Impact of Soil Moisture Data Characteristics on the Sensitivity to Crop Yields Under Drought and Excess Moisture Conditions

Catherine Champagne 1,*,†, Jenelle White 1,2, Aaron Berg 2,§, Stephane Belair 3,∥ and Marco Carrera 3

1 Agriculture and Agri-Food Canada, Ottawa, ON K1A0C6, Canada; jwhite18@uoguelph.ca
2 Department of Geography, University of Guelph, Guelph, ON N1C1A6, Canada; aberg@uoguelph.ca
3 Environment and Climate Change Canada, Dorval, QC H4S1Y9, Canada; stephane.belair@canada.ca (S.B.); marco.carrera@canada.ca (M.C.)

* Correspondence: catherine.champagne@canada.ca
† Her Majesty the Queen in Right of Canada, as represented by the Minister of Agriculture and Agri-Food Canada.

Abstract: Soil moisture is often considered a direct way of quantifying agricultural drought since it is a measure of the availability of water to support crop growth. Measurements of soil moisture at regional scales have traditionally been sparse, but advances in land surface modelling and the development of satellite technology to indirectly measure surface soil moisture have led to the emergence of a number of national and global soil moisture data sets that can provide insight into the dynamics of agricultural drought. Droughts are often defined by normal conditions for a given time and place; as a result, data sets used to quantify drought need a representative baseline of conditions in order to accurately establish a normal. This presents a challenge when working with earth observation data sets which often have very short baselines for a single instrument. This study assessed three soil moisture data sets: a surface satellite soil moisture data set from the Soil Moisture and Ocean Salinity (SMOS) mission operating since 2010; a blended surface satellite soil moisture data set from the European Space Agency Climate Change Initiative (ESA-CCI) that has a long history and a surface and root zone soil moisture data set from the Canadian Meteorology Centre (CMC)’s Regional Deterministic Prediction System (RDPS). An iterative chi-squared statistical routine was used to evaluate each data set’s sensitivity to canola yields in Saskatchewan, Canada. The surface soil moisture from all three data sets showed a similar temporal trend related to crop yields, showing a negative impact on canola yields when soil moisture exceeded a threshold in May and June. The strength and timing of this relationship varied with the accuracy and statistical properties of the data set, with the SMOS data set showing the strongest relationship (peak $X^2 = 170$ for Day of Year 145), followed by the ESA-CCI (peak $X^2 = 89$ on Day of Year 129) and then the RDPS (peak $X^2 = 65$ on Day of Year 129). Using short baseline soil moisture data sets can produce consistent results compared to using a longer data set, but the characteristics of the years used for the baseline are important. Soil moisture baselines of 18–20 years or more are needed to reliably estimate the relationship between high soil moisture and high yielding years. For the relationship between low soil moisture and low yielding years, a shorter baseline can be used, with reliable results obtained when 10–15 years of data are available, but with reasonably consistent results obtained with as few as 7 years of data. This suggests that the negative impacts of drought on agriculture may be reliably estimated with a relatively short baseline of data.

Keywords: soil moisture; drought; agriculture; SMOS; ESA-CCI
1. Introduction

Agricultural droughts occur when water shortages lead to reductions in crop productivity, and are most often associated with increases in evapotranspiration coupled with reductions in root zone soil moisture. This type of drought has been traditionally difficult to monitor with tools such as meteorological stations, since neither soil moisture nor evapotranspiration are typically measured. As a result, indicators of agricultural drought are most often modelled using water budget models [1–3]. While these approaches have shown robust results in monitoring global changes in drought occurrence, they are often limited at more regional to local scales due to the uncertainty in both the data and the models at finer spatial resolutions. In recent years, soil moisture data sets derived from microwave satellites have emerged as a new source of information to support drought monitoring. Satellites collecting passive microwave emissions at X-, C-, and L-band frequencies have been used to derive soil moisture with varying levels of accuracy at coarse spatial resolutions every 1–3 days globally. Newer satellites, such as the Soil Moisture Ocean Salinity (SMOS) mission and the Soil Moisture Active Passive (SMAP) mission were designed explicitly for soil moisture monitoring. These soil moisture data sets have been validated around the world using in situ station networks. Validation results have shown that the volumetric soil moisture derived from L-band frequencies is more accurate than those using higher radiometric frequencies such as X- and C-band, but the temporal trend in soil moisture (i.e., wetting and drying cycles) can be captured reasonably well using all three types of data [4,5]. Satellites are sensitive to the moisture in a thin layer of soil at the surface, limiting their use for monitoring root zone soil moisture most closely associated with agricultural droughts. Data assimilation into land surface models has been used successfully to improve the estimation of water storage in the root zone, particularly in areas where soil properties are not well characterized, or precipitation estimates have low accuracy [6,7]. The greater availability and accuracy of soil moisture data sets provides a promising tool for agricultural drought monitoring.

