Quantifying the uncertainty in spatially-explicit land-use model predictions arising from the use of substituted climate data

M. Rivington, K.B. Matthews, and K. Buchan.
Macaulay Institute, Craigiebuckler, Aberdeen, AB15 8QH, Scotland.
m.rivington@macaulay.ac.uk

Abstract: There can often exist a significant gap between the sophistication of land use systems models and our ability to provide the required biophysical input data. This can be particularly significant when systems models are used as components for decision support. The lack of spatially-explicit input parameters means that models' site-specific predictions have potentially large uncertainties that are frequently unquantified. For crop production models, solar radiation (SR) is a significant parameter that is often sparsely measured. In the absence of site-specific data, modellers often substitute data from other sources: the nearest meteorological station; data derived from measurements made on the site (e.g. solar-radiation interpolated from temperature and rainfall); or data derived from weather generators. This paper investigates the impact on a land use model of substituting on-site measured data with that from sites at increasing distances. This analysis quantifies the changes in both data bias and model variability introduced by the process of data substitution and forms a baseline against which it is possible to evaluate the alternative methods of data provision. To measure the relationship of changing uncertainty with distance to a data source, a database of observed climate data was created for 24 locations in the U.K. The root mean square error (RMSE) between each location's SR data was calculated for each day of the year for corresponding years. The CropSyst crop production model was used to calculate yield, for a standardised spring barley scenario, for each year of available climate data for the 24 locations. The RMSE and each site's yield estimates were then compared to determine the impact. The results show that there can be a complex relationship between the rate of decay in data similarity and impact on model output, governed by factors such as the density of the network of measurement locations and topography. Fundamentally, however, the results show that data substitution methods have a profound impact on the reliability of model results.

Keywords: climate data; land use systems; crop models; solar radiation; CropSyst.

1. INTRODUCTION

There is an increasing demand for the site-specific application of simulation models and decision support systems. This places an emphasis on being able to provide appropriate spatially and temporally representative biophysical data. There is a serious limit on the application of agricultural, hydrological and ecosystem models if such data is not directly available (Hoogenboom 2000). One solution is to identify a suitable data substitute, and to quantify the impact on the quality of the model output. Daily climate represents one of the more commonly used data types in models of biophysical processes. There is a wide range in the availability and site-specific representation of daily climate data. In Britain, as elsewhere, daily solar radiation (SR) data from meteorological stations is considerably more restricted than for precipitation and temperature. Solar radiation is the key data type for photosynthesis, plant and soil evaporative processes and surface energy budgets, in conjunction with temperature. In the absence of site-specific climate data, substitute data from nearby meteorological stations, derived from interpolation methods or weather generator models are often used. When using alternative meteorological stations, clearly problems can arise when trying to synthesise temporally and spatially synchronised data to provide a single data set (i.e. precipitation, temperature and SR).

This paper investigates the consequential impacts of using substitute climate data from alternative meteorological stations on the output of the CropSyst crop production model (Stöckle and Nelson 1998). Identifying the influence of a single data type becomes difficult when a range of climate data types are used as inputs in synchronisation. This is particularly true within a model representing a range of non-linear processes. Our aim was to illustrate the prediction uncertainty that using substitute data introduces to a range of assessment metrics. In a previous study substitute data from the nearest meteorological station did not necessarily provide the best replacement (Rivington *et al* 2002). Hunt *et al*

(1998) working in Ontario, Canada, determined a threshold distance for substitution of 390km. Beyond this distance they recommended the use of an interpolation method to provide missing SR data. We investigated the decay in similarity of SR between sites and the change over distance. From this we hoped to determine a baseline of SR similarity decay with distance, against which other forms of data substitution (interpolation, weather generators) can be compared. A practitioner can then determine their own acceptable distance threshold beyond which they will not use observed substitute SR data, but an alternative instead. An investigation was then made to see whether there was a correlation between SR data representation decay with distance and the differences in a range of metrics derived from the CropSyst model output.

