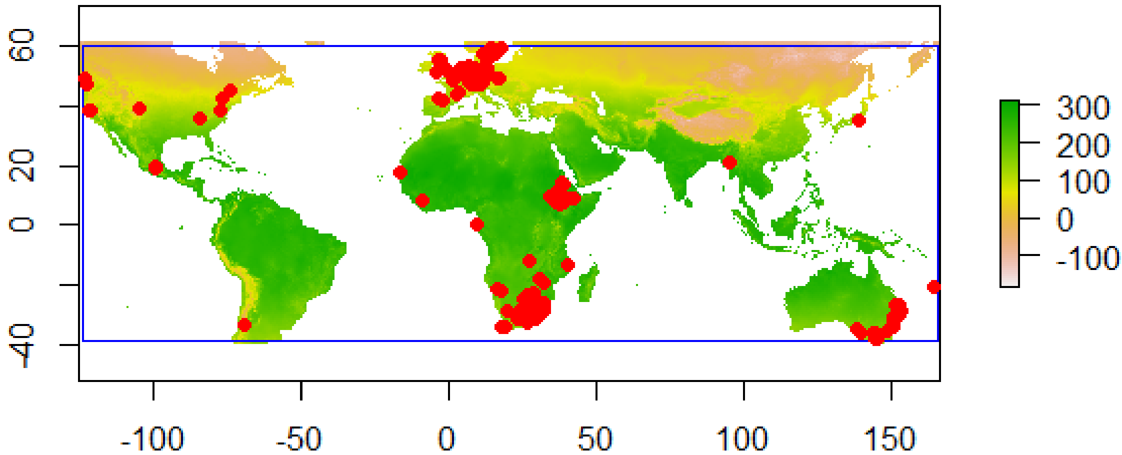

Figure 2: Global distribution of *E. tef* occurrence data points (red), showing the area outside of the shapefile (blue) where coordinates are to be cropped from the area of interest raster.
Created in R, using GBIF occurrence data.

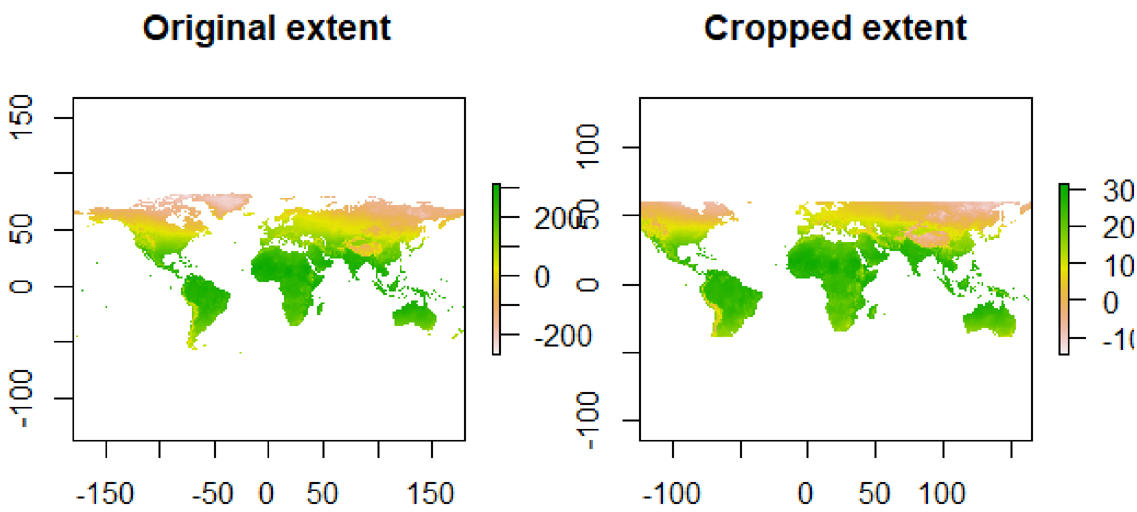

Figure 3: Comparison of the original extent and cropped extent of the raster containing all *E. tef* coordinates, on a global (WGS84 projection) map.
Created in R.