Droughts are most often characterized not by aridity or absolute thresholds in moisture but by the relative dryness in the context of historical conditions for a given time and place. For this reason, most drought indicators are reliant on creating a baseline of normal conditions using historical data, and then determining drought severity by comparing to this normal [8,9]. Current soil moisture conditions alone, therefore, are rarely enough to fully characterize drought. Two general approaches have been used to estimate drought severity using satellite soil moisture data sets. The first set of approaches use a baseline of historical conditions to define statistical characteristics of soil moisture for each location to develop a relative indicator of drought severity [10–12]. Another approach to involves using the field capacity, wilting point or the available water holding capacity of soils to determine water storage in a particular soil [13,14]. A key limitation to the first approach is the lack of historical satellite soil moisture data to define these baseline statistical conditions [10]. A limitation of the second approach is the need to define soil characteristics which are not often available at a suitable spatial scale and which can be statistically incompatible with the satellite data sets. [14]. The historical baseline for a single satellite is typically short, spanning the operating lifetime of an instrument, (~1 to 20). To address the need to contextualize and quantify long term changes in soil moisture, the European Space Agency (ESA) developed a multi-sensor satellite soil moisture data set under the Climate Change Initiative (CCI) essential climate variables program (hereafter referred to as ESA-CCI) providing a statistically consistent soil moisture data set from 1979 to the present [15]. This opens up the potential to better quantify relative drought conditions using earth observation data [16].

With a large number of potential models and satellite data sets available for monitoring droughts, it raises the question on what characteristics of the data are critical to adequately monitoring drought and drought impacts. There are multiple modelled and satellite soil moisture data sets that exist, and each has its flaws: satellites only monitor the surface conditions; models can provide root zone soil moisture but often have higher errors and biases; L-band sensors provide the highest accuracy, but they have only been available for a short period of time; longer satellite blends have a relatively long temporal record, but may not be as accurate or timely for monitoring. The impact of these data
characteristics for quantifying agricultural droughts has not been clearly identified based on existing research. Different satellite soil moisture data sets have different accuracies relative to in situ networks, have different historical baselines and are only sensitive to surface soil moisture. The impact of these differences on drought monitoring and assessment are not clear from the current research.

The emerging research on satellite soil moisture and drought has shown that there is a broad sensitivity to drought risk, shown in several studies in Canada, the US, Europe and Africa [17–20]. Indices calculated from the longer-term ESA-CCI data set have shown good correspondence with drought events or agricultural yields, making use of longer, more robust historical baselines on which to establish the relationship with drought or yield [11,12,21]. As an absolute measure of soil moisture, ESA-CCI does show variations in accuracy and may provide less accuracy than SMOS or SMAP as an absolute measure of soil moisture [15]. Satellite soil moisture-based drought indices have been found to be primarily indicative of short-term dry conditions, which may be expected given that they measure only the surface conditions [22,23]. Soil moisture from satellites have large uncertainties in areas with dense vegetation (forests), complex topography or organic soil types [15]. However, satellite soil moisture data sets have been shown to provide higher accuracy than modelled data sets, particularly where the data inputs for models have higher uncertainty [24]. Satellite soil moisture data sets have been shown to exhibit similar error structures to other satellite soil moisture data sets, whereas models have different error structures than satellite data sets, but are often similar to other modelled data sets [25].

Based on this limited research, an ideal data set for monitoring agricultural droughts would be one with high and consistent accuracy over time, and provide a measure of soil moisture at both the surface and root zone. Currently, this data set does not exist, so there is a need to evaluate the tradeoffs in using the data sets that currently do exist. While both modelled and satellite derived soil moisture datasets have shown sensitivity to drought, it is not clear how they differ in capturing these conditions and what the advantage of using one data set over another provides. The objective of this research is to evaluate the sensitivity of earth observation and modelled data sets to agricultural drought conditions in terms of three conditions: accuracy, length of the temporal baseline and estimation of root zone conditions. This was evaluated by examining changes in the statistical relationship between soil moisture and crop yield using an iterative chi-squared modelling approach. This approach was selected since it has been used in the past for examining the relationship between soil moisture and crop yield so the expected behaviour is well-understood [26–28]. Three different data sets were evaluated: a short record soil moisture data set from a dedicated soil moisture sensor (SMOS), a longer term but potentially lower accuracy soil moisture data set from ESA-CCI and a modelled soil moisture data set from the Canadian Meteorology Centre (CMC)’s Regional Deterministic Prediction System (RDPS) that captures soil moisture at the surface and in the root zone. The three data sets were compared over a coincident time period as well as over the full period of record for each data set.