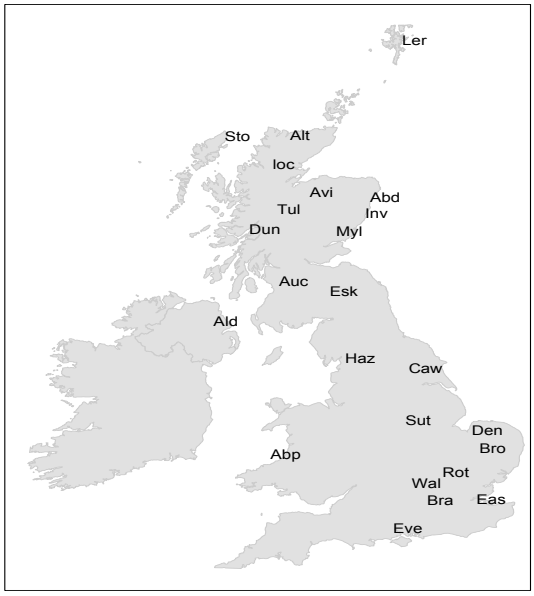

Figure 1. The distribution of sites in the UK

MATERIALS AND METHODS

2.1 Dissimilarity in solar radiation

A database of daily climate data (precipitation, max and min temperature and SR) was created for 24 locations in the UK (Figure 1). The overall period of data coverage was from 1964 to 1999, with each location having different lengths of data record. Original data was provided by the Meteorological Office via the British Atmospheric Data Centre (BADC). These sites were selected as they were the only ones to have all four data types for more than five years. Original data contained missing values, where years containing >30 consecutive days of a single data type were excluded. A replacement strategy was developed for missing data (one or several data types), using a data base search and optimisation method. Comparisons of observed solar radiation ($MJ/m^2/day$) data similarity were made between a site and remaining 23 sites within the data base by calculating the Root Mean Square Error (RMSE):

$$\sqrt{\sum \frac{(O - E)^2}{n}} \qquad \text{Eq. 1}$$

where O is the observed SR for the day of year (1 to 365) at a site and E is the observed SR for the same day at another site, n is the number of days in the year. This process was applied for each site, i.e. O for a fixed site, E for the nearest site, then second nearest, third nearest etc, for all sites in the data base. This produced a matrix data set from which it was possible to derive the mean, maximum and minimum RMSE for each site. The mean RMSE was then plotted for each site to produce regressions (Figure 2 – selected sites).

2.2 Impacts on CropSyst output

A standardised spring barley CropSyst simulation was run for each year of available climate data (the only variables) for all 24 sites. The yield estimates were then compared between a site and the three nearest sites for all corresponding years of available climate data. The metrics used to compare yield estimates were: difference in mean yield / n (t/ha); probability of means being equal (2 sample t-test); difference in total yield / n (t/ha); absolute difference / n (t/ha) (sum of over- and under-estimates); and size of maximum single error (t/ha), where n is the number of corresponding years. From the results, it was possible to identify the substitute site that produced the best fit for each metric.

2.3 Comparison of solar radiation and yield

The results from the two methods (2.1 and 2.2) were then compared by testing the correlation between yield difference from each site and its three nearest neighbours with the RMSE in SR dissimilarity for each year. The purpose was to identify a relationship between the differences in yield and dissimilarity in SR.

3. RESULTS

3.1 Solar radiation dissimilarity between sites

The comparisons of the RMSE between a single site and remaining 23 sites showed a range in the rates and patterns of increase in dissimilarity with distance. Sites in south-east Britain, i.e. Rothamstead (Rot) (Figure 2) had steeper rates of increase in dissimilarity than sites in the north or west, i.e. Aberdeen, Aberporth (Abp) and

