Australian Government
Department of Agriculture and Water Resources
ABARES

The Agricultural Data Integration Project

Neal Hughes, Mihir Gupta, Wei Ying Soh, Chris Boult, Kenton Lawson, Michael Lu, Tim Westwood

Research by the Australian Bureau of Agricultural and Resource Economics and Sciences

Research Report 20.21
December 2020

Cataloguing data

This publication (and any material sourced from it) should be attributed as: Hughes, N, Gupta, M, Soh, W, Boult, C, Lawson, K, Lu, M, Westwood, T (2020) *The Agricultural Data Integration Project*, ABARES research report, Canberra, December. DOI: https://doi.org/10.25814/r8be-kt14 CC BY 4.0.

ISSN 1447-8358

Department of Agriculture, Water and the Environment
GPO Box 858 Canberra ACT 2601
Telephone 1800 900 090
Web agriculture.gov.au

Contents

Tables

Figures

Introduction

The Agricultural Data Integration Project (AgDIP) is a long-term collaboration between ABARES and the Australian Bureau of Statistics (ABS) to develop, integrate and analyse new farm level agricultural data sets. During 2019-20 the project was supported by the Data Integration Partnership of Australia (DIPA).

The AgDIP establishes a new national database of Australian farms, which includes information on agricultural production, business financial outcomes, weather conditions and commodity prices over the period 2000-01 to 2017-18. This database has significant long-term value to government and could inform a wide range of agricultural and environmental issues of relevance to Australian farms including productivity, industry structure, drought, climate change and water policy.

This report provides a summary of the achievements of the AgDIP to-date, including the construction and integration of new data sets and the application of these data sets to develop insights on the effects of drought on Australian farms.

Firstly, this report documents the construction of the Farm-level Longitudinal Agricultural Data set (FLAD), which combines data from all ABS Agricultural collections from 2000-01 to 2017-18. The FLAD provides consistent data on Australian agricultural production for a wide-range of commodities, dryland and irrigated crops and horticulture. FLAD is supplemented with both location specific weather data (e.g., rainfall and temperature) and commodity price data.

Secondly, this report documents the integration of FLAD with the ABS Business Longitudinal Analysis Data Environment (BLADE). This integration allows farm data from the FLAD to be combined with business financial data from the Australian Tax Office, including farm revenue, costs and profits.

Third, these data sets are applied to develop new models, which (similar to models recently developed at ABARES, see Hughes et al 2019) link farm production and profits with climate conditions and commodity prices. In particular, a model of broadacre cropping farms is developed which simulates the production and value of key dryland crops (e.g., wheat, barley, canola, sorghum etc.). In addition, a model of irrigation farms in the southern Murray-Darling Basin is developed which simulates irrigation water use and production.

Finally, some illustrative applications of these datasets and models are presented in a series of case studies. These include: *Trends in Australian crop production* (case study 1), *Small area statistics for WA* wheat (case study 2), *The effects of drought on cropping farms* (case study 3*), Index-based drought insurance for cropping farms* (case study 4) and *Water productivity in the Murray-Darling Basin* (case study 5).

This report concludes by discussing some of the potential directions for future research.

Summary

The Agricultural Data Integration Project (AgDIP) is a long-term collaboration between ABARES and the Australian Bureau of Statistics (ABS) to develop, integrate and analyse new large-scale farm level agricultural data sets. During 2019-20 the project was supported by the Data Integration Partnership of Australia (DIPA).

The AgDIP establishes a new national database of Australian farms, including information on agricultural production, business financial outcomes, weather conditions and commodity prices over the period 2000-01 to 2017-18. This database has significant long-term value to government and could inform a wide range of agricultural and environmental issues of relevance to Australian farms.

The key achievements of the AgDIP to date include the construction of the *Farm-level Longitudinal Agricultural Dataset* (FLAD), the integration of FLAD with the ABS *Business Longitudinal Agricultural Data Environment* (BLADE) and the development of new predictive models linking farm outcomes with climate conditions.

The Farm-level Longitudinal Agricultural Dataset

The FLAD combines farm-level micro-data from all ABS Agricultural surveys and census' undertaken between 2000–01 and 2017–18. The construction of FLAD accounts for variation in ABS collections over time, to provide consistent information on the production of a wide range of agricultural commodities (including dryland and irrigated crops and horticulture). FLAD also contains information on water use, livestock holdings and farm characteristics (e.g., land area and location). The FLAD contains more than 200 individual data items (variables) and nearly 800,000 sample points between 2001-01 and 2017-18, typically covering more than 90% of farms (around 100,000) in census years and 20% (around 20,000) in non-census years.

Integrating with the BLADE

The ABS BLADE provides detailed information on all Australian businesses including Australian Tax Office (ATO) administrative data drawn from Busines Activity Statement (BAS) and Business Income Tax (BIT) filings. For this project FLAD was integrated with the BLADE for the period 2005-06 to 2016-17, predominantly through simple matching of Australian Business Numbers (ABNs). The resulting FLAD-BLADE database can be applied to generate farm level information on production and financial outcomes for essentially every farm business in Australia.

New farm-scale predictive models

This FLAD-BLADE database was combined with climate and commodity price data to develop new statistical models, which can predict agricultural production at a farm-scale given information on prevailing climate conditions (e.g., rainfall and temperature), commodity prices and farm characteristics (location, size etc.). The methodology applied to develop these models follows that of ABARES *farmpredict* model (Hughes et al. 2019). To date, modelling has focused on two sectors: Australian cropping farms and irrigation farms in the Murray-Darling Basin.

Case studies

Five illustrative case studies are presented to demonstrate the potential of the AgDIP data / models. In each case more research would be required to confirm, test and expand the results.

Trends in Australian crop production

In this case study, trends in the area planted and yields for major Australian crops are presented, controlling for the effects of climate variability. This analysis replicates recent ABARES research (Hughes et al. 2017) but covers a wider range of crops and offers higher spatial resolution.

Small area statistics for WA wheat

This case study demonstrates how the AgDIP data and models could be applied to generate experimental small-region crop statistics for public release, overcoming limitations in current public statistics.

Effects of drought on cropping farms

In this case study, the AgDIP data and models are applied to quantify the effects of drought on the production and revenue of Australian cropping farms. This analysis replicates some recent ABARES research (Hughes et al. 2019) but again offers higher spatial resolution.

Index-based drought insurance for cropping farms

This case study provides an illustration of index-based farm insurance and how it could be applied to mitigate drought risk for cropping farms. The case study provides estimates of insurance pay-outs for a hypothetical insurance scheme and shows how these vary across regions and over time.

Water productivity in the Murray-Darling Basin

In this case study trends in water productivity (crop output per unit of water used) are presented for a range of irrigation crops in the Murray-Darling Basin, controlling for annual variability in climate and water prices.

Future development and applications

There are a number of opportunities for further development of the FLAD-BLADE datasets, including improvements to data quality and the continual addition of new years of data as they become available. There also remains significant potential to improve both the performance and coverage of the predictive models developed in this project.

The FLAD / BLADE data sets do have some gaps and limitations which mean they are not a ready-made replacement for existing farm survey-based data collections. Nevertheless, the AgDIP datasets and related models have many potential applications. In the medium term, further refinement of the cropping farms models, could enable small area crop statistics to be produced on a national scale for all major crops. These models could also be linked with BOM seasonal outlook data to generate annual crop production forecasts.

In the longer-term, these data sets could be used to help improve agricultural statistics, support government policy analysis and inform the agriculture and rural finance sectors in a wide range of ways. In particular, these data sets could support detailed evaluations of government policy programmes (i.e., measuring farm-level 'treatment' effects of specific government interventions). The datasets could also be applied to support the development of drought insurance markets.

1 Data integration

1.1 The Farm-level Agricultural Dataset (FLAD)

A key achievement of the AgDIP has been the construction of the Farm Level Agricultural Dataset (FLAD). The FLAD combines micro-data from 18 separate ABS Agricultural collections undertaken between 2000–01 and 2017–18 as listed in Table 1.

Table 1 ABS censuses and surveys used to construct the FLAD

Year	ABS collection
2000-01	Agricultural census
2001-02	Agricultural survey
2002-03	Agricultural survey
2003-04	Agricultural survey
2004-05	Agricultural survey
2005-06	Agricultural census
2006-07	Agricultural survey
2007-08	Agricultural Resource Management Survey / Land Management Practice Survey
2008-09	Agricultural survey
2009-10	Agricultural Resource Management Survey / Land Management Practice Survey
2010-11	Agricultural census
2011-12	Agricultural Resource Management Survey / Land Management Practice Survey
2012-13	Rural Environment and Agricultural Commodities Survey
2013-14	Rural Environment and Agricultural Commodities Survey
2014-15	Rural Environment and Agricultural Commodities Survey
2015-16	Agricultural census
2016-17	Rural Environment and Agricultural Commodities Survey
2017-18	Rural Environment and Agricultural Commodities Survey

ABS agricultural collections were never originally intended for micro-data analysis. For various reasons ABS agricultural collections have varied over time in terms of their scale, coverage and methodology. A key challenge in constructing the FLAD was establishing consistent variable definitions over time, taking into account changes in ABS collections between years. This involved mapping ABS data items from each separate annual collection to a single consistent set of FLAD variables.

In some cases, this mapping from ABS to FLAD data items involved a simple one-to-one match in each year. For example, many common agricultural commodities (such as wheat) have been collected by the ABS in the same format each year and could be mapped directly to a single FLAD variable. In other cases, multiple ABS data items had to be aggregated into a single FLAD variable, particularly where the level of commodity detail collected by the ABS has varied between years.

The set of variables included in FLAD was chosen to reflect the core data items collected by the ABS on a consistent basis over time (excluding other ABS data items if they were collected too infrequently or inconsistently). Most of the FLAD data items are available consistently between 2000-01 and 2017-18; however, for some commodities gaps exist where data was not collected by the ABS in a given year (these missing data are detailed in the FLAD metadata).

The final FLAD data items are summarised in Table 2 and described in detail in the FLAD metadata. The core data items include crop areas and production, livestock holdings and water use for a wide range of agricultural commodities (Table 2). In addition, FLAD contains various data on farm characteristics such as total land area and land composition (grazing land, cropping land etc.), industry type and location (described further in the metadata).

Table 2 Summary of key variables in the FLAD

Category	Variables	Commodities
Broadacre crops	Area planted (ha) Production (t)	Wheat, Barley, Oats, Sorghum, Triticale, Maize, Rice, Cotton, Peanuts, Sugar cane, Canola, Other cereals, Other legumes
Fruit and Nut Orchards	Area planted (ha) Bearing trees (no.) Non-bearing trees (no.) Production, (t)	Mandarins, Oranges, Apples, Pears, Mangoes, Peaches, Cherries, Nectarines, Avocadoes, Olives, Almonds, Macadamias
Other fruit	Area bearing (ha) Area non-bearing (ha) Production, (t)	Bananas, Strawberries, Pineapples, Grapes
Vegetables	Area planted (ha) Production (t)	Carrots, Mushrooms, Onions, Potatoes, Tomatoes, Beans, Lettuce, Melons
Livestock	Livestock on-hand (no.)	Beef cattle, Sheep, Pigs, Dairy cows, Other dairy cattle, Chickens (layer hens), Chickens (meat)
Irrigation	Area watered (ha) Volume of water applied (ML)	Cotton, Pasture, Fruits and nuts, Grapevines, Almonds, Hay, Rice, Sugar cane, Vegetables, Other broadacre, Other cereals, Other crops

For the purposes of linking FLAD units to spatial data (such as weather data), geographic co-ordinates (latitude and longitude) were obtained for each farm in FLAD. The ABS obtain co-ordinates for farms in their surveys by geo-coding address information (via the Geographic National Address File GNAF). This geo-coding was not available prior to 2006 or in the years 2007 to 2010. In these years farm locations were imputed by matching farms between years (for example using 2006 locations for farms that appeared in both 2007 and 2006) or otherwise approximated based on regional identifiers. The mapping of climate data to FLAD is summarised in the next section.