The point shapefile created from the distribution of *E. tef* (Figure 2) only cropped a small margin from the edge of the full WGS84 projection (Figure 3), due to the global extent of its distribution.

Selecting Bioclimatic Variables

In selecting bioclimatic variables literature was first reviewed, to suggest those variables likely to be of most influence on the success of *E. tef*, this is summarised in the introduction. This was followed by Pearson's pairwise correlation statistical analysis. Table 1 shows the r^2 values for the bioclimatic variables that were selected as having the least correlation. These variables were selected from the full 19 variables, as having the fewest pairwise correlation scores outside of the threshold range (-0.7 to 0.7).

Table 1: Pearson's pairwise correlation statistical analysis showing the bioclimatic variables with the lowest r^2 correlation values. All values fall within the -0.7 to 07 range.

	Bio02	Bio05	Bio08	Bio14	Bio15	Bio19
Bio02	1	0.444264	0.128405	-0.46274	0.453563	-0.38522
Bio05	0.444264	1	0.643395	-0.18131	0.34698	-0.01296
Bio08	0.128405	0.643395	1	0.003467	0.360996	0.068094
Bio14	-0.46274	-0.18131	0.003467	1	-0.53516	0.650407
Bio15	0.453563	0.34698	0.360996	-0.53516	1	-0.30519
Bio19	-0.38522	-0.01296	0.068094	0.650407	-0.30519	1

As seen in table 1, the greatest correlation occurs between Bio14 and Bio19, and between Bio5 and Bio8, however this is not beyond the (0.7) threshold, and should therefore not cause any issue. Analysis of the multicollinearity of these six bioclimatic variables using the VIF statistic confirmed their independence, with VIF scores well below the threshold of 10 (Table 2).

Table 2: Variance inflation factor (VIF) values for the dependence of each of the six selected bioclimatic variables in relation to the other five. With descriptions of each bioclimatic variable.

	Variables	**Variable description**	**VIF**
1	Bio02	Range (Mean of monthly (max temp - min temp))	1.778870
2	Bio05	Max Temperature of Warmest Month	2.313029
3	Bio08	Mean Temperature of Wettest Quarter	2.099883
4	Bio14	Precipitation of Driest Month	2.391017
5	Bio15	Precipitation Seasonality (Coefficient of Variation)	1.846960
6	Bio19	Precipitation of Coldest Quarter	1.861895

Species Distribution Models (SDMs)

SDMs of the *E. tef* revealed predicted that at preset (Figure 4) the distribution of *E. tef* could be spread across much of North and South America, Europe, East Asia and Southern Africa. Whereas future conditions seem to allow for an increased concentration within Europe and North Eastern USA, South Eastern Canada and Northern Asia (especially Russia) but a substantially reduced range in Africa and Southern America. What is perhaps also surprising is the similarity between the rpc45 (Figure 5) and rpc85 (Figure 6) distributions, there is little difference, especially when the two are contrasted to the present distribution, implying that the range reduction is apparently sensitive to change in much of the distribution where *E. tef* is predicted to be able to live in present conditions, but should be quite stable in many places, where it is predicted to thrive similarly under present, rpc45 and rpc85 conditions.