2. Methodology

2.1. Study Area

The province of Saskatchewan in the western Canadian prairies was selected for this study. This area is largely agricultural and has a low vegetation biomass that is consistent with high accuracy satellite soil moisture retrievals [4]. Located in the North American Great Plains, Saskatchewan is characterized by fertile soils that support the largest land area of agricultural production in Canada, consisting primarily of small grains and cool season oilseeds [29]. Inter-annual variations in crop yield in this region are largely driven by moisture variability due to its geographic position east of the Rocky Mountain range, and a generally semi-arid climate. Canola, a variant of oilseed rape and the largest area crop grown in Canada, was selected to evaluate the impacts of soil moisture data characteristics. Canola is a cool season crop that is highly sensitive to drought conditions, making canola yields a good indicator of the occurrence of agricultural droughts [27,30,31].
Crop yield data in Saskatchewan is distributed through administrative districts known as Census of Agriculture Regions (CARs) [32]. For this study, the 20 CARs from the 2011 census by Statistics Canada were used (Figure 1), ranging in size from approximately 9000 to 350 000 km\(^2\). Canola yield data is collected annually through a statistical survey of farmers by Statistics Canada in kilograms per hectare for each year from 1992 to 2015 [32]. Crop yields tend to trend positively over time, reflecting improvements in farming practices and seed hybrid development, so for the purposes of this study, a linear detrending was applied to the yield data prior to analysis. For this analysis, all soil moisture data sets were averaged to the CAR level to facilitate assessment against crop yield.

![Figure 1. Census agricultural regions (CARs) of the province of Saskatchewan. Canola growing regions across Canada are depicted in light grey.](image)

**2.2. Soil Moisture Data Sets**

Three soil moisture data sets were evaluated against crop yields that had different periods of record, different accuracies, and representing different soil depths.

**2.2.1. SMOS**

Soil moisture data from the Soil Moisture and Ocean Salinity (SMOS) mission was used covering a period from 2010 to 2015. SMOS collects passive microwave emitted radiation from a surface footprint of approximately 40km, with repeat coverage over most locations in Canada every 1–2 days. Volumetric soil moisture from SMOS was used from version 6.50 of the SMOS soil moisture processor, which uses the tau-omega model to quantify soil dielectric constant and vegetation opacity using multi-angular brightness temperatures [33]. Daily volumetric soil moisture measurements were interpolated to a 0.25-degree spatial grid and masked for the occurrence of snow, rain, high radio frequency interference, and frozen soil temperatures at the time of acquisition using the data quality flags supplied. Daily values were averaged for each Census of Agriculture Region (CAR) in Saskatchewan over the study.
period. For this study, data covering a period from 2010 to 2015 were used to analyze the relationship between soil moisture and crop yield as described in [28]. This previous study found a strong statistical relationship between high soil moisture in the late spring (June) and low crop yield, which was associated with extreme wetness over the study period, leading to water logging that could negatively impact crop growth and development. This data set was used as a baseline to determine if similar trends could be found using other soil moisture data sets.

2.2.2. ESA-CCI

The ESA-CCI soil moisture data was developed to provide a long-term climate data set for evaluating long term trends. For this study, version 4.2 of the data was used, and covers a period from 1979 to 2016. Past research has indicated that the period from 1992 to the present provides the most consistent data, and for the purposes of this study, the data set was restricted to this time period. This data set includes a blend of active and passive soil moisture data sets, from the Special Sensor Microwave Imager (SSM/I), the Advanced Microwave Scanning Radiometer (AMSR2 and AMSR-E), Windsat and the Metop Advanced Scatterometer (ASCAT) [15]. Level 2 soil moisture products are derived using two sets of models: for passive sensors, the Land Parameter Retrieval Model (LPRM) model is applied, which uses a multi-frequency approach to separate emissions from vegetation and soils, and for active sensors, the TU Wien method is applied, which uses a change detection approach to normalize for scattering from different surface elements [15,34,35]. Level 2 soil moisture data sets are statistically rescaled by the data providers to a consistent long-term modelled soil moisture data set from the NASA Global Land Data Assimilation System (GLDAS) using a set of decision rules based on the error characteristics of each data set and gridded to a 0.25 degree spatial grid. For this study, daily volumetric soil moisture observations from the ESA-CCI active-passive data set were spatially averaged for each CAR in Saskatchewan over the period from 1992 to 2015. This data set represents a long-term, temporally stable data set that was used to evaluate the impact of baseline length on determining crop yield sensitivity to soil moisture.

2.2.3. Canadian Meteorological Centre Soil Moisture

Surface and root zone soil moisture from the Canadian Meteorological Centre (CMC) was used from the Regional Deterministic Prediction System (RDPS) land surface model. The RDPS uses the Interactions between Soil–Biosphere–Atmosphere (ISBA) land surface scheme, which comprises two soil layers with associated soil characteristics variables and describes the evolution of temperature and water content based on a “force–restore” mechanism [36,37]. The ISBA land surface model is forced with atmospheric data from short-range forecasts from the Canadian Meteorological Centre’s Numerical Weather Prediction (NWP) models. The atmospheric forcing variables required are short- and long-wave radiation incident at the surface, air temperature, specific humidity, wind, surface pressure, and precipitation. Air temperature, specific humidity, and wind forcing are taken from the lowest vertical level in the NWP model (40 m), with the other variables representing surface values [8]. The soil moisture analysis from the model is derived after assimilation of screen level observation of temperature and relative humidity to correct for model errors. The surface is modelled at a 10km grid resolution. Soil moisture estimates at depths of 0–10 cm and 0–100 cm were output by the model and used for this analysis. Data were analyzed over a period from 2011 to 2015 since 2010 data was not available on a 10 km grid.