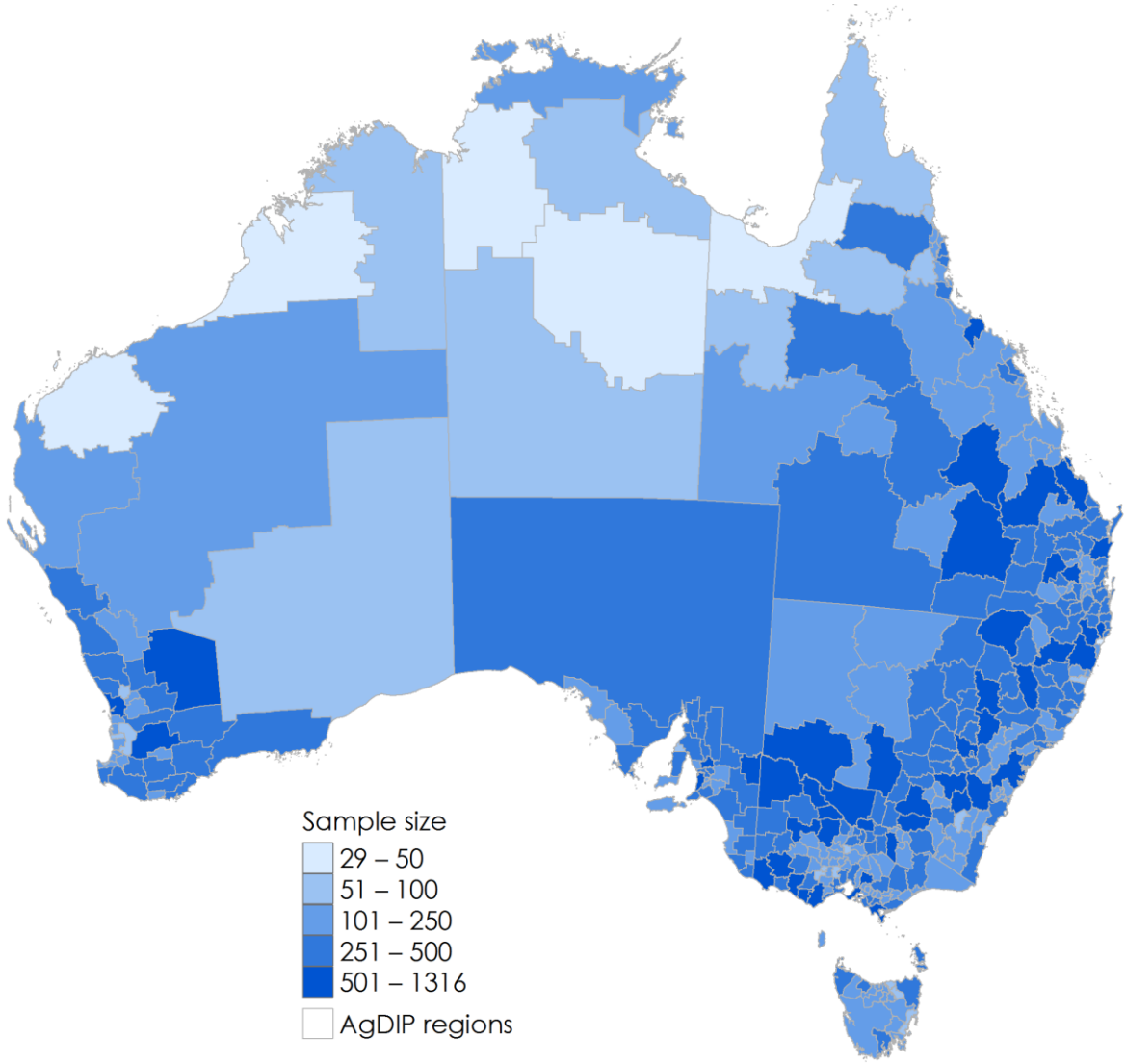
The sample size of FLAD is summarised in Table 3 and Figure 1. Some farm characteristic information (including farm location) are available for the larger ABS target sample. This target sample includes all units selected from the frame (ABS business register) to be sampled by the ABS in a given year (of which survey responses are typically received from around 90%). Note that sample sizes were smaller in the most recent (2015-16 census) due to the higher Expected Value of Agricultural Operations (EVAO) threshold ($40,000 rather than $5000) now adopted by the ABS.

Farm units in the FLAD are assigned longitudinal identifiers: allowing individual farms to be tracked across multiple years. FLAD contains approximately 249,000 unique farm units, with an average of 3.1 years of data available for each farm. Around 22% of the FLAD units are present in the sample for 5 or more years.

Table 3 Sample size of the FLAD

Year	Target sample	FLAD
2000-01	118,180	115,757
2001-02	26,164	24,025
2002-03	22,214	19,245
2003-04	24,251	20,594
2004-05	21,677	21,175
2005-06	164,554	138,300
2006-07	29,497	26,910
2007-08	25,001	23,011
2008-09	29,054	26,895
2009-10	29,205	26,648
2010-11	135,840	119,446
2011-12	29,982	27,776
2012-13	28,011	25,430
2013-14	26,068	23,393
2014-15	25,055	22,479
2015-16	82,932	71,062
2016-17	24,391	22,257
2017-18	21,573	18,524
Total	**863,649**	**772,927**

Figure 1: Average sample size of the FLAD by AgDIP region (census years)



Note: AgDIP regions are aggregations of ASGS 2016 SA1 regions that broadly respect SA2 boundaries, with some modifications to ensure reasonable farm sample sizes and to maintain consistency with AAGIS survey regions

1.1.1 Example: wheat production

Some example statistics for Australian wheat production drawn from FLAD are shown in Figure 3 and **Error! Reference source not found.**. Figure 2 demonstrates the annual volatility in Australian wheat yields driven largely by weather conditions. For example, 2002-03 and 2006-07 were severe drought years with very low wheat yields, while 2016-17 saw a very wet winter which led to exceptionally high yields for wheat. Figure 2 also shows the significant variation in wheat yields across farms in a given year (driven partly by regional differences in rainfall).

Figure 3 shows spatial variation in Australian wheat yields (averaged over the period 2002-03 to 2017-18). Again these regional differences in yield are explained primarily through differences in rainfall, with higher average yields in higher rainfall zones closer to the coast (with some of the highest yields obtained in Tasmania).

Figure 2: Annual wheat yield percentiles, 2000-01 to 2017-18

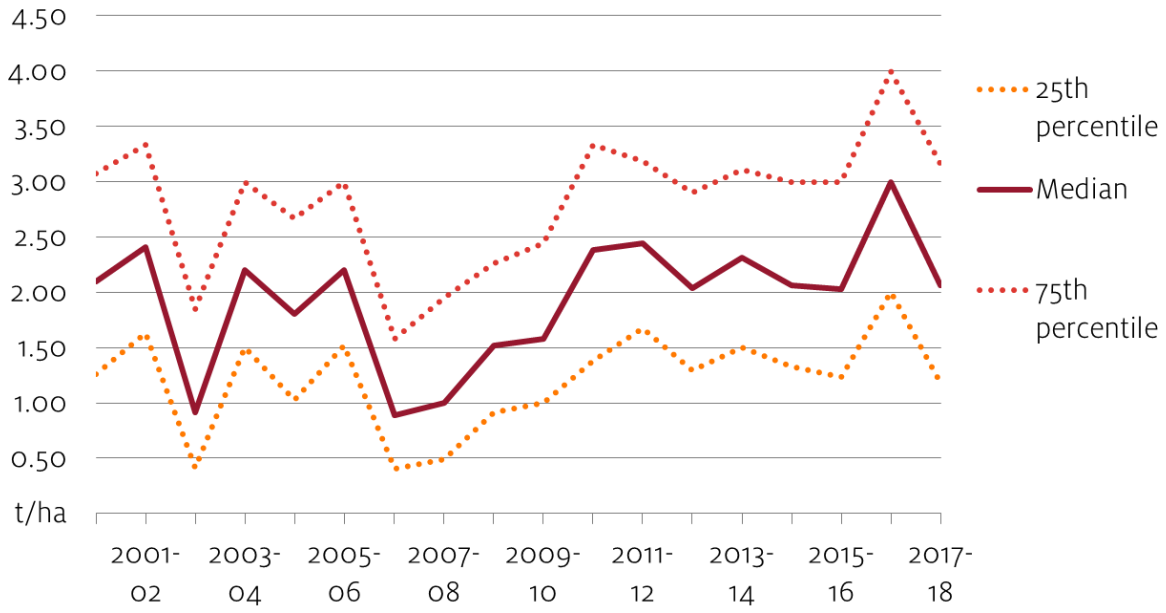
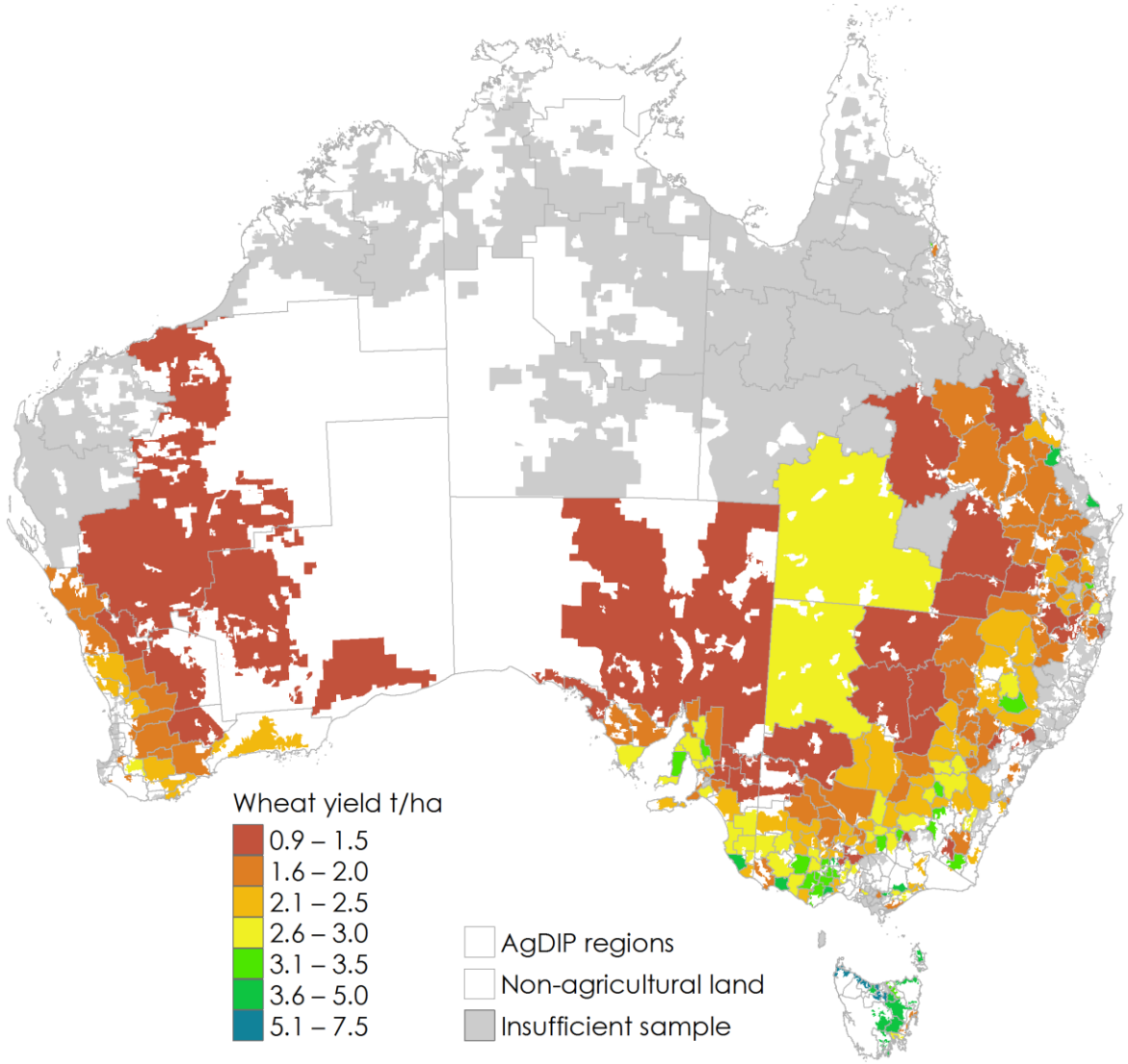


Figure 3 Annual average wheat yield by AgDIP region (2000-01 to 2017-18)



1.2 Climate and price data

1.2.1 Climate data

Climate data were matched to each farm in the FLAD target sample, following a methodology similar to Hughes at al. (2019). Spatial climate data were matched to FLAD farm units, using farm co-ordinates (latitude / longitude). Climate data are obtained from a number of sources. Monthly rainfall and temperature data are sourced from the Australian Water Availability Project (AWAP) (Raupach et al. 2009). Soil moisture data are obtained from the Bureau of Meteorology (BoM) Australian Water Resources Assessment Landscape model (AWRA-L) (Frost et al. 2016). Data on hail storms were obtained from the BoM Severe Storms Archive.

A variety of variables are constructed from these sources, for a combination of different climate measures (Table 3) and time periods / seasons (Figure 2) of relevance to Australian broadacre farms.