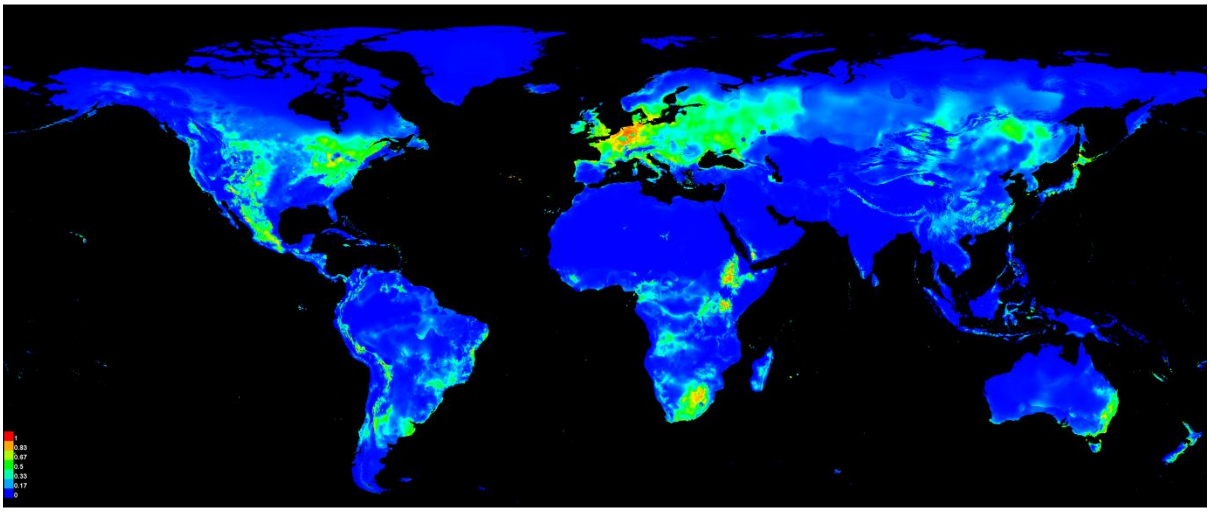

Figure 4: Point-wise data for the distribution of E. tef based on present world data. Taken from the mean of the 5 model repeats.

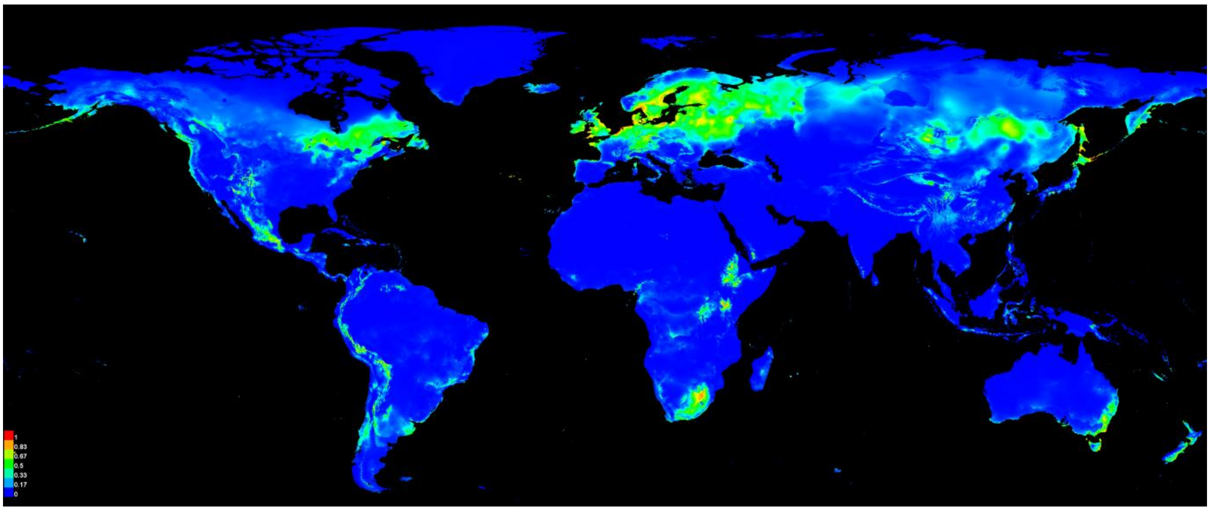

Figure 5: Point-wise data for the distribution of *E. tef* based on future rcp45 world data. Taken from the mean of the 5 model repeats.