2.3. Iterative Chi-Squared Modelling

The iterative chi-square analysis, developed by Caprio [38], is a statistical procedure designed to investigate the association between climate observations and biological data, including crop yields. The technique has been applied to numerous applications, including relating temperature and precipitation observations to wheat yield records in Montana [39]; apple and grape yields in British Columbia [40,41]; canola yields in Saskatchewan [28]; and cabbage, onion, and rutabaga yields in
Ontario [26,42]. The technique has also been used to relate temperature and precipitation variables to tree-ring growth in southern Arizona [43] and in comparing temperature, precipitation and modelled soil moisture estimates to grasshopper pest populations in Alberta [44].

The iterative chi-square analysis identifies the timing, magnitude and direction (positive or negative) of the relationship between individual daily climate observations and crop production throughout the growing season and determines the climate threshold values above or below which the most significant associations occur [26,27,38,44]. The technique iteratively compares the number of days that meet a threshold condition in high- or low-yield years (observed) to normal-yielding years (expected) within a three-week moving window, generating an overall chi-squared statistic (Equation (1))

\[ X^2 = \frac{(O - E)^2}{E} \]  

where \( O \) is the observed number of days that meet the condition and \( E \) is the expected, or ‘theoretical’, number of days meeting the condition. If the number of days meeting the threshold condition has a low probability of occurring by chance, the returned chi-squared value departs from zero indicating a statistically significant relationship. For this study, chi-squared values greater than +10 or less than −10 were significant at \( p < 0.01 \), or where more than twice the number of observed cases meet the condition relative to the expected [43]. Larger chi-squared values represent a stronger relationship between the two variables being examined. Positive chi-squared values indicate a positive relationship between high values of the variable being assessed (in this case soil moisture), whereas negative values indicate a relationship with low values. For example, low yields associated with high soil moisture would be indicated by high positive chi-squared values, whereas low yields associated with low soil moisture would be shown as strong negative chi-squared values. These values can then be plotted over time to indicate the timing of the peak relationship between crop yields and climatic variables. Two previous studies using this technique are noteworthy for establishing the robustness of this technique over the study area. A study by Kutcher [27] found a significant chi-squared relationship between high temperatures and low precipitation during canola flowering (early July in this region) that resulted in lower than average yields when examined over a 35-year period from 1967–2001. A follow-up study over this same study area examined the relationship between SMOS soil moisture, temperature and precipitation over a shorter time period (2010–2015) and found a similar relationship between extreme high temperatures during flowering and low yields, but could not replicate the trend in precipitation data [28]. This same study found a significant relationship between high soil moisture from SMOS in late May/early June and canola yield, a result of several seasons of excess moisture [28].

The present study uses these established relationships between climatic variables and canola yield in Saskatchewan to evaluate the impacts of soil moisture data characteristics and the strength and trend in this relationship. To do this, the iterative chi-square technique was utilized to compare daily soil moisture observations from the SMOS, ESA-CCI, and RDPS data sets in high- or low-yielding years between 2010 and 2015 across the 20 CARs of Saskatchewan. The impact of data set baseline length was evaluated using the ESA-CCI data set from 1992–2015. Low-, normal-, and high-yield classes were determined based on quartiles, where low-yield years were defined as the bottom 25% of samples, normal-yield years between 25% and 75%, and high-yield years the top 25%. Thresholds were searched in steps of 0.2% from 0 to 50% for daily percent volumetric soil moisture by scanning the data from high-to-low to establish the relationship.

2.4. Climate Conditions

The soil moisture results were interpreted using standard meteorological indices as well as the Canadian Drought Monitor from the Agriculture and Agri-Food Canada National Agroclimate Information Service [45]. The period under examination showed a range of drought conditions and excess moisture conditions that would impact canola yield in this region. Figure 2a shows the ratings from the Canadian Drought Monitor from 2003 (when assessments began) to 2015. A severe
drought impacted the area from 2000–2005, having both high severity and widespread conditions. Drought events were periodic but regular after 2005, with extreme drought event in 2008 and 2009, and relatively few events in terms of both extent and severity from 2012 onwards. The longer-term drought severity conditions are shown in Figure 2b, depicting the average Palmer Drought Severity Index (PDSI) conditions over the study region for the period of 1992–2015. The PDSI was calculated using a customized model that includes a soil water balance model to reflect general soil moisture conditions [46]. The data show the severe drought conditions from 2000–2005, as well as a number of wetter than average periods, including peaks in 2010, 2011, and 2014.