Table 4: Climate variable measures

Name	Description	Units	Source
rain	Rainfall volume	mm	AWAP
tmax	Average maximum temperature	degrees C	AWAP
tmin	Average minimum temperature	degrees C	AWAP
moist	Root zone soil moisture	index (0-1)	AWRA-L
hail	Exposure to hail storms	index (0-1)	BoM

1.2.2 Price data

Annual commodity price data were also construed on a state and national basis for each FLAD commodity, for the years 2000-01 to 2018-19. This data is sourced primarily from the ABS, based on data used to construct the Value of Agricultural Commodities Produced (VACP) series, with adjustments applied to match the FLAD commodity groupings. Missing values were imputed using commodity price indexes obtained from ABARES. For major livestock activities (Beef, Sheep, Dairy) prices are specified as 'average farm revenue per livestock number' (based on ABARES farm survey data) as the FLAD does not contain data on livestock outputs (i.e., Beef cattle / lamb / wool / milk sold) only livestock holdings.

1.3 BLADE integration

The ABS Business Longitudinal Analysis Data Environment (BLADE) contains data on all active Australian businesses between 2001-02 and 2016-17. BLADE provides data from Australian Tax Office sources including: Business Activity Statements (BAS), Business Income Tax (BIT) filings and Pay as You Go (PAYG) summaries (a detailed description of BAS, BIT and PAYG data items is provided in the BLADE metadata).

The BLADE is based on the ABS Business Register (ABSBR), which provides a longitudinal business frame (based on Australian Business Numbers ABNs) tracking the characteristics of individual business over time. Business entities on the ABSBR fall into one of two groups:

- **Non-profiled population:** Simple business structures (e.g., where the ABSBR unit matches to a single ABN)
- **Profiled population:** Large complex business structures (e.g., where the ABSBR unit maps to multiple ABNs.).

Data for businesses in the profiled population is available at the whole Enterprise Group (EG) level and the lower Type of Activity Unit (TAU) level: each EG can have multiple TAUs (and each TAU can have multiple ABNs).

Farm units in the FLAD target sample between 2005-06 and 2016-17 were linked to the BLADE primarily on the basis of ABNs (linking prior to 2005-06 is difficult as this predates the use of ABNs in agricultural data collections). Approximately 87% of units on FLAD could be matched to a non-profiled business in the BLADE by an ABN. The remaining 13% were matched to the profiled population at the TAU level on the basis of ABN and industry classification (Australian and New Zealand Standard Industrial Classification).

While the majority of FLAD to BLADE links are 'one-to-one', there are some 'many-to-one' links (e.g., where a large business owns multiple farms) and a small number of 'one-to-many' links

(where a single farm has multiple business entities). In total, the 194,184 unique farm units in FLAD from 2005-06 to 2016-17, link to 186,476 unique BLADE units.

1.3.1 Example: farm production (FLAD) and farm business revenue (BLADE)

This section compares estimates of farm production value based on the FLAD with BLADE business revenue measures ('turnover' from the BAS data and 'income' from the BIT data (see Appendix B). Here farm value of production is defined as the sum of production times price for each commodity produced (i.e., the Value of Agricultural Commodities Produced or VACP, see Appendix B).

Differences between farm VACP and business revenue are expected for a variety of reasons. Firstly, business revenue may include non-farming related activities (which could be significant, particularly where farms are owned by large corporate entities). Secondly, farm VACP is only a proxy for farm revenue as some production may be used or stored rather than sold. This is likely to be a significant issue in the case of grain farms, due to both on-farm storage and centralised grain marketing schemes. Further, livestock production is currently only represented approximately, as data on livestock sales is not collected in ABS census or surveys.

Despite this, a high level of correlation is still observed between estimated farm VACP and BLADE revenue for most farms (Table 5). Typically, higher correlations are observed for horticultural farms and cropping farms, with lower correlations found for livestock farms (as would be expected). Both BLADE revenue measures (BAS and BIT based) show similar levels of correlation with farm production.

Figure 4 compares annual farm VACP and business revenue for grain growing farms (for more detail see case study 3). As would be expected BAS revenue estimates are generally higher than BIT as, as BIT estimates used here excludes 'non-primary production' revenue for some businesses. The annual figures show expected effects of climate variability, with a drop during the 2006-07 drought and a spike upward in 2016-17 (a year of high rainfall and crop yields). Business revenue data show less volatility than farm production values (see Case Study 3 for more detail).

Figure 4: Median farm production value and revenue per hectare, 'other grain growing' farms, 2001-02 to 2016-17

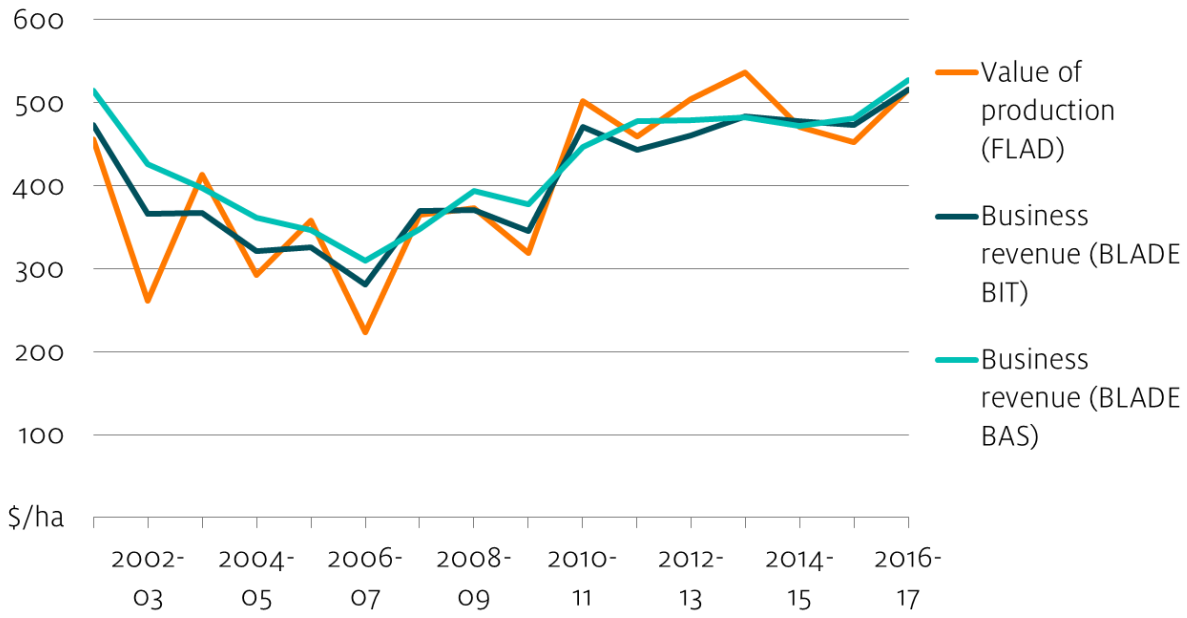


Table 5: Correlation between value of farm production (FLAD) and business revenue (BLADE)

	BLADE business revenue measure	
Industry	**Income (BIT)**	**Turnover (BAS)**
Mushroom	0.82	0.82
Vegetable (undercover)	0.72	0.47
Vegetable (outdoors)	0.50	0.53
Grape	0.35	0.38
Kiwifruit	0.50	0.43
Berry fruit	0.72	0.74
Apple and pear	0.43	0.47
Stone fruit	0.32	0.54
Citrus fruit	0.59	0.69
Olive	0.59	0.35
Other fruit and tree nut	0.76	0.73
Sheep farming	0.24	0.23
Beef cattle farming	0.47	0.30
Beef cattle feedlot	0.42	0.54
Sheep-beef cattle farming	0.32	0.31
Grain-sheep or grain-beef farming	0.59	0.52
Rice growing	0.52	0.52
Other grain growing	0.67	0.73
Sugar cane	0.69	0.68
Cotton	0.62	0.69
Other crop	0.37	0.35

Dairy cattle farming	0.50	0.51
Poultry farming (meat)	0.17	0.14
Poultry farming (eggs)	0.61	0.55
Deer farming	0.36	0.69
Horse farming	0.02	0.02
Pig farming	0.72	0.59
Beekeeping	0.18	0.12
Other live stock farming	0.18	0.23

Note: excludes one-to-many and many-to-one FLAD/BLADE links, profiled / enterprise group BLADE units and other outliers.

1.3.2 Approximating the farm population

While the BLADE administrative data provides complete coverage of all Australian businesses (at least between 2001-02 and 2016-17) the sample size of FLAD fluctuates greatly between years (with around 90% coverage in census years and around 20% in survey years). For this project the FLAD and BLADE data were combined to generate an approximate farm business register, which attempts to represents the full population of farms operating in each year. Further detail on this register is provided in Appendix A.

This method identifies all active businesses on BLADE in each year associated with agriculture. In each year, agricultural data for any unobserved farm units is approximated based the nearest observation of that farm in the FLAD. For example, a farm unit which appears in FLAD in the 2005-06 census year but is not sampled in 2006-07 (and is linked to a BLADE unit which remains active in 2006-07) could be included on the register in 2006-07.

While this assumption-based approach involves some approximation, it is able to generate a realistic farm population that results in national aggregates comparable to those produced in public ABS agricultural statistics (see Appendix A). While further testing is required, this approach has some potential advantages over the traditional statistical weighting approaches used by the ABS to produce population estimates. In particular, it could enable higher resolution (small region) estimates to be generated on an annual basis (a possibility we explore later in this report).

1.4 ABS *datalab* and deidentification

For this project, deidentified versions of the above FLAD-BLADE (and climate/price) data were constructed for use within the ABS secure *datalab* environment. This involved the removal of identifying variables including ABN numbers and farm locations (latitude and longitude). Farm locations were replaced with geocoding to the ABS SA1 level. Further information on the processes for accessing *datalab* is available on the ABS website.

2 Model development

As part of this project, the FLAD/ BLADE datasets were used to develop a number of statistical models. These models can predict farm level outcomes, including agricultural production and farm revenue, given information on climate conditions (e.g., rainfall and temperature), commodity prices and farm characteristics (location, size etc.). A non-parametric machine learning methodology is applied to develop these models, similar to that applied by ABARES for the *farmpredict* model (Hughes et al. 2019). When combined with the approximate farm register (Appendix A) these models have the potential to generate custom results for essentially all farms in Australia.

The FLAD/BLADE data sets enable analysis of a wide range of agricultural regions / industries; however for this initial project modelling has been limited to two farming sectors:

- **Crop farm model:** Australian broadacre ('dry-land') cropping farms
- **Irrigation farm model:** irrigation farms in the Murray-Darling Basin

There remains potential to expand these models to cover a wider range of activities, regions and variables. These models have a range of possible applications including farm forecasting, drought and climate change policy analysis, financial sector (insurance) and ABS statistical applications. Some of these are explored with case studies in section 3. Further options for future research are discussed in section 4.

A brief overview of the current models is provided below. For a detailed description of the methodology see Appendix B.

2.1 Crop farm model

The crop farm model predicts the area planted, production and value for major Australian broadacre crops (such as wheat and barley). The crop production model is a data-driven predictive model estimated using historical data on crop production from the FLAD, along with linked climate and price data. This model predicts area planted and crop yield at a farm level for a range of crops (Table 6). These predictions are combined with commodity price data to estimate farm value of production.

Table 6: Crop farm model overview

Model inputs	Model outputs
Farm characteristics	Area planted (*A*), production (Q), value (*V*)
Location	Wheat
Land area / type of land use	Barley
Livestock holdings	Oats
Fruit / nut trees	Sorghum
Industry	Triticale
Climate variables	Maize
Rain	Canola
Soil moisture	
Temperature	
Commodity prices	

2.2 Irrigation farm model

The irrigation farm model focuses on irrigated farms in the Murray-Darling Basin, and involves a statistical model which predicts water use and production of key irrigated crops (*Rice, Cotton, Grapes, Almonds, Oranges*), given information on farm characteristics, climate conditions and commodity prices.