Altnaharra (Alt). However the south-east sites were starting from lower RMSE values, attributable to the denser network of meteorological stations and greater geographical similarities. The south-east sites also had lower regression intercepts. There was a range of distances that provided the closest match. For 16 sites the nearest substitute provided the lowest mean RMSE. Conversely, i.e. at Cawood (Caw), the fourth nearest neighbour, Brooms Bar (Bro) (209km) gave the lowest mean RMSE. For Dunstaffnage (Dun), five other sites had lower mean RMSE's than the nearest site. Across all sites' comparisons, Denver (Den), the nearest neighbour to Brooms Barn (40km) had the lowest mean RMSE at 1.89 MJ/m^2/day. The lowest and highest intercept of the regressions were 2.59 at Sutton Bonington (Sut) and 5.01MJ/m^2/day a t Tulloch Bridge (Tul) respectively. The mean of all lowest RMSE values for all sites was 2.43 MJ/m^2/day. The lowest individual years' RMSE value was found between Bracknell (Bra) and Wallingford (Wal) (33km), at 1.47 MJ/m^2/day. These results show some spatial consistency, but the analysis has not considered site elevation or proximity to the sea. These two factors combined may explain some of the inconsistencies, i.e. at the west coast site of Dunstaffnage the nearest site is Tulloch Bridge, an inland mountain location. Dunstaffnage's most similar site is Auchencruive, also on the west coast. Comparisons between sites of similar geographical characteristics did not show the same degree of inconsistency.