**Figure 2.** Temporal trend in the Canadian Drought Monitor ratings from 2003–2015 for Saskatchewan (a) and temporal trend of the Palmer Drought Severity (PDSI) ratings for Saskatchewan from 1992–2015 (b).
3. Results and Discussion

3.1. Soil Moisture Data Characteristics

The statistical characteristics of each data set are given in Figures 3 and 4 and Table 1. For the surface soil moisture data sets covering the core period where all three data sets overlap from 2010 to 2015 (Figure 3; Table 1 top), the SMOS and RDPS data sets have slightly lower means that the ESA CCI data set, and the SMOS data set has a slightly larger distribution (standard deviation = 7.1%). The RDPS data has a higher skew and kurtosis than the other two data sets, likely because of the large number of low moisture values at the extremes of the distribution. The distribution of the three soil moisture data sets using the full record length of each data set are shown in Figure 4. The full ESA-CCI data set has a lower mean than the 2010–2015 subset, indicating a wet bias in the more recent years, but other statistical characteristics of the 2010–2015 data subset are similar. Both the long- and short-term ESA-CCI data sets show a skew towards lower soil moisture values compared with the SMOS data set. Examining other six-year subsets from the ESA-CCI data set, the period from 1998–2003 has the lowest mean value and a slightly higher skew and kurtosis than the 1992–2015 data set, with the 2004–2009 subset being the most similar statistically to the full ESA-CCI data set. The RDPS root zone soil moisture has a higher mean than the surface value, with a lower standard deviation and a negative skew, indicating a large number of extreme values in the lower end of the soil moisture range.

**Figure 3.** Comparison of statistical distributions of three surface soil moisture data sets for overlapping period 2010 (2011)–2015.

**Figure 4.** Statistical distribution of three surface soil moisture data sets for full data record. The ESA-CCI data set from 1992–2015, the SMOS data set from 2010–2015 and the RDPS data set from 2011–2015.
Table 1. Statistical characteristics of different soil moisture data sets used in this study.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Sd</th>
<th>Median</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDPS surface 2011–2015</td>
<td>15300</td>
<td>16.0</td>
<td>5.1</td>
<td>15.2</td>
<td>1.2</td>
<td>1.8</td>
</tr>
<tr>
<td>SMOS 2010–2015</td>
<td>21410</td>
<td>17.3</td>
<td>7.1</td>
<td>17.2</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>ESA CCI 2010–2015</td>
<td>18349</td>
<td>19.6</td>
<td>4.1</td>
<td>19.3</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>ESA CCI 1992–2015</td>
<td>61770</td>
<td>18.9</td>
<td>4.5</td>
<td>18.6</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>ESA CCI 1992–1997</td>
<td>7990</td>
<td>19.2</td>
<td>4.9</td>
<td>19.1</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>ESA CCI 1998–2003</td>
<td>12873</td>
<td>17.7</td>
<td>4.9</td>
<td>17.1</td>
<td>0.7</td>
<td>0.6</td>
</tr>
<tr>
<td>ESA CCI 2004–2009</td>
<td>18208</td>
<td>18.8</td>
<td>4.3</td>
<td>18.4</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>RDPS rootzone 2011–2015</td>
<td>15300</td>
<td>17.6</td>
<td>2.8</td>
<td>17.7</td>
<td>−0.4</td>
<td>1.3</td>
</tr>
</tbody>
</table>

3.2. Impact of Data Type on the Relationship Between Soil Moisture and Canola Yield

The iterative chi-square results relating satellite soil moisture to canola yield for Saskatchewan are shown in Figure 5A–C for low and high yielding years. The surface soil moisture conditions show a relatively consistent pattern between all three data sets, with the SMOS soil moisture showing a strong association between low yield and high soil moisture, particularly in the early season period. As was discussed in a previous study, this relationship is likely driven by excess moisture events in several years within this time window that resulted in water logging at the surface and substantial damage to the growing crop [28]. It should be noted that the period from 2010–2015 show, half of the years showed much higher than average precipitation, representing an anomaly for this region. The chi-squared statistic indicating the statistical strength of this relationship is higher for the SMOS data set (peak $X^2 = 170$ for Day of Year 145), somewhat lower for the ESA-CCI data set (peak $X^2 = 89$ on Day of Year 129), and weakest for the RDPS surface data (peak $X^2 = 65$ on Day of Year 129). There is also a difference in the timing of the peak between the three data sets, with the ESA-CCI and RDPS showing a larger peak at day 129 (early May) and a weaker peak around day 142 ($X^2 = 84$ on Day of Year 142 for ESA-CCI and $X^2 = 65$ on Day of Year 142) in late May, consistent with the SMOS data set. This timing would be consistent with the seeding and germination growth stages for canola in this region. It is not clear why this earlier peak in May is not as pronounced in the SMOS data. This could be a result of the statistical distributions of the ESA-CCI and RDPS data sets, which appear to capture more low soil moisture values than high soil moisture values. For this reason, the earlier peak may be in reality less significant (as seen in the SMOS results in Figure 5). The later peak in the relationship between low yields and high soil moisture seen by SMOS may be more agriculturally significant, since soils that may already be saturated from spring rains become over-saturated if heavy rainfall persists, and that this persistence in wet conditions negatively impacts the yield more than wet conditions earlier in the season. In other words, high soil moisture in early May could be less problematic for crop yields if there is time for the soils to dry within a two-week window, but if they do not dry, yields will be negatively impacted by saturated soils. This difference in the magnitude of these two peaks may not be well-captured by ESA-CCI and RDPS because of the bias towards low soil moisture values.