Table 7: Irrigation farm model overview

Model inputs	Model outputs
Farm characteristics	Water use (*W*), production (Q), value (*V*)
Location	Rice
Land area / type	Cotton
Livestock holdings	Almonds
Fruit / nut trees	Grapes
Industry	Oranges
Climate variables	
Rain	
Soil moisture	
Temperature	
Commodity prices	
Water market prices	

3 Results

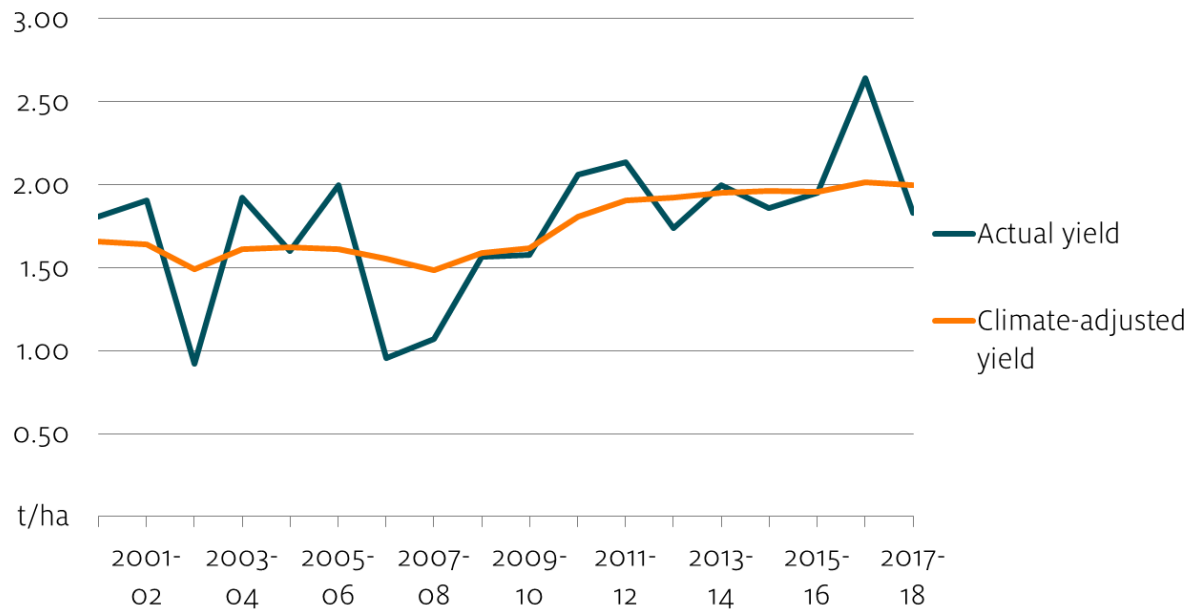
To demonstrate the potential of the FLAD / BLADE datasets and related statistical models five case-studies were developed. These case-studies focus on the effects of drought and climate variability on farm outcomes similar to recent ABARES research (particularly Hughes et al. 2017, Hughes et al. 2019, Hughes et al. 2020). However, the case studies remain illustrative and, in each case more research would be required to confirm, analyse and expand on the results. Further, detail on the methods and assumptions applied in these case studies is presented in Appendix C.

3.1 Case study 1: Trends in Australian crop production

Measurement of long-term crop production trends in Australia is complicated by the effects of climate variability and climate change. A number of previous studies (see Hochman et al. 2017, Hughes et al. 2017) have applied different types of crop production models to control for these effects and produce 'climate adjusted' estimates of Australian wheat yields. This research has shown significant growth in Australian wheat yields over the last 20 years after controlling for the negative effects of hotter and drier conditions (Hochman et al 2017, Hughes et al 2017). Evidence has also emerged of shifts in the location of crop activity across Australia, in response to the changing climate (see Hughes et al. 2017).

The new FLAD dataset allows for a more detailed consideration of these crop yield and area trends. Here the crop production model (see Appendix B) was applied to estimate 'climate adjusted' crop yields and areas (yields and areas under average climate conditions based on the 2000-01 to 2017-18 period, see Appendix C). Figure 5 below shows average annual climate adjusted wheat yields from 2000-01 to 2017-18. These results are similar to Hochman et al. (2017) and Hughes et al. (2017) with strong growth in yield between 2008 and 2015.

Figure 5: Climate adjusted wheat yield 2000-01 to 2017-18



The larger sample sizes in FLAD enable higher resolution results to be generated, showing differences in wheat yield trends by region (Figure 6). Strong growth in wheat yields can be seen in a number of WA and SA regions, while limited gains in wheat yield are observed in northern Vic and South-West NSW (Figure 6). Trends for other major crops are summarised in Table 8.

Figure 6: Growth in climate adjusted wheat yield (2000-01 to 2017-18) by region

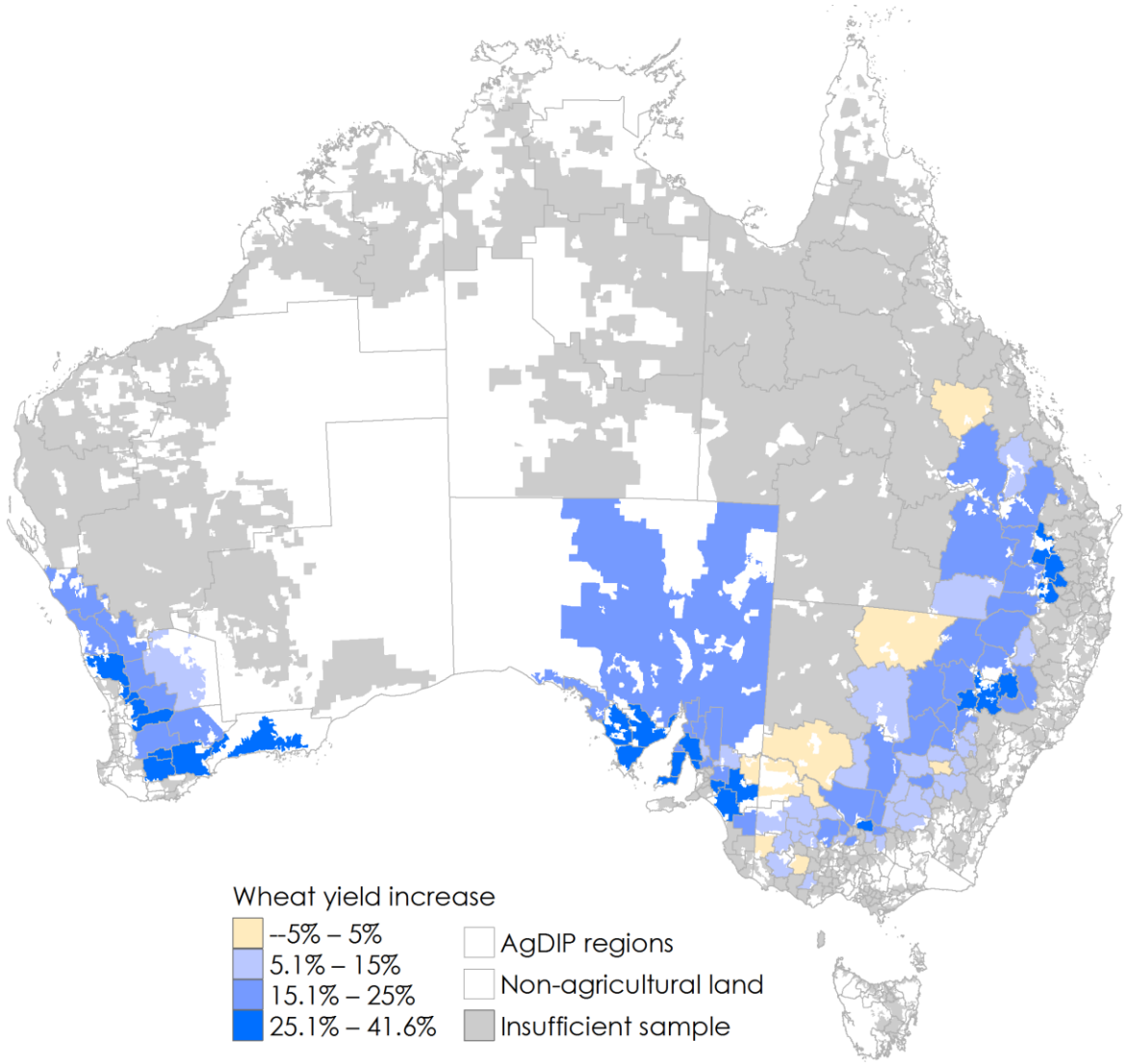


Table 8: Climate adjusted crop production trends

Crop	Climate adjusted area ('000 ha)			Climate adjusted yield (t / ha)		
	2000-01	2017-18	% change	2000-01	2017-18	% change
Wheat	12,273	11,206	-8.7%	1.66	2.00	20.7%
Barley	3,579	4,347	21.5%	1.72	2.22	29.4%
Canola	1,132	2,843	151.2%	1.14	1.24	8.3%
Maize	76	54	-28.9%	4.40	6.90	56.8%
Oats	683	858	25.7%	1.48	1.57	6.7%
Sorghum	739	573	-22.5%	2.55	3.21	25.8%
Triticale	386	55	-85.6%	1.68	1.69	0.7%

The period 2000-01 to 2017-18 has seen some significant changes in crop areas planted, including strong growth in Canola (Table 8). Regional level crop area trends are summarised in Figure 7 and Figure 8. Nationally, there was a slight decline in wheat areas over the period, partly due to substitution towards canola crops (particularly in Western Australia, see Figure 7). A number of regions, have seen increases in total broadacre crop areas over the period (due to substitution away from livestock farming) particularly in WA and southern Vic (Figure 8). However, most regions have seen decreases in cropping activity since 2008-09 (Figure 8).

Figure 7: Change in (climate adjusted) area planted to wheat (2000-01 to 2017-18) by region