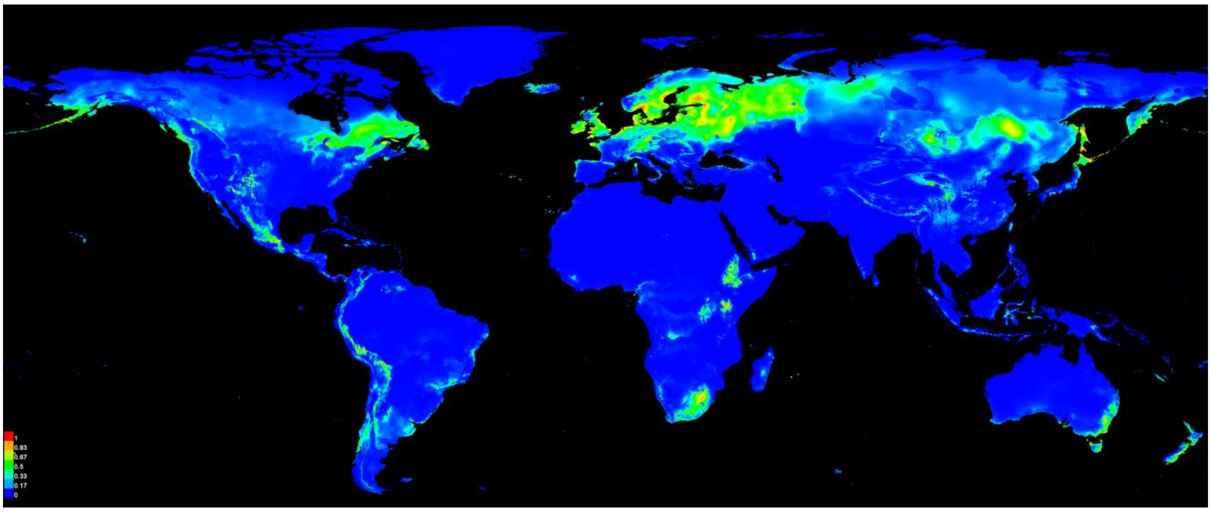

Figure 6: Point-wise data for the distribution of *E. tef* based on future rcp85 world data. Taken from the mean of the 5 model repeats.

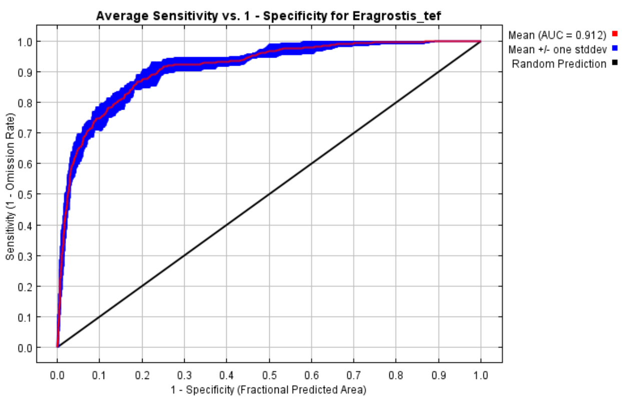

Figure 7: Receiver Operator Characteristic (ROC) curve, showing the AUC to be 0.912.

The Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) curve (Figure 7) was 0.912 (with a standard deviation of 0.009), which suggests the species distribution model is reliable, as is well above the (AUC>0.7) threshold.

Table 3: Analysis of variable contribution for each bioclimatic variable to the SDM.

Variable	Percent contribution	Permutation importance
_Bio05	25.6	45.4
_Bio14	25.6	9.7
_Bio02	17.3	17.3
_Bio08	12.9	7.9
_Bio15	11.6	7.2
_Bio19	7	12.4

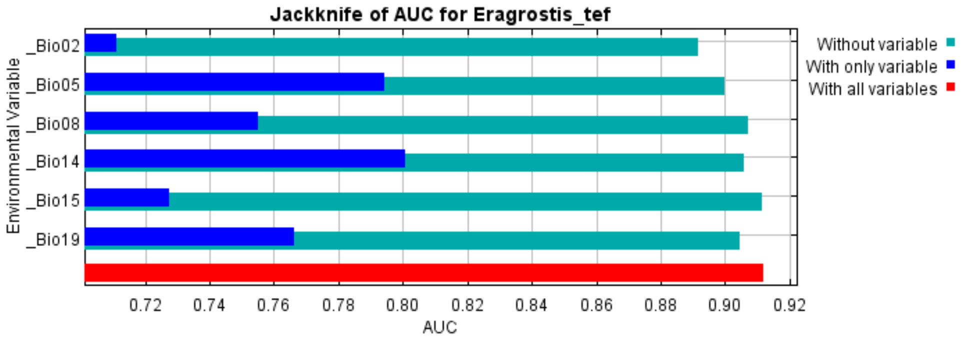

Figure 8: Jackknife scores based on the AUC test data, showing the relative impact of each environmental variable