The root zone soil moisture (Figure 5D) showed the weakest relationship with crop yields. There may be several reasons for this. Since the dominant trend over this time period was the impact of excess moisture leading to low yields, it may be that the root zone soil moisture may be less relevant since this is caused by water logging at the surface rather than water storage at depth. Drought conditions would likely be better reflected in root zone soil moisture deficits, which occur when water
storage in the root zone is inadequate to support crop growth [47]. Unfortunately, we cannot evaluate this in this study due to the lack of severe drought impacts during this time period.

Figure 5. Chi-squared statistics between canola yield and soil moisture for three surface soil moisture data sets and one root zone soil moisture data set in the 2010–2015 time period: (A) Satellite surface soil moisture from SMOS; (B) Satellite surface soil moisture from ESA-CCI; (C) Modelled surface soil moisture from RDPS; and (D) Modelled root zone soil moisture from RDPS. Chi-squared values higher than 10 or lower than −10 are statistically significant at $p < 0.01$.

The consistency of the relationship for all three surface soil moisture data sets shows that they can all capture the dominant trend between crop yields and soil moisture over this time period, which is characterized by excess moisture having a negative impact on canola yield. The differences in the strength and timing of the statistical relationship for all three data sets suggest that the SMOS data set, which presumably has the highest accuracy since the sensor characteristics were designed specifically to measure soil moisture, is capturing this trend better than the other two data sets. SMOS soil moisture has been shown to have a higher accuracy than ESA-CCI in previous studies [5]. The weaker relationship with the ESA-CCI data set may be due to the statistical rescaling of that data set to the GLDAS model, which may be reducing the dynamic range of the data set, with a bias toward drier soil moisture values (Figure 4, Table 1). The reason for this bias is not known and is noteworthy since the ESA-CCI data set includes SMOS data in its long-term data blend; this could also be a result of the inclusion of both active and passive soil moisture data sets in the blend, and is worth further investigation. The weaker chi-squared relationship with the RDPS surface data set may be due to a similar skew in the soil moisture distribution. The reason for this skew could be due to simplifications in the model physics to provide a suitable estimate of soil moisture for meteorological applications, which may result in less accurate overall soil moisture estimates. These could likely be improved with future enhancements to the land surface scheme, including changes to the land surface model

and assimilation of satellite soil moisture to improve the accuracy [48,49]. There is also likely some discrepancy in the depths being represented in the surface satellite soil moisture data sets and the model; the model represents a depth of 10 cm, but the sensing depth of the satellite products is likely shallower than this, representing the top 5 cm or less [50].

3.3. Impact of Soil Moisture Baseline Length on Impact Assessment

To examine the impact of the length of the soil moisture baseline on the statistical relationship with crop yield, the chi-squared trend was examined over different six-year periods in the ESA-CCI data set, as well as the full 1992–2015 24-year time period. Using the longest baseline, a very strong relationship is found between low yields and low soil moisture after day 200 (mid to late July, which coincides with the flowering or reproductive growth phase) (Figure 6d). This is consistent with the relationship found in [27] between low rainfall in the first two weeks of July and low canola yields in this region. Similarly, there is a positive association between high soil moisture and high crop yields somewhat later in the season, approximately in August. This relationship was stronger than what has been shown in previous studies [27]. There is a weaker relationship between low yields and excess moisture in the spring (May) than when using the 2010–2015 subset (Figure 5B), but the positive relationship is still apparent even when a longer soil moisture baseline is used.