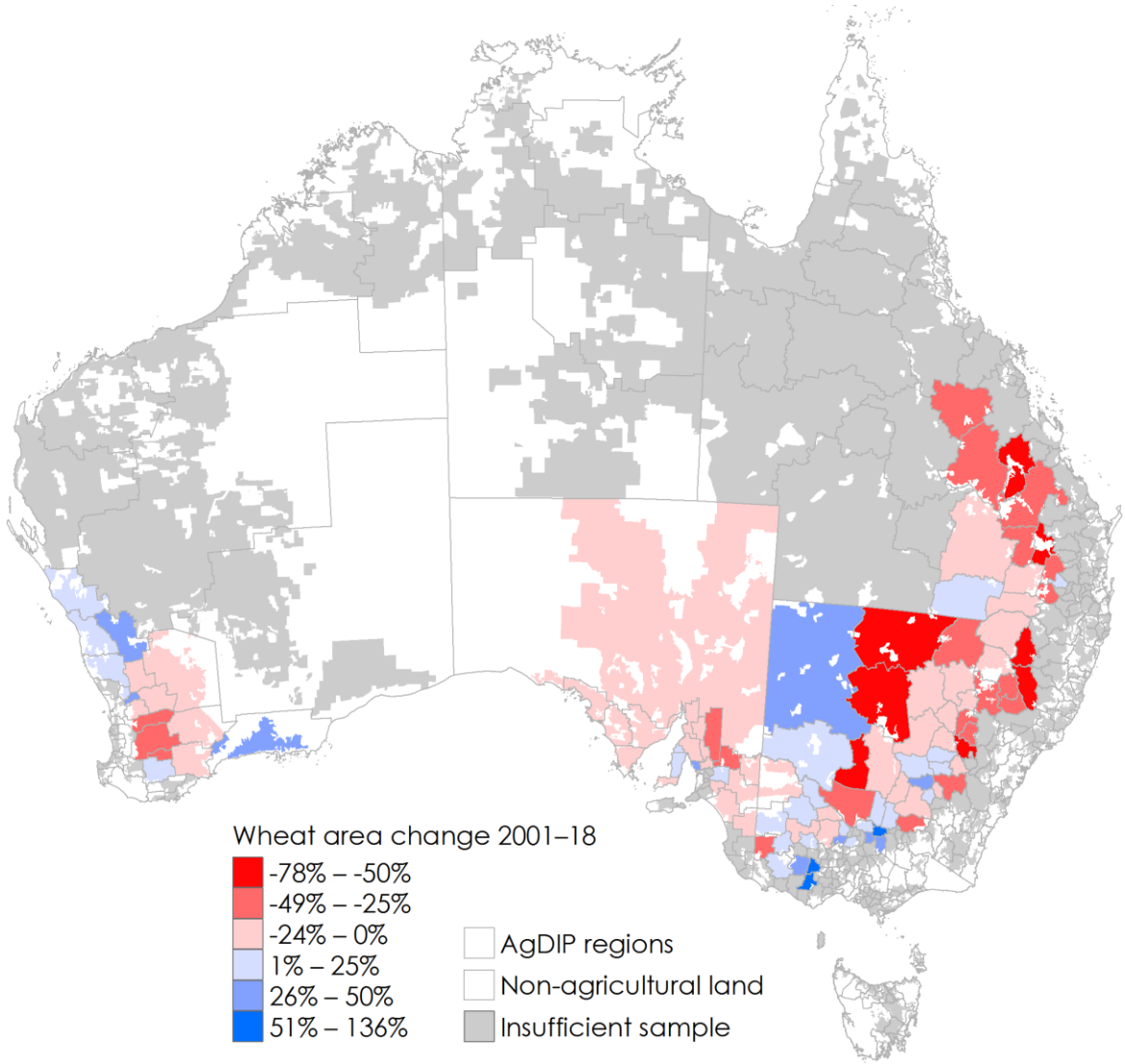
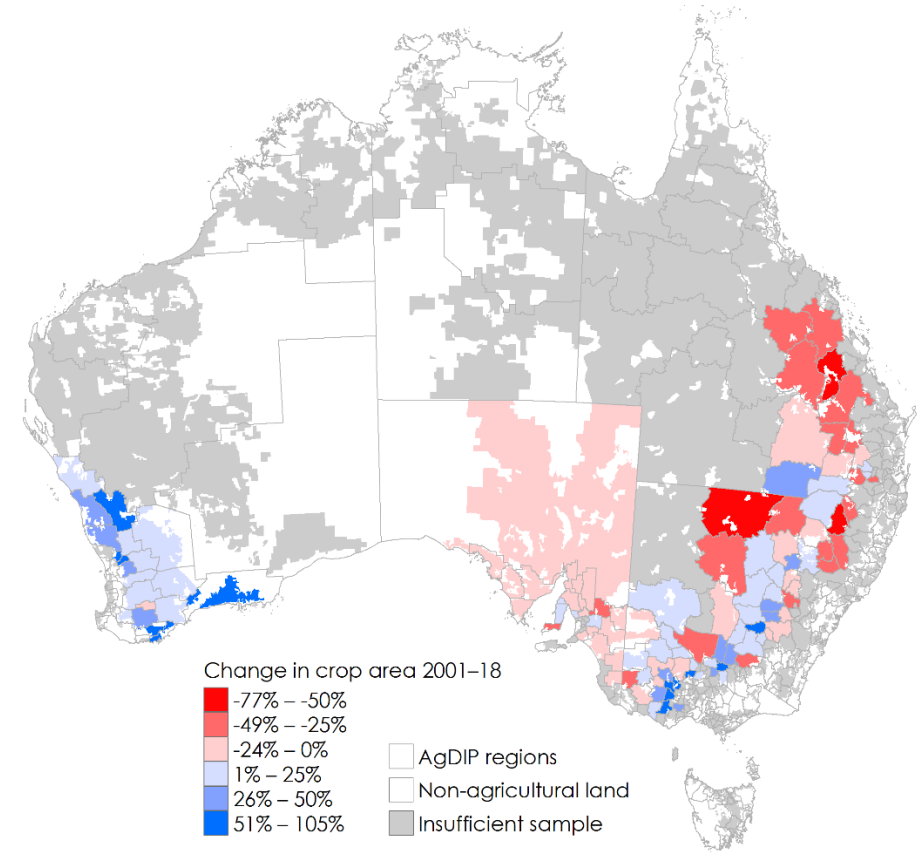
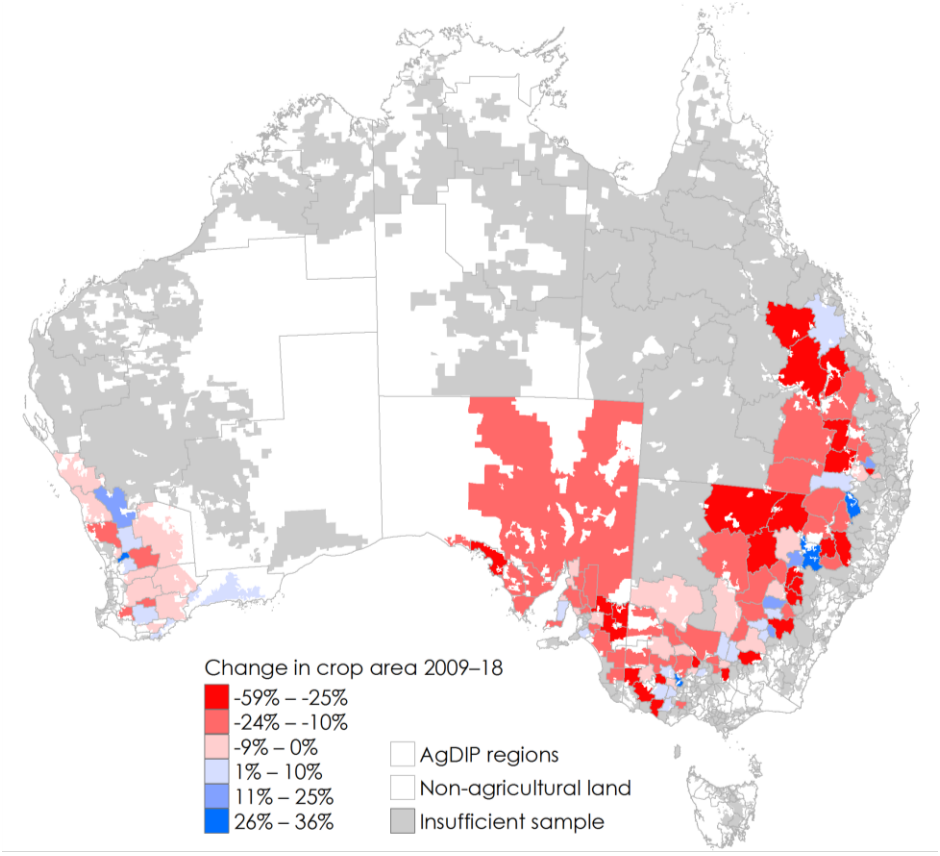


Figure 8: Change in (climate adjusted) area planted to all broadacre crops by region

2000-01 to 2017-2018



2008-09 to 2017-2018



3.2 Case study 2: Small area statistics for WA wheat

Under current ABS methods, regional crop production data are limited, with small regions (ABS SA2) only available in census years (every 5 years). In addition, regular changes in ABS region boundaries make it difficult to construct consistent regional time-series data. Further, the confidentialisation approach applied by the ABS to aggregate statistics can limit the usefulness of some regional data.

In this case study a set of experimental small region data are developed detailing wheat production in Western Australia between 2000-01 and 2017-18. This data set is based on simulated wheat production data generated from the crop production model. Here the approximate farm register (see Appendix A) is applied in order to simulate crop production for the entire farm population. This approach allows for census like sample coverage in all years (by generating simulated results for non-surveyed farms). This model-based approach provides an alternative to the survey weighting methods, normally used to by the ABS to 'scale up' sample data to population estimates. Further detail on the methodology is provided in Appendix C.

Figure 9 demonstrates the small region dataset for WA wheat in 2017-18, with results for each AgDIP region contrasted against the larger NRM regions used in the published ABS statistics for that year (Figure 10). The full dataset (containing wheat area and production by AgDIP region for WA from 2000-01 to 2017-18) is available as a spreadsheet.

Figure 9: WA wheat yields by region, 2017-18 (AgDIP regions)

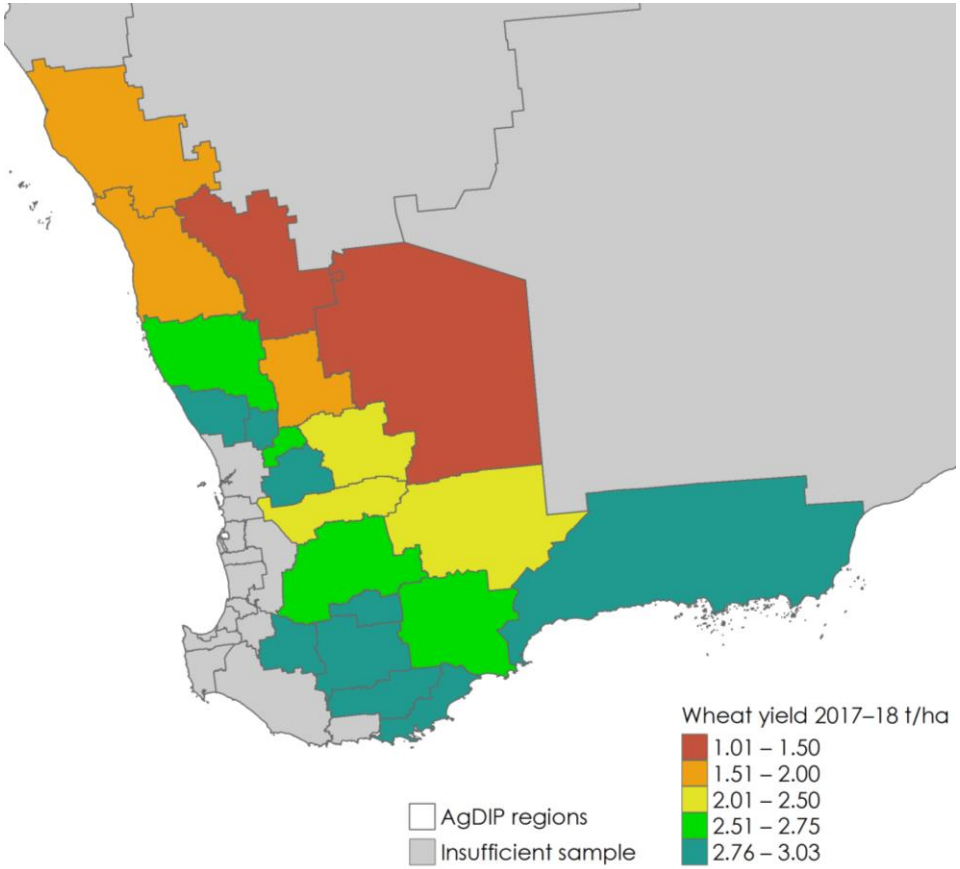
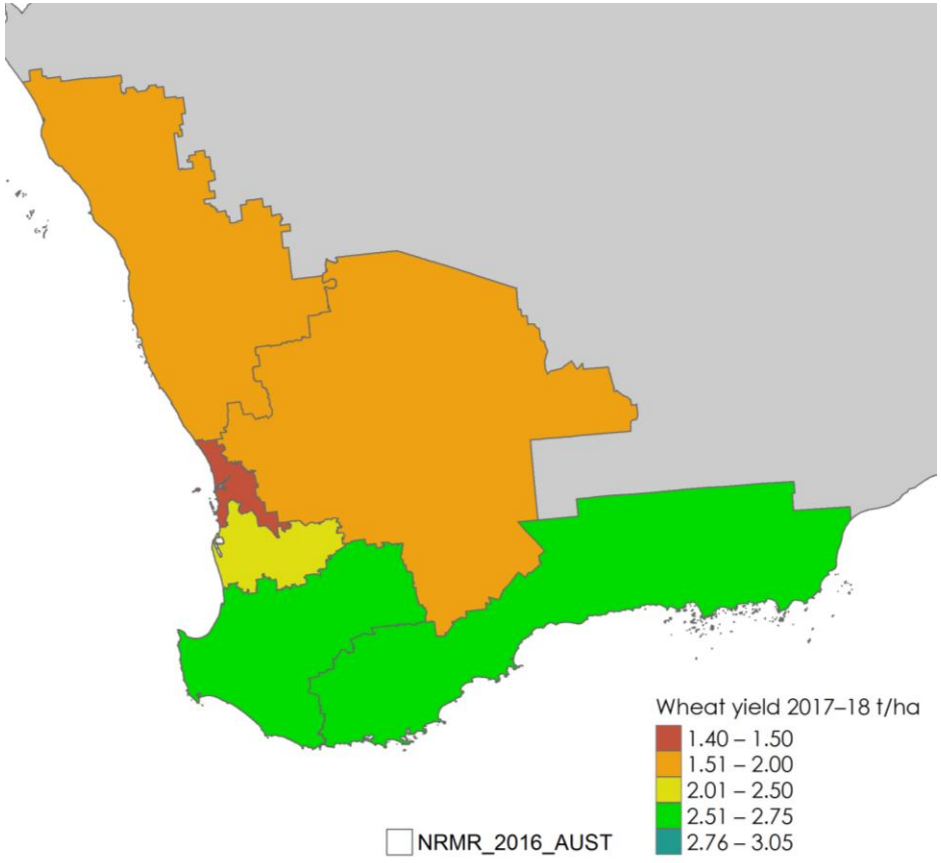


Figure 10: WA wheat yields by region, 2017-18 (ABS NRM)

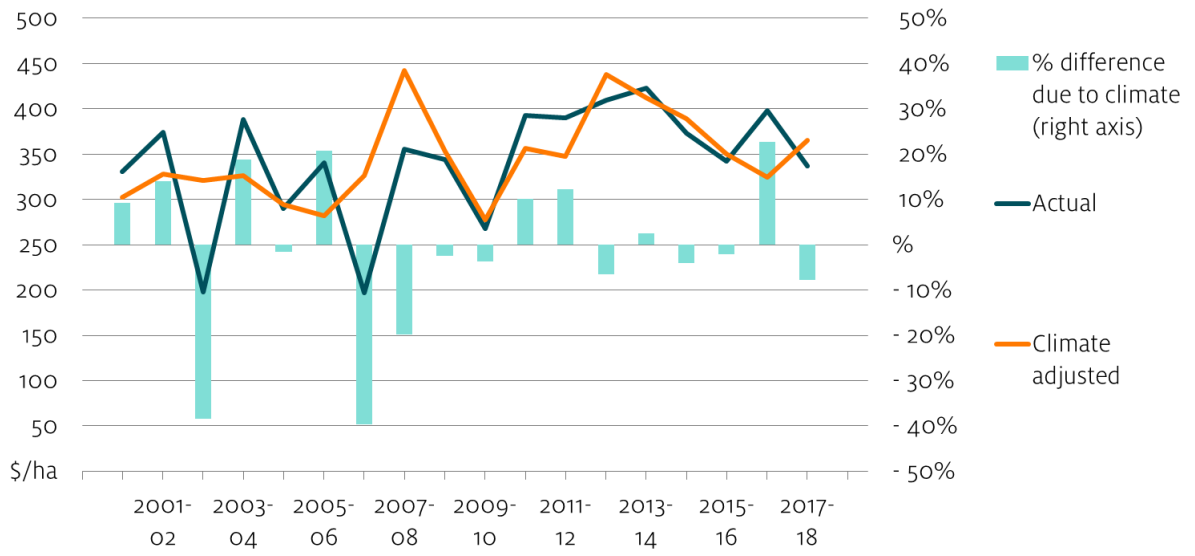


3.3 Case study 3: Effects of drought on cropping farms

Drought conditions can lead to dramatic reductions in production and profits for affected cropping farms. The crop production model developed in this study can be applied to isolate the effects of drought on farm crop production from other key factors, such as commodity price changes, over the period 2000-01 to 2017-18.