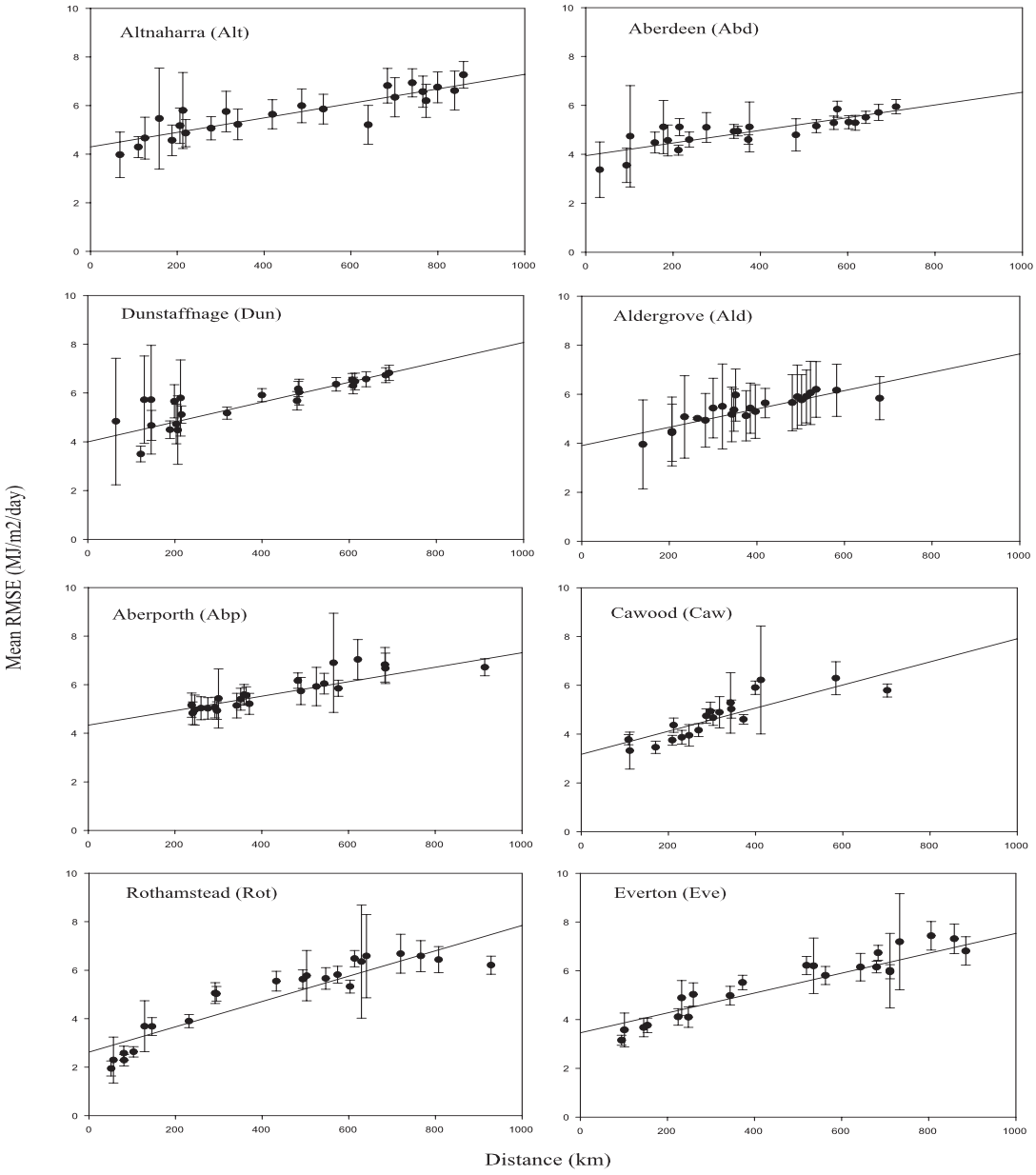

Figure 2 . Changes in mean RMSE (MJ/m2/day) with increasing distance (km)

Table 1. CropSyst yield (t/ha) estimate comparison between selected sites and three nearest substitutes. P value is the probability of means being equal. Shaded areas indicate best fitting value per metric. Correlation coefficient is for difference in yield per year (t/ha) and SR RMSE ($MJ/m^2/day$) (*P= 0.004)