When different six-year subsets are used, the relationship between soil moisture and canola yield is quite variable. The 1992–1997 subset shows the weakest relationship between soil moisture and both low and high crop yields (Figure 6a) compared with the 1998–2003 and 2004–2009 subsets (Figure 6b,c). These latter subsets show a similar trend to the 1992–2015 long term data set, but the strength and timing of the relationship, particularly between low soil moisture and low yields, differs (peak $X^2 = -141$ on Day of Year 197 for 1998–2003; peak $X^2 = -104$ on Day of Year 200 for 2004–2009 compare to $X^2 = -290$ on Day of Year 206). There is however, a similar trend with low soil moisture being associated with low yields on or around day 200 (mid-July), and high yields associated with high soil moisture after Day of Year 230, which is in approximately the middle of August. At this time of year, canola growth stages would range from peak leaf development to pod development, depending on the timing of seeding. This variable relationship shows that the 1992–1997 period did not capture a wide enough range of conditions to build an accurate relationship between canola yield and soil moisture. This is likely due to the frequency of drought conditions in this area in this time period (Figure 2). There were few droughts in the 1992–1996 period and the 2010–2015 period, making the trend between low soil moisture and low yields difficult to detect. The similarities between the 1996–2003 and 2004–2009 subsets to the 1992–2015 suggest that this relationship can be captured using a shorter climatological baseline, but that the years in question need to be representative of the broader long term trends that would impact crop yields. Since it is unknowable if a short time series representative of these longer-term trends unless a longer term data series exists to verify this, short time series should not generally be considered representative unless otherwise shown to be. The skew in the ESA-CCI data set towards dry soil moisture values discussed earlier may make this data set better for capturing dry extremes than wet extremes.

To further investigate the impact of changing record length on the relationship between crop yield and soil moisture, each year was removed iteratively from the sample to create multiple subsets of data ranging from 6–23 years. Figure 7 shows the iterative chi-squared statistics between these soil moisture data subsets and canola yield. As the length of the data set gets smaller, the strength of the relationship between soil moisture and yield at key periods is weaker, with the drop in the peak season relationship between soil moisture and high crop yield substantially lower when the record is 18 years or less. The relationship between low yielding years and soil moisture is somewhat less reliant on having a longer record length, with the strength of the relationship dropping most when the data record is 12 years or less. To better assess this, Figure 8 shows the correlation between each data subset and the calculated chi-squared statistics using the longest data set (1992–2015). The correlation between different combinations of years becomes much more varied when the data subset is shorter, meaning
that as the sample size decreases, the chances of getting a sample that accurately reflects the relationship detectable from a 24-year sample is much less likely. The correlation between the chi-squared statistic calculated using a 24-year soil moisture data set and a shorter subset for high yielding years starts to weaken (with a correlation of 0.75 or higher) when the record is 20 years or less. For the chi-squared relationship between soil moisture and low yielding years, the relationship is much more stable, with a high correlation for data sets of 7 years or more, and a strong relationship (correlation of 0.9 or higher), when records are 15 years or more. The data requirements for establishing a statistically strong relationship between low yields and high yields likely differ due to the multiplicative factors that influence crop yield. Low yields are often associated with a single constraint that controls crop growth, in this case an inability to seed fields when soil moisture conditions are wet in the spring, or conversely when soil moisture is low during reproductive phases when water controls the rate of seed development. High yielding years, on the other hand, occur when a number of climate related conditions are ideal, and so the need for ideal water storage becomes only one factor contributing to good growth, but not the only one. For this reason, a longer data set may be needed to establish what ideal soil moisture conditions are for high yielding years because it is not the only factor contributing to higher crop productivity, which might also include factors such as warm springs to allow for early seeding or ideal temperatures during key growth stages to reduce heat stress.

Figure 6. Chi-squared statistic between canola yield and satellite surface soil moisture for the ESA-CCI data set over four time periods: (A) Six year period from 1992–1997; (B) Six year period from 1998–2003; (C) Six year period from 2004–2009; (D) 24-year period from 1992–2015. Chi-squared values higher than 10 or lower than −10 are statistically significant at \( p < 0.01 \).
yielding years starts to weaken (with a correlation of 0.75 or higher) when the record is 20 years or less. For the chi-squared relationship between soil moisture and low yielding years, the relationship is much more stable, with a high correlation for data sets of 7 years or more, and a strong relationship (correlation of 0.9 or higher), when records are 15 years or more. The data requirements for establishing a statistically strong relationship between low yields and high yields likely differ due to the multiplicative factors that influence crop yield. Low yields are often associated with a single constraint that controls crop growth, in this case an inability to seed fields when soil moisture conditions are wet in the spring, or conversely when soil moisture is low during reproductive phases when water controls the rate of seed development. High yielding years, on the other hand, occur when a number of climate related conditions are ideal, and so the need for ideal water storage becomes only one factor contributing to good growth, but not the only one. For this reason, a longer data set may be needed to establish what ideal soil moisture conditions are for high yielding years because it is not the only factor contributing to higher crop productivity, which might also include factors such as warm springs to allow for early seeding or ideal temperatures during key growth stages to reduce heat stress.

Figure 7. Chi-squared statistic for different data subsets based on the number of years using the ESA CCI satellite surface soil moisture. The chi-squared statistic quantifies the relationship between soil moisture and high yielding years (top) and low yielding years (bottom). Chi-squared values higher than 10 or lower than $-10$ are statistically significant at $p < 0.01$. 