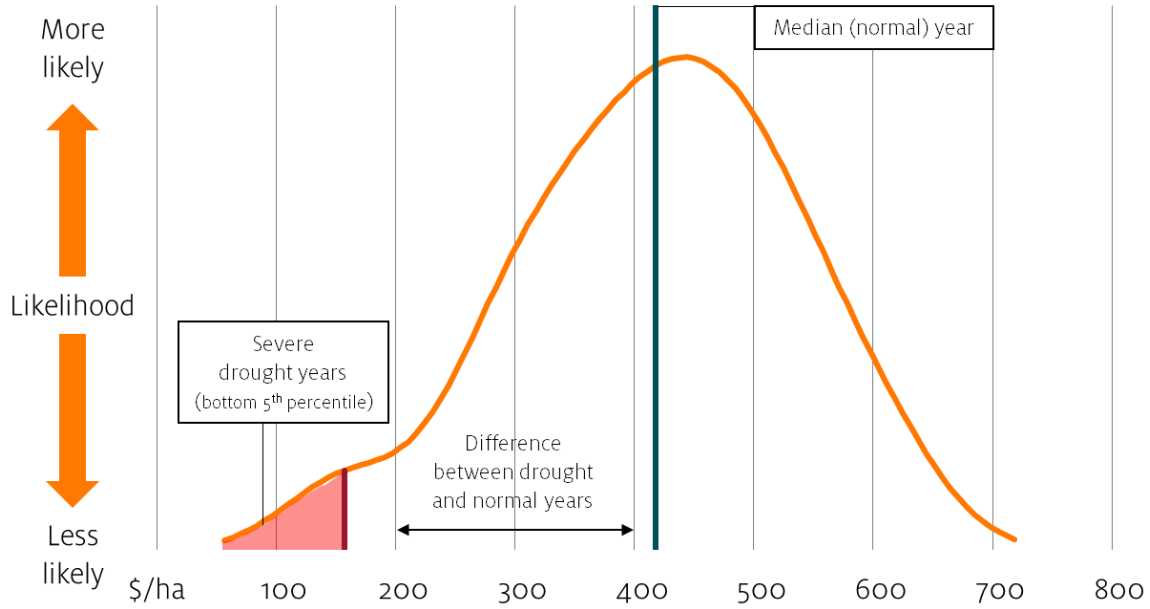
Figure 11 shows the average value of crop production both before and after adjusting for climate variability. The percentage difference between the two series is also shown, which highlights the effect of climate on crop production. For example, in the 2002-03 and 2006-07 drought years, the average value of crop production per hectare was reduced by around 40 per cent. These drought impacts reflect the combined effect of changes in yields and areas planted for each of the crops included in the model (wheat, barley, sorghum, canola, oats, triticale, maize).

Figure 11: Effect of climate on the average value of crop production, 2000-01 to 2017-18



These average industry wide effects understate the effects of drought on affected individual farms. Figure 12 shows the climate related variation in crop production value per hectare for a typical Australia cropping farm. Here production value declines by over 60% between a 'normal' (median) climate year and a severe drought (based on the period 2000-01 to 2017-18).

Figure 12: Effect of climate variability on production value per hectare for a 'typical' cropping farm



The extent of these drought effects varies across different farming types and regions, depending on the types and amounts of crops grown, farm management practices and the variability of the local climate. Figure 13 shows the relative sensitivity of farm production value to climate variability by region (the percentage decrease between a median and 5th percentile 'drought' year). These results show higher drought sensitivity for cropping farms in north-western NSW and QLD and lower sensitivity in WA.

Figure 14 compares median annual farm production value with median annual business revenue (based on BLADE BIT data, see Appendix C) for Australian cropping specialist (grain growing) farms. Here we can see that farm business revenue shows less volatility than farm production value. This is expected given the various income smoothing opportunities available to farm businesses including, for example, centralised crop marketing schemes (where farm crop revenue is smoothed across multiple years) and non-farm revenue sources (e.g., non-farm business activity, government drought assistance payments).

Figure 13: Sensitivity of crop production value to drought by region

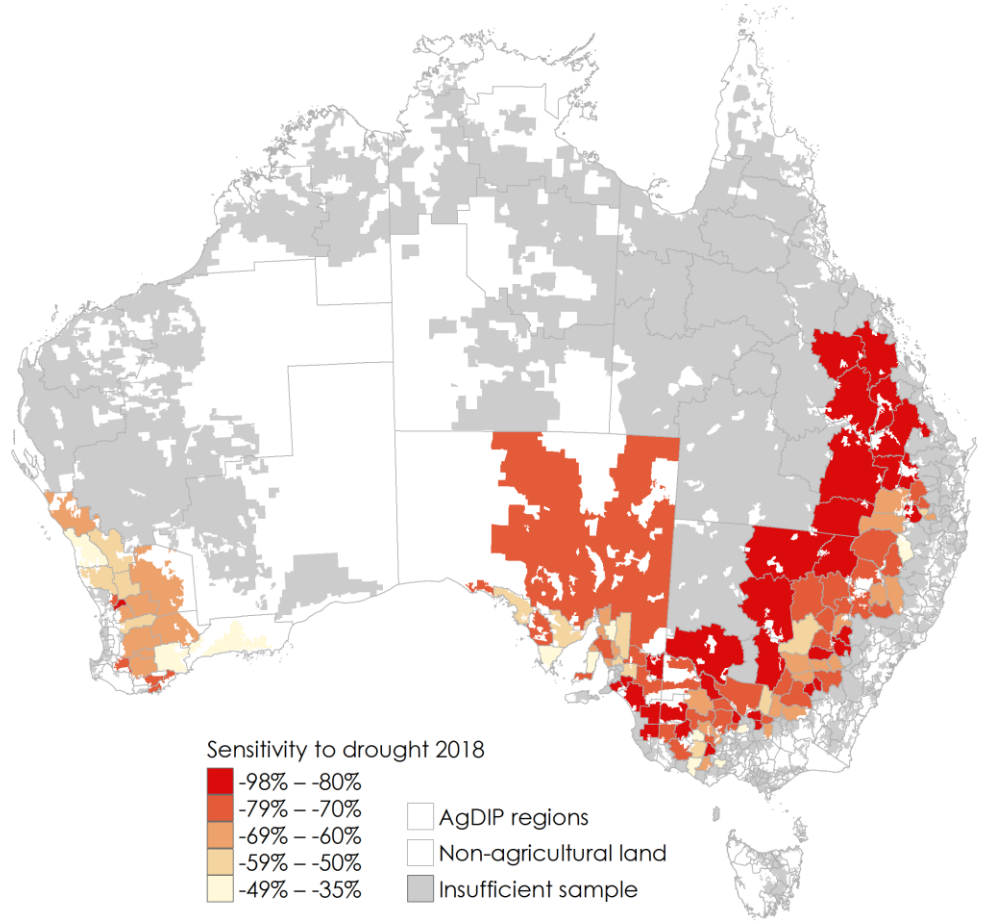
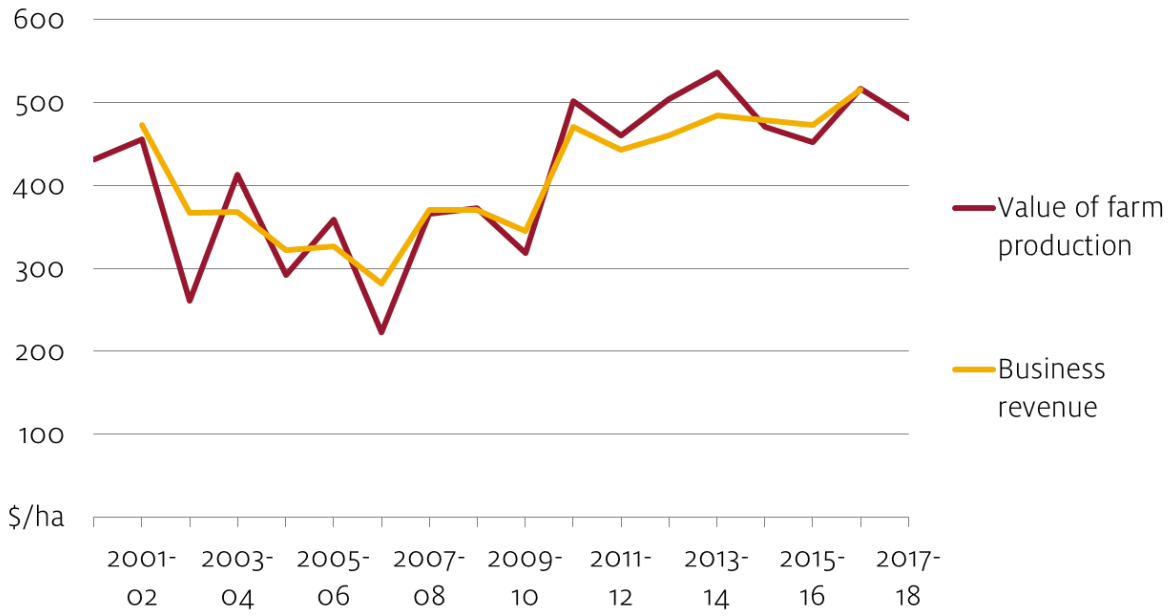


Figure 14: Median cropping farm production value and business revenue per hectare, 2000-01 to 2017-18



3.4 Case study 4: Index-based drought insurance for cropping farms

In recent years there has been growing interest in the potential for drought insurance products to mitigate farm income variability, and in-turn help to limit the demand for government drought assistance (see Hatt et al. 2012, Hertzler 2005, Hughes 2018). Of particular interest are index-based or parametric insurance products where pay-outs are tied directly to weather data.

This case study presents some scenario results for a hypothetical index-based insurance product designed specifically for Australian cropping farms. This scenario remains purely illustrative (requiring a number of strong simplifying assumptions, see Appendix C) and further research would be required to assess the feasibility of this approach (which in practice would depend on whether the benefits of this risk mitigation are larger than the premiums required by insurers to cover potential pay-outs) However, the results serve to at least illustrate the potential size of payouts and demonstrate how insurance can act to minimise variability in farm incomes[1].

Under this hypothetical insurance scheme (for details see Appendix C) the crop production model is used to develop an index measuring the effect of climate variability on the value of farm production per hectare of cropping land. The resulting 'multi-variate' index insurance provides cropping farms with protection from climate related variation in farm revenue (but not from price related variation). This insurance provides simultaneous coverage of both yield and area variation for all major crops (as such this form of insurance is designed to be held over the long term, regardless of the amount / mix of crops a farm grows in a given year).

Figure 1 shows the simulated total insurance pay-outs under the assumption that this insurance is held by every cropping farm in Australia (with a 20% percentile payout threshold, see Appendix C). As would be expected the largest payouts occur in the 2002-03 and 2006-07 drought years.

Further, detail on the simulated pay-outs are presented in Figure 15Figure 16 and Table 9. Nationally pay-outs over the entire period average to $18.70 per hectare of broadacre crops planted, with significant differences over time and across regions. Regional differences in average payouts reflect a combination of factors (farm sensitivity to drought and typical crop yields / mix). Generally, higher pay-outs per hectare are observed in regions with higher crop yields.