Site	Subst	Dist km	Mean yield Diff / n	Total yield Diff / n	Absolute Diff / n	Max error	P value	correlation coeff	n
Dun	Auc	121	0.0084	0.1508	0.5677	-1.49	0.354	-0.354	19
	Avi	129	-0.0873	-0.9028	0.9028	-1.58	0.001	0.213	10
Loc	Alt	68	0.0077	0.0542	0.7548	2.67	0.160	0.371	7
	Avi	86	0.0499	0.1498	0.4130	0.84			3
	Tul	96	0.0984	0.4920	0.5543	1.14	0.225	-0.977*	5
Abd	Inv	32	0.0068	0.0272	0.1892	-0.33	0.941	-0.048	4
	Myl	94	0.0064	0.1475	0.5471	1.10	0.437	0.353	23
	Avi	102	-0.0792	-0.9503	1.0651	-5.20	0.060	-0.082	12
Inv	Abd	32	-0.0068	-0.0272	0.1892	-0.33	0.941	0.048	4
	Myl	65	0.0361	0.2164	0.5084	0.89	0.567	0.073	6
	Avi	102	-0.6264	-1.8791	1.8791	-5.53			3
Myl	Inv	65	-0.0361	-0.2164	0.5084	-0.89	0.567	-0.073	6
	Ab1	94	-0.0064	-0.1475	0.5471	-1.10	0.437	-0.353	23
	Avi	95	-0.0902	-1.0822	1.3483	-6.29	0.049	-0.118	12
Auc	Esk	88	-0.1081	-1.8408	1.1547	-2.97	0.000	0.417	19
	Dun	121	-0.0088	-0.1591	0.5992	1.49	0.354	0.354	18
	Ald	140	-0.0029	-0.0494	0.3619	1.38	0.824	0.417	17
Esk	Auc	88	0.1027	1.9894	1.9894	6.12	0.000	-0.064	20
	Myl	128	0.0810	1.7810	1.9671	6.29	0.001	-0.572	22
	Haz	147	0.1805	2.3465	2.3465	6.24	0.001	-0.497	13
Ald	Auc	140	-0.0022	0.2988	0.3418	-1.38	0.824	-0.417	18
	Esk	206	-0.0735	-2.3417	1.3298	-3.17	0.000	-0.392	18
	Dun	206	-0.0029	-0.0649	0.4134	-1.61	0.660	-0.129	22
Haz	Caw	109	-0.0904	-1.3566	1.3566	-3.27	0.000	0.390	17
	Esk	147	-0.1805	-2.3465	2.3465	-6.24	0.001	0.497	13
	Sut	166	-0.0700	-0.9796	1.0031	-2.09	0.000	0.402	14
Abp	Sut	238	-0.0304	-0.8199	1.0830	-2.51	0.000	-0.202	27
	Haz	241	-0.0017	-0.0273	0.8020	-1.55	0.891	-0.438	16
	Wal	246	-0.0554	-1.9776	1.5061	-3.01	0.000	-0.185	25
Caw	Haz	109	0.1025	1.5375	1.5375	3.27	0.000	-0.390	15
	Sut	111	0.0131	0.3274	0.6693	2.46	0.010	-0.062	27
	Den	171	0.0120	0.1919	0.3656	1.88	0.280	-0.143	18
Sut	Caw	111	-0.0131	-0.3274	0.6693	-2.46	0.010	0.062	27
	Den	114	-0.0129	-0.2069	0.7735	-2.32	0.280	-0.274	16
	Rot	129	-0.0106	-0.2746	0.8754	-3.36	0.321	0.282	26
Den	Bro	40	0.0112	0.1686	0.3539	0.95	0.000	0.065	15
	Rot	103	0.0138	0.1979	0.4032	1.21	0.432	-0.313	17
	Sut	114	0.0129	0.2069	0.7735	2.32	0.436	-0.285	16
Bro	Den	40	-0.0112	-0.1686	0.3539	-0.95	0.568	-0.065	15
	Rot	81	-0.0051	-0.0758	0.4294	-1.19	0.909	-0.067	15
	Eas	109	-0.0111	-0.1771	0.6697	-2.16	0.531	0.242	16
Rot	Bra	51	-0.0051	-0.1378	0.5043	3.31	0.615	0.150	27
	Wal	57	-0.0160	-0.6113	0.6472	2.45	0.182	-0.141	26
	Eas	81	0.0028	0.0776	0.5494	3.39	0.836	0.150	29
Wal	Bra	33	0.0110	0.2738	0.4468	1.38	0.414	0.105	25
	Rot	57	0.0144	0.3840	0.6232	-2.45	0.182	0.141	27
	Eve	101	-0.0370	0.9211	0.9868	2.00	0.004	-0.322	27
Eas	Rot	81	-0.0062	-0.0776	0.5494	-3.39	0.836	-0.150	29
	Bra	84	-0.0068	-0.1842	0.4617	-1.37	0.522	-0.057	27
	Bro	109	0.0111	0.1771	0.6697	2.16	0.531	-0.242	16
Bra	Wal	33	-0.0092	-0.2736	0.4468	-1.38	0.414	-0.105	25
	Rot	51	0.0056	0.1378	0.5043	-3.31	0.615	0.131	27
	Eas	84	0.0068	0.1842	0.4617	1.37	0.522	0.057	27
Eve	Brac	95	-0.0282	-0.7616	0.9220	-1.88	0.022	0.029	27
	Wal	101	-0.0415	-0.9957	1.0696	-2.00	0.004	0.326	24
	Rot	146	-0.0228	-0.4270	0.8433	-4.33	0.069	0.060	28

3.2 CropSyst output

There is a varied response in terms of which substitute provided the best replacement when considering the range of metrics assessed (Table 1). In four cases the nearest substitute provided the best fit for difference in mean and total yield, absolute difference and smallest maximum error. At two sites the second nearest substitute and at three sites the third nearest substitute provided the best replacement. The remaining sites had a mixture of best fitting substitutes for each metric. At some higher latitude sites the yield failed (zero t/ha), due to low temperatures, delaying the accumulation of thermal time (which drives phenological development within CropSyst). This

resulted in large maximum errors and impacted on other metrics: at Aberdeen (Abd) substituted by Aviemore (Avi), a failed yield gave a maximum error of –5.20 (t/ha). Aberdeen and Inverbervie (Inv) (32 km) had the smallest maximum error of -0.33 (t/ha). The mean maximum error for all sites, excluding failed crops, was 1.90 (t/ha). The relationship of distance with ability to match the model output per metric used did not follow a consistent overall pattern. The yield estimates produced here by CropSyst are a function of the combination of precipitation, temperature and SR. Temporal and spatial variability of these climate data can thus impact on the metrics used here, but with differing patterns (Table 1), i.e. best fit for mean yield but not total yield.