The relative contributions according to the analysis of each bioclimatic variable upon the maxent model, suggests that the Bio05, Bio14 and Bio02 contributed most significantly. Jackknife testing, on the AUC test data revels that both the Bio02 and Bio15 variables explained the least when considered alone, while Bio05 and Bio14 explained the distribution substantially greater. However, when excluded, Bio02 decreases the AUC the most, indicating that it contains the most unique information, not contained within the other variables.

The map of range change (figure 9) shows a clear shift in available range away from the equator, this is an incredibly extreme prediction, especially given the native range of the *E. tef* species. It seems that such extreme predictions may be a result of under representation on samples around the equator, which produce a model trained on present day distribution data taken primarily from South Africa, The Netherland, Belgium, Germany and Eastern Australia. The suggestion that much of the Horn of Africa itself, is shown as “never suitable” (grey) is clearly inaccurate, given the documented native range of the species in this area.

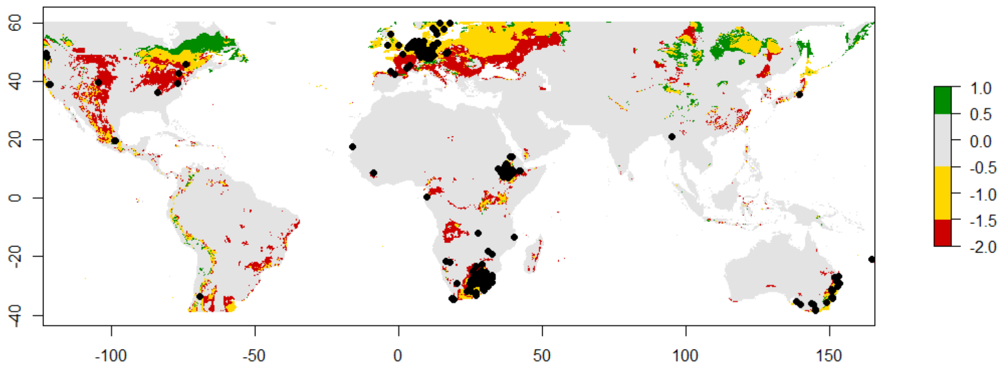

Figure 9: Change in the suitability of areas between present and future rcp45 climate conditions. Green = gained suitability, Red = lost suitability, yellow = remains suitable, Grey = never suitable.

Biological interpretations:

Of the original 470 records, three were removed due to their coordinates showing them to be in the sea, thus implying incorrect coordinate input upon collection. All other points were left in, including those, such as that seen in Japan (Figure 1) that exist beyond the current documented range of *E. tef*. This is because the record in not necessarily incorrect, it is entirely possible that the occurrences are genuine. There are also very few record in the Americas, despite cultivation and farming of *E. tef* in several US states since the '90s (Ghebrehiwot et al., 2008). The analyses were carried out on the data available. However, it should be noted that the irregularities in the occurrence distribution measured are at odds with expectations from the literature, and the final distribution prediction could potentially be improved with a more accurate occurrence distribution data. Further there is a very strong bias towards presence in the central Europe, simply because more samples have been submitted there, representation in Ethiopia is relatively underrepresented, while not a single observation is recorded for Eritrea or Egypt, despite evidence of *E. tef* being cultivation in these countries more than 5000 years ago (Ketema, 1986).

Values collected for the Pearson's correlation statistic, all of which were within our -0.7 to 0.7 range, and variance inflation factor (VIF), all of which were well below even a conservative threshold of dependence (5), suggested that there was little correlation between the influence of these bioclimatic variables on *E. tef*'s distribution. This should limit any bias on the outcome of the Maxent model, based on stronger selection on any single climatic condition.