Figure 17 illustrates how holding this hypothetical index-based insurance can reduce volatility in cropping farm production value and revenue, particularly decreases in drought years. As discussed, observed farm business revenue data contain limited variability (even before applying the simulated insurance payouts) as these data include various forms of self-insurance (and government drought assistance) which farms turn to in the absence of a viable insurance

[1] In practice, the availability of insurance may enable farmers to change their production systems and accept more risk. In this case the benefits of insurance may be realised as higher productivity and profit levels rather than reductions in income volatility (see Hughes et al. 2020).

market. Table 9 compares farm revenue volatility outcomes by state, showing that the biggest reduction in farm revenue volatility (due to the simulated insurance payouts) are achieved for Victorian cropping farmers.

Figure 15: Simulated total annual insurance pay-outs, 2000-01 to 2017-18

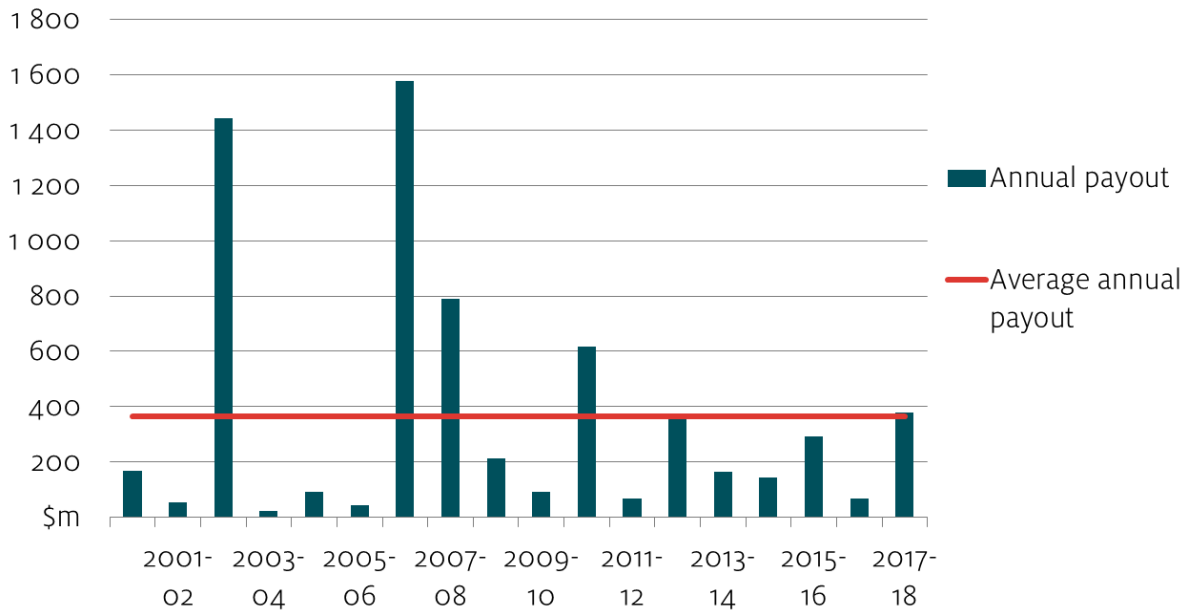


Table 9: Simulated crop revenue insurance outcomes by state

Region	Insurance payouts		Annual farm income volatility b					
	Average per ha a	Average per year	Farm production value			Business revenue		
	$ / ha	$m	No Ins.	Ins	Change	No Ins.	Ins	Change
NSW	21.0	115.50	27.6%	22.3%	-5.4%	14.7%	12.6%	-2.1%
Vic.	16.8	48.16	46.3%	36.3%	-10.0%	27.3%	21.7%	-5.6%
QLD	20.0	27.85	25.2%	20.0%	-5.2%	16.1%	13.9%	-2.3%
SA	17.4	60.21	27.7%	21.0%	-6.8%	13.5%	10.8%	-2.7%
WA	16.0	112.70	29.1%	23.8%	-5.3%	8.3%	10.6%	2.2%
Australia	18.7	365.21	23.3%	18.9%	-4.4%	10.1%	8.7%	-1.5%

a Average pay-outs (2000-01 to 2017-18) relative to total broadacre crop area planted.
b Annual farm income volatility is measured as the mean absolute annual deviation in median farm production value / business revenue.

Figure 16: Simulated average insurance pay-outs per hectare (total crop area planted) by region (2000-01 to 2017-18)

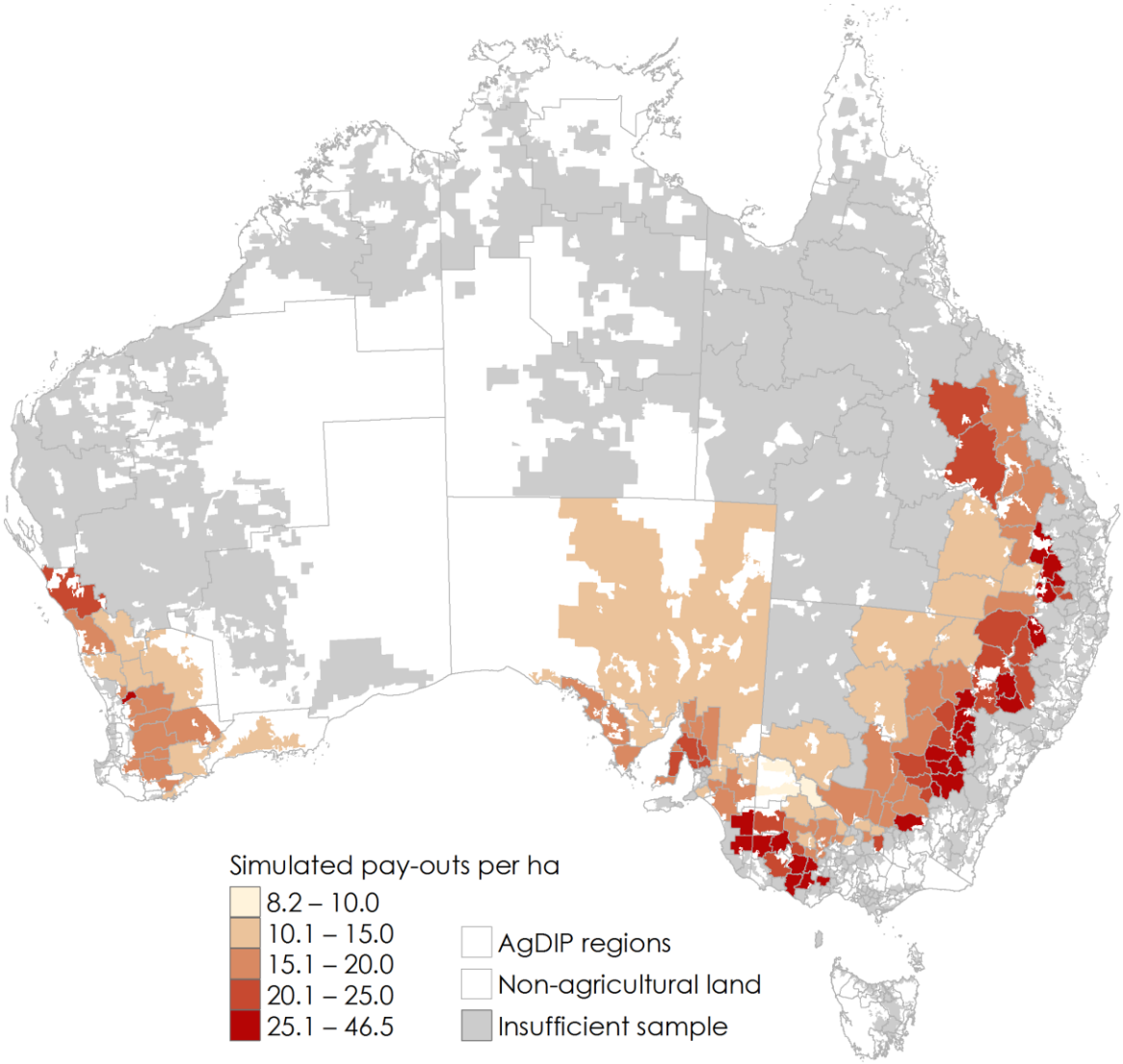
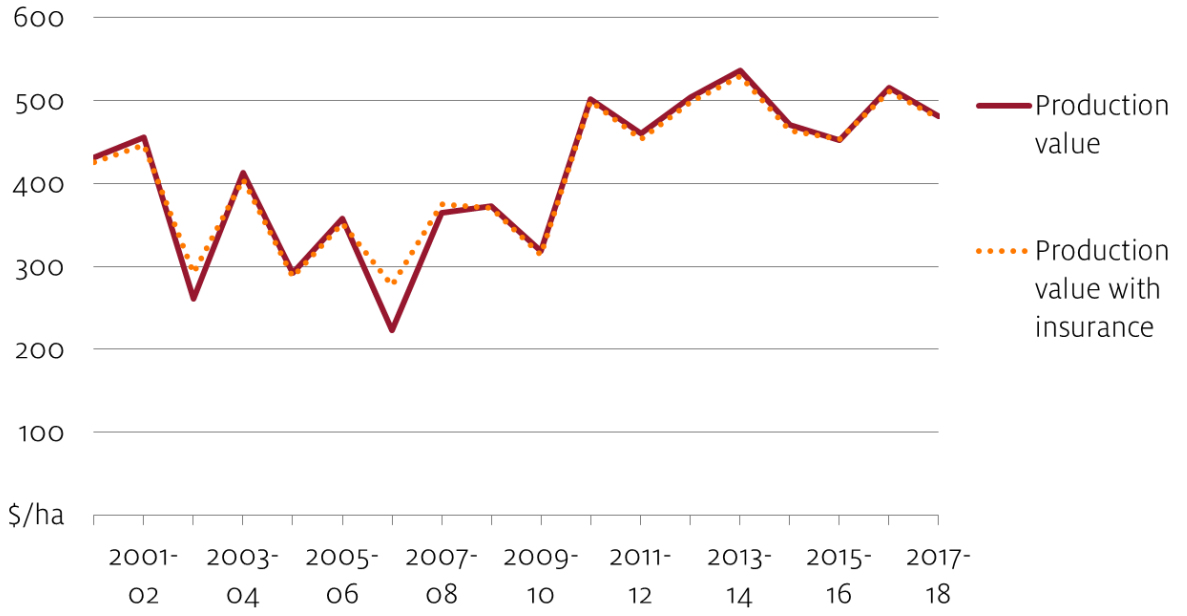
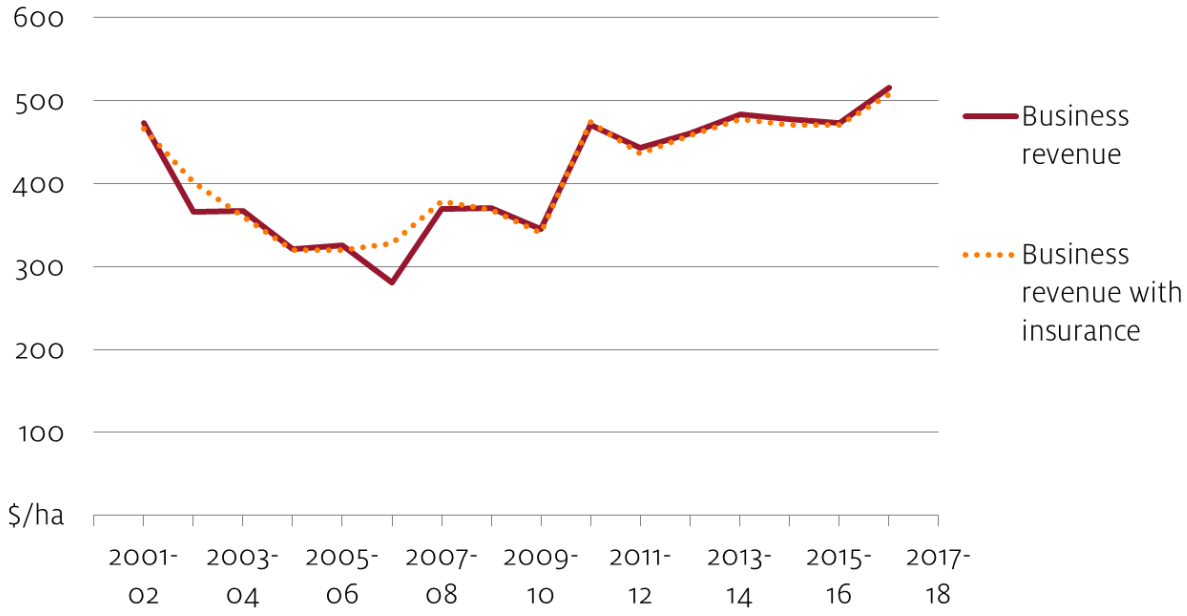


Figure 17: Median cropping farm revenue per hectare with simulated insurance pay-outs, 2000-01 to 2017-18

Farm production value



Farm business revenue



3.5 Case study 5: Water productivity in the Murray-Darling Basin

Water productivity refers to amount of crop produced per unit of water use on irrigation farms (i.e., 'crop per drop'). Given increasing water scarcity in the Murray Darling-Basin (MDB) and significant government investment in irrigation infrastructure (such as the on-farm irrigation efficiency program) there is much interest in measuring water productivity trends. However, irrigation activity in the MDB is complex and influenced by a range of factors, including weather, commodity markets and water policy (Goesch et. al. 2020), making it difficult to observe underlying trends in productivity.