The scales of differences also reflect the spatial similarity between sites. At the sites in the dense meteorological network of the geographically similar south-east Britain, i.e. Denver, Bracknell (Bra) and Sutton Bonington, comparing the mean yield difference / n (t/ha) for the three nearest sites, all values were similar. Conversely, at the more isolated, topographically diverse sites, i.e. Auchencruive (Auc) and Eskdalemuir (Esk), the values were considerably different (Table 1). Some sites did show an increase in absolute difference / n (t/ha) with distance, i.e. Aberdeen. At Cawood the inverse situation occurred. Ranking the absolute difference / n (Table 2) results, there is a varied, inconsistent response with distance. There are substantial yield estimate differences (absolute difference / n), the lowest ten having a mean of 0.379 (t/ha).

Table 2. Substitute sites providing the ten lowest absolute differences / n in yield (t/ha).

Site	Subst	Distance (km)	Absolute diff / n (t/ha)
Abd	Inv	32	0.189
Ald	Auc	140	0.342
Den	Bro	40	0.354
Caw	Den	171	0.366
Den	Rot	103	0.403
Loc	Avi	86	0.413
Ald	Dun	206	0.413
Bro	Rot	81	0.429
Alt	Loc	68	0.436
Bra	Wal	33	0.447
		Mean	0.379

3.3 Impacts of solar radiation data source on CropSyst output.

From the method used here, there was no discernable impact of the SR data source alone on the output from CropSyst. One pair of sites, Loch Glascarnoch (Loc) and Tulluch Bridge had a significant negative correlation (-0.997, $p = 0.004$, $n = 5$) between crop yield and SR RMSE.

4 DISCUSSION

The dissimilarity in SR data between sites shows complex relationships with topography, geographic location and distance. Figure 1 shows a clear trend of increasing dissimilarity with distance. The impacts of substituting climate data from alternative meteorological stations into a land use model do not however, always show direct relationships with distance. An explanation of the relationship is given in the shape of the RMSE regression curve for each site. These indicate a function of the density of the meteorological site network, the topographical characteristics of the site being examined and the amount of data (years) per site. Sites in the south east of Britain, i.e. Rothamstead (Figure 1) are in a denser network, hence a greater probability of similarity with nearby sites. The nearest site to Rothamstead is Bracknell (Bra) (51km) with a mean RMSE of 1.94 $MJ/m^2/day$. Conversely, Aberporth is coastal and isolated; its nearest site, Sutton Bonington (238km), had a mean RMSE of 5.16 $MJ/m^2/day$. The length of record of neighbouring sites impacts upon the estimated error associated with the data. For example, sites near to Dunstaffnage have only 4 – 5 years of data, hence they have larger errors associated with them (Figure 2).

The RMSE comparison of SR dissimilarity includes positive and negative differences between one site and another. Hence the accumulative difference is not represented. Further study is required to determine the temporal differences in substitute data, as these could have substantial variations between sites. This study has used data from a whole year, but the most influential differences may occur during the growing season. When using the mean RMSE value, a period of constantly negative differences during the growing season can be balanced by positive differences outside of the growing season. The impact on model output though can be substantial.

In this investigation, the impact on CropSyst output has no readily identifiable pattern (Table 1), other than a general trend of an increase in yield metric differences with distance. However, there are significant exceptions which prevents the assumptions being made that the nearest meteorological stations will automatically provide suitable replacement data. Substitution of data will incur a minimum of estimation error per year (Table 2), which is inconsistent with distance to data source. The method used here was unable to detect the role of SR alone in contributing to the differences in yield estimates, which were attributed to all four data types used acting in conjunction.