The jackknife results showed that *Max Temperature of Warmest Month* and *Precipitation of Driest Month* (bio05 and bio14 respectively) explained the distributions most effectively, while *Range (Mean of monthly (max temp - min temp))* and *Precipitation Seasonality (Coefficient of Variation)* (bio02 and bio15 respectively) explained the least. This suggests that extremes of high temperature and drought are more influential towards the distribution of *E. tef* than period averages. Furthermore, extremes of wet and cold (bio08 and bio19) seem to be less influential than extremes of drought or heat.

The AUC of the ROC curve (Figure 7) was 0.912 suggesting the species distribution model to be reliable. Although this value is very strong (well above the threshold of AUC>0.7), the maps generated for the distribution of *E. tef* (Figures 4-6) are still limited by the distribution data for *E. tef* taken from the GBIF database. This data shows clear underrepresentation of areas such as Eritrea, where there are numerous literary records of the distribution teff, but none appearing in GBIF. Consequently the predictions, although well supported in the computational stages, are perhaps a poor representation of the potential future range of *E. tef*, as they are derived from a very incomplete picture of the contemporary distribution of the species. Nevertheless, these models can be useful, especially with regards to predicting the most influential bioclimatic variables upon teff distribution. The fact that extremes may disproportionally influence the distribution potential of *E.tef* may well result in a substantial reduction in the suitability of its native range, around the Horn of Africa and close to the equator, as this is where extreme drought and temperature fluctuations are likely to manifest in the coming decades. The prediction that teff may become a far more prevalent crop in the temperate zones, particularly in Northern America, Europe and Asia, may yet prove accurate.

References:

Bellouin, N., Boucher, O., Haywood, J., Johnson, C., Jones, A., Rae, J. and Woodward, S. (2007) Improved representation of aerosols for HadGEM2. Meteorological Office Hadley Centre, Technical Note 73.

Delden, S.H. van, Vos, J., Stomph, T.J., Brouwer, G. and Struik, P.C. (2012) Photoperiodism in Eragrostis tef: Analysis of ontogeny and morphology in response to photoperiod. Euro. J. Agron. **37**: 105- 114.

GBIF.org (06 December 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.xy3dm6

Ghebrehiwot, H. M., Kulkarni, M. G., Kirkman, K. P. and van Staden, J. (2008) Smoke-Water and a Smoke-Isolated Butenolide Improve Germination and Seedling Vigour of Eragrostis tef (Zucc.) Trotter under High Temperature and Low Osmotic Potential. J. Agron. Crop Sci. **194(4)**, 270-277.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. and Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology. **25**: 1965-1978.

Kassier, S. B. (2002) Comparative Responses of Fodder and Grain teff (Eragrostis tef (Zucc.) Trotter) Cultivars to Spatial, Temporal and Nutritional Management. MSc Thesis, University of Natal, Pietermaritzburg, South Africa.

Ketema, S. (1986) Production trends, germplasm resources, breeding and varietal improvement of small millets, with special emphasis on teff in Ethiopia. In: Seetharam A, Riley K & Harinarayana G (eds.) Small Millets in Global Agriculture, Oxford & IBH Publishing Co. Pvt . Ltd . New Delhi, pp 309 - 314.

Ketema, S. (1993) Tef (*Eragrostis tef*): Breeding, Genetic Resources, Agronomy, Utilization and Role in Ethiopian Agriculture. Institute of Agricultural Research, Addis Ababa, Ethiopia.

National Research Council (1996) Lost Crops of Africa, Vol. I, Grains. 215–235. National Academy Press, Washington, DC.

Tefera, T. (2002) Allelopathic effects of Parthenium hysterophorus extracts on seed germination and seedling growth of *Eragrostis tef*. J. Agron. Crop Sci. **188**, 306-310.

Appendix:

Appendix Figure 1: Maxent input parameters, used in generating the *E. tef* species distribution models (SDMs)