Figure 8. Spearman rank correlation between the chi-squared statistics for the 24-year data record and shorter data subsets. The chi-squared statistics were calculated between each data subset from the ESA-CCI soil moisture data set and high yielding years (left) or low yield years (right).
The strength of the relationship between yield and the soil moisture at key periods is stronger the more years there are in the data record (Figure 9). The maximum chi-squared value (shown here as a negative, since it indicates a relationship between low soil moisture and low yields) gets stronger as more years are included in the data record. The timing of the peak statistical relationship also changes somewhat, with the relationship primarily in the mid-range of the data (~DOY 200) as more years are in the sample size, and the relationship drifting either earlier or later in the season when the sample size is smaller. The reason for this may be that when fewer years are included in the baseline, the statistical relationship is influenced to a greater degree by the characteristics of the individual years in the sample rather than a general pattern. Since the shifts in timing of the peak occur both earlier and later in smaller samples, this suggests that a longer-term average is more suitable for establishing a generalizable pattern than a shorter baseline, but that there is a range in timing of the peak which is critical for interpreting the results. Year to year differences in the timing of seeding and growth phases may account for this range. For high yielding years, the peak statistical relationship is similarly stronger and less variable when there are more years included in the baseline. In this case the relationship is between high yield and high soil moisture, so the chi-squared value is expressed as a positive. There is no clear pattern of drift in the timing of this relationship for high yielding years.
Figure 9. Peak chi-squared value calculated between ESA-CCI soil moisture data subsets for high yielding years (top) and low yielding years (bottom). Chi-squared values higher than 10 or lower than −10 are statistically significant at \( p < 0.01 \).

4. Conclusions

The relationship between canola yield and different soil moisture data sets was examined to better understand how data characteristics effect the assessment of how soil moisture impacts crop yields.
Three soil moisture data sets were examined: a high accuracy satellite surface soil moisture data set from the Soil Moisture and Ocean Salinity (SMOS) mission, a longer term satellite surface soil moisture data set from the European Space Agency Climate Change Initiative (ESA-CCI) and a modelled soil moisture data set at the surface and at root zone from the Environment and Climate Change Canada Regional Deterministic Prediction System (RDPS). All three data sets showed a similar pattern in the relationship between low crop yield and high spring surface soil moisture for the 2010–2015 period, but the timing and strength of the relationship differed with each data set. For the 2010–2015 period where all three data sets were coincident, the SMOS data set showed the strongest relationship (peak $X^2 = 170$ for Day of Year 145), followed by the ESA-CCI (peak $X^2 = 89$ on Day of Year 129) and then the RDPS (peak $X^2 = 65$ on Day of Year 129) between low yielding years and high soil moisture. The SMOS soil moisture had the highest dynamic range of the three data sets, with modelled data set from RDPS having the smallest. Both the ESA-CCI and the RDPS data set were skewed towards drier soil moisture values, and there is some indication that they captured dry extremes better than wet extremes in this study. The modelled root zone soil moisture was not related to any dominant trend in crop yield in the 2010–2015 time period; this may be due the lack of drought events during this time window, which would be more reflected in the root zone soil moisture than at the surface.

Using short baseline soil moisture data sets can produce consistent results compared to using a longer data set, but the characteristics of the years used in the baseline is important to determine the strength and timing of the relationship. Soil moisture baselines of 18–20 years or more are needed to reliably estimate the relationship between high soil moisture and high yielding years. For the relationship between low soil moisture and low yielding years, a shorter baseline can be used, with reliable results obtained when 10–15 years of data are available, but with reasonably consistent results obtained with as few as seven years of data. This suggests that the negative impact of soil moisture deficits on crop yield are more reliably assessed than the positive impacts. This is important for crop yield modelling, suggesting that even soil moisture data sets with a relatively short number of years of data can be used to identify soil moisture deficits or drought events that will negatively impact crop yields. The better performance of L-band measurements than other data sources suggest the importance of data continuity for missions such as SMOS and SMAP, and the importance of long term, dedicated measurements in order to properly evaluate the impact of droughts and other climate extremes on agriculture.

Author Contributions: Conceptualization, C.C., J.W. and A.B.; Methodology, C.C., J.W. and A.B.; Formal Analysis, C.C. and J.W.; Modelling, S.B. and M.C. Writing-Original Draft Preparation, C.C.; Writing-Review & Editing, J.W., A.B., S.B. and M.C.

Funding: This research was funded by Agriculture and Agri-Food Canada and Environment and Climate Change Canada.

Acknowledgments: The authors wish to thank Richard Warren and Trevor Hadwen from Agriculture and Agri-Food Canada for providing climate and drought data sets. The authors also wish to thank Jon Warland from the University of Guelph for providing the Iterative Chi Squared tools. Yield data was provided from Statistics Canada with processing by Yinsuo Zhang, Aston Chipanshi, Frederic Bédard and Gordon Reichert.

Conflicts of Interest: The authors declare no conflict of interest.

References


© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).