5 CONCLUSIONS

5.1 Dissimilarity in solar radiation data

There is a clear trend of increasing dissimilarity in SR with distance. However, there is sufficient variation to exclude the assumption that the nearest data source provides the closest match. The rate at which the dissimilarity between sites increases is a function of two key factors: the density of the data source network and geographical location. At sites within a dense network of meteorological stations, there is a greater probability of a neighbouring site providing similar SR data. Isolated sites have a greater probability of having dissimilar data from the nearest neighbours. A third factor is the length of data record, which impacts upon the magnitude of variability between data similarity. There is a minimum automatic increase in dissimilarity with distance (2 - 4 $MJ/m^2/day$), varying with network density and geographical position.

5.2 Relationship of distance to data source and impact on model output

The impacts of substitute climate data sources on CropSyst output show a complex relationship. The nearest meteorological station with a full compliment of climate data types does not necessarily provide the best substitute. The suitability of a substitute depends on the metric used for comparison. Whilst some substitute sites are able to provide the best replacement climate data for all metrics, others only satisfy one or several. Differences in model output were a function of the accumulative impact of all climate variables. Whilst some nearest substitute sites' data provided good replacements for the range of metrics used, it is clear that distance to data source is not the prime factor. In this study the use of substitute data will introduce a minimum error of approx +/- 0.4 t/ha and a mean maximum potential error of 1.90 t/ha to the yield estimate. It was not possible to quantify the impact of SR alone on model output. However, the spatial and temporal patterns in amount of SR received would have contributed to the differences in model output. Further studies will investigate the influence of other climate data types and their spatial dissimilarities.

5.3 Impacts on model output interpretation

There was no consistent relationship between the distance to data source and ability to provide the best match for each of the yield metrics. There is a complex relationship between the data similarity between sites and the ability of substitute data to provide representative model output. This implies that practitioners interpreting model output need to consider the impacts arising from using substitute data. No one single substitute will provide a best overall representation when considering a range of output metrics. In this context, the distance to the substitute data source becomes just one of several considerations. However the model output is used (decision support, tactical response), it becomes increasingly important to quantify the impacts arising from the use of substitute data when applied to a site-specific study.

5.4 Recommendations

A combination of meteorological station network density and geographical information can be used to help identify suitable data substitutes. Based on these results, it is not feasible to specify a single threshold in the UK for the distance – data decay relationship and the need to use alternative SR data sources. Instead a range of thresholds are required, due to the maritime climate and diverse topographical spatial arrangement. Localised thresholds would be more appropriate, using network density, data record length and topographical characteristics as guides. South-east sites that exist in a dense meteorological station network can have a higher threshold than isolated sites in more diverse topography. The uncertainty arising from the use of substitute data within simulation models needs to be quantified to maintain credibility in the models reliability.

6 AKNOWLEDGEMENTS

The authors gratefully acknowledge the funding support of the Scottish Executive Environment and Rural Affairs Department and the Meteorological Office for use of the climate data.

7 REFERENCES

Hoogenboom, G., Contribution of agro-meteorology to the simulation of crop production and its applications. *Agriculture and Forest Meteorology*, 103, 137-157. 2000.

Hunt, L.A., L. Kuchar, and C.J. Swanton. Estimation of solar radiation for use in crop modelling. *Agricultural and Forest Meteorology 91*, 293-300, 1998.

Rivington, M., K.B. Matthews, and K. Buchan. A Comparison of Methods for Providing Solar Radiation Data to Crop Models and Decision Support Systems. Proc. Int. Environmental Modelling and Software Society, Lugano, Switzerland, 24-27 June. Vol 3, 193-198.

Stöckle, C.O. and R. Nelson. *CropSyst Users Manual*. Washington State University, Pullman, WA. 1998.