In this case study, predictive models are applied to measure trends in irrigated farm water productivity, after controlling for external factors such as climate variability and water prices. Climate adjusted yield, application rate and water productivity (see Appendix B) are estimated for a range of irrigated commodities – rice, cotton, grapes, almonds and oranges.

In this section, Figures 18-20 show results for rice farms in Australia (predominantly located in the southern connected Murray-Darling Basin). Under average climate conditions (based on the 2000-01 to 2017-18 period), there is a marginal decrease in the application rate and an increase in the yield. As a result, there is growth in water productivity for rice farms.

Increases in water productivity are also observed for all other commodities (Figure 21). Cotton water productivity in the southern MDB is below that of the northern MDB, but has shown higher water productivity growth in recent years: increasing water productivity by around 8 per cent between 2011 and 2018.

Figure 18: Rice water application rate (climate adjusted) 2003-04 to 2017-18

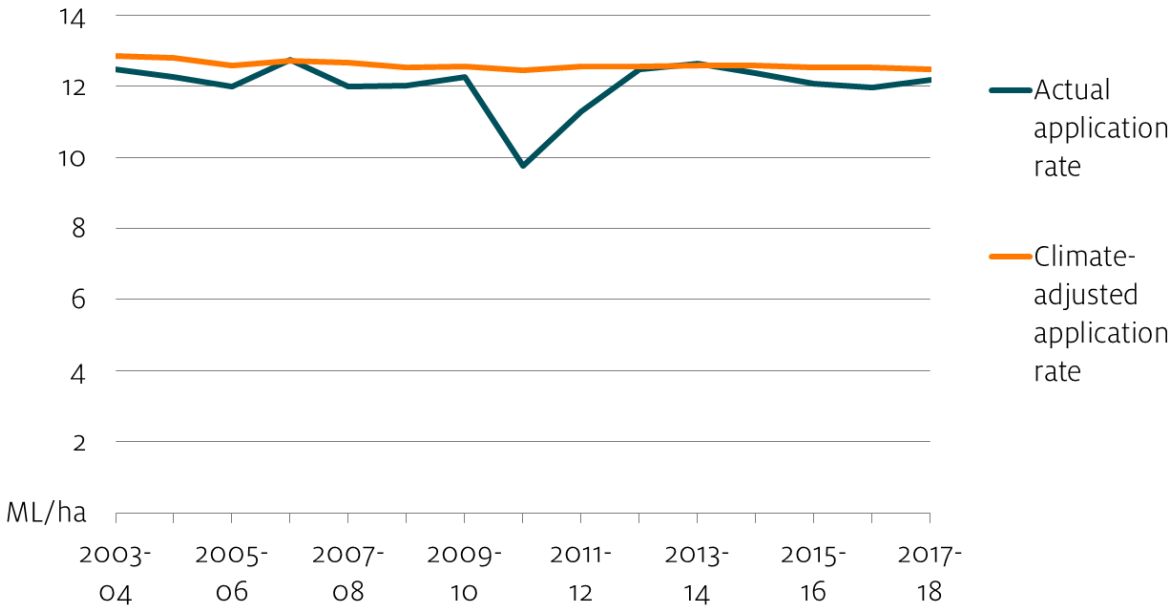


Figure 19: Rice yield (climate adjusted) 2003-04 to 2017-18

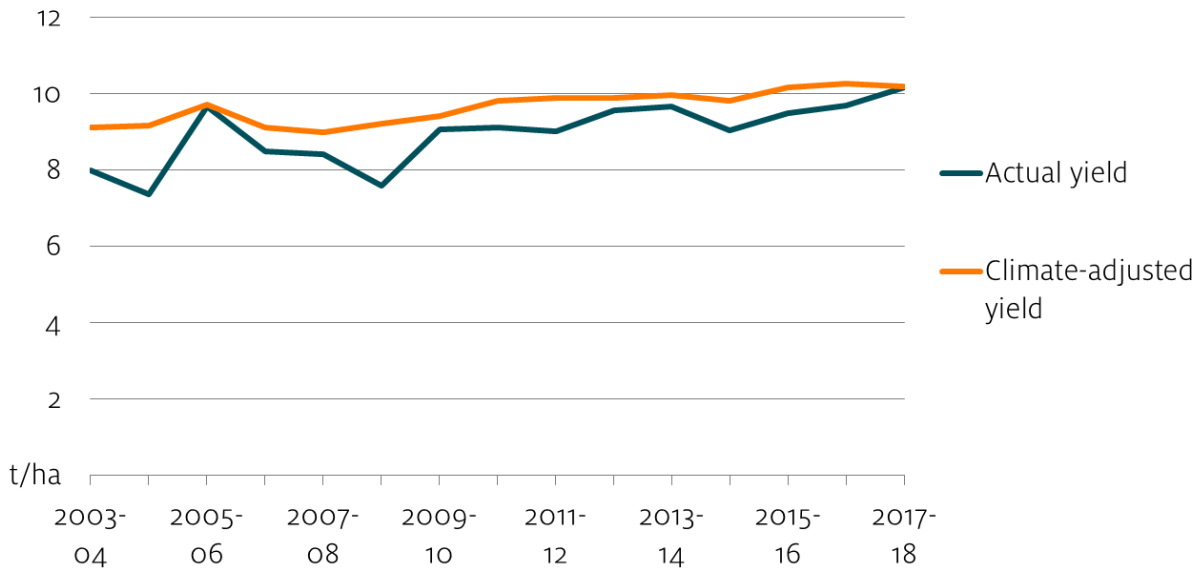


Figure 20: Rice water productivity (climate adjusted) 2003-04 to 2017-18

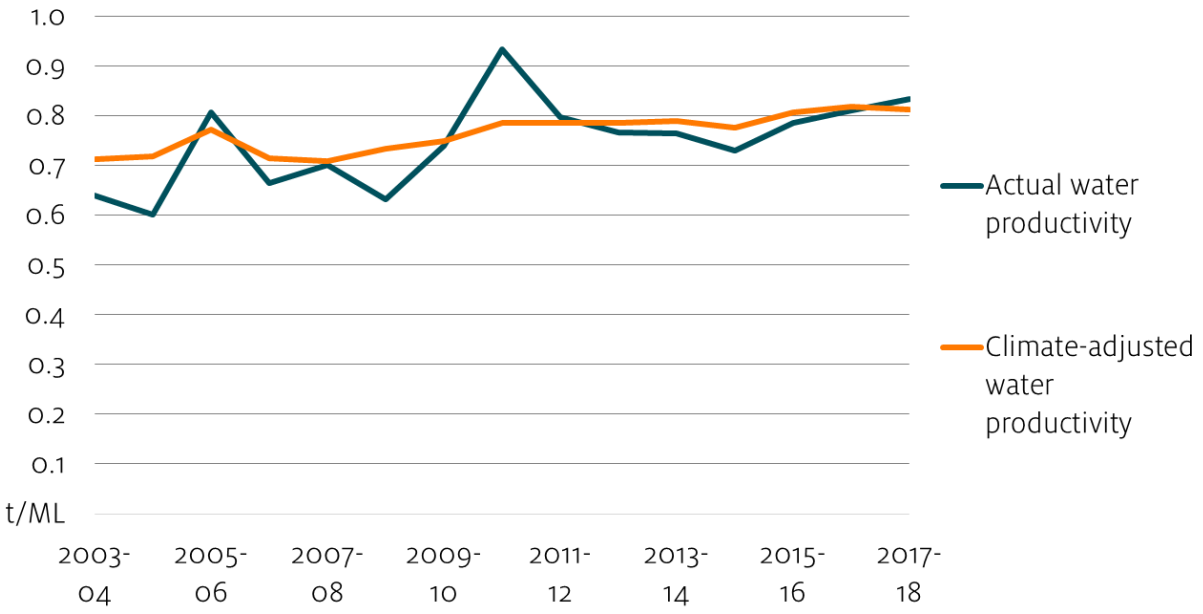


Table 10: Climate adjusted water productivity results by activity

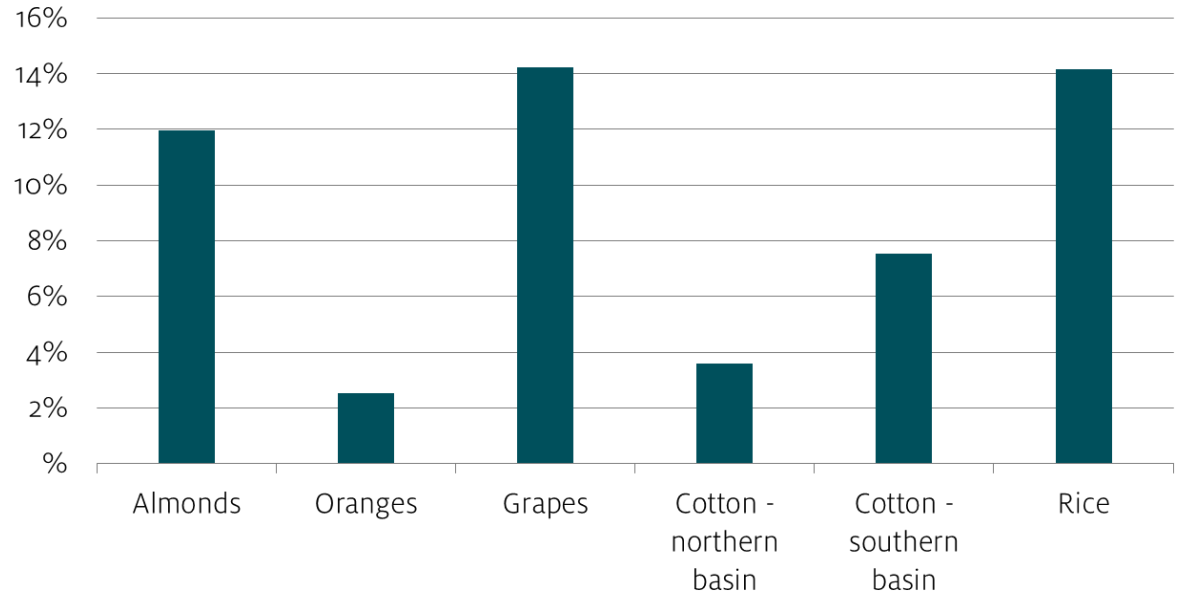
		Application rate		
Irrigation activity	**Units**	**Start year**	**2017-18**	**% change**
Almonds	ML/ha	10.88	9.62	-11.6%
Oranges	ML/ha	7.44	7.48	0.4%
Grapes	ML/ha	4.19	4.20	0.2%
Cotton - northern basin	ML/ha	6.26	6.86	9.7%
Cotton - southern basin	ML/ha	8.93	8.97	0.5%
Rice	ML/ha	12.85	12.49	-2.8%

Table 10: Climate adjusted water productivity results by activity (continued)

		Yield		
Irrigation activity	**Units**	**Start year**	**2017-18**	**% change**
Almonds	kg/tree	8.42	7.15	-15.1%
Oranges	kg/tree	61.65	61.84	0.3%
Grapes	t/ha	11.62	13.24	13.9%
Cotton - northern basin	t/ha	2.18	2.39	10.1%
Cotton - southern basin	t/ha	2.13	2.31	8.7%
Rice	t/ha	9.10	10.17	11.7%
		Water productivity		
Irrigation activity	**Units**	**Start year**	**2017-18**	**% change**
Almonds	t/ML	0.19	0.21	12.0%
Oranges	t/ML	2.63	2.70	2.5%
Grapes	t/ML	2.75	3.14	14.2%
Cotton - northern basin	t/ML	0.34	0.35	3.6%
Cotton - southern basin	t/ML	0.24	0.26	7.5%
Rice	t/ML	0.71	0.81	14.1%

Note: Start year for almonds, oranges, grapes and rice is 2003-04. Start year for cotton - northern basin is 2004-05. Start year for cotton - southern basin is 2010-11. For more detail see Appendix C.

Figure 21: Percentage change in climate adjusted water productivity by activity 2003-04 to 2017-18



4 Future development and applications

4.1 Data

4.1.1 FLAD

There are a number of opportunities for further development of the FLAD-BLADE datasets, including the improvement of data quality and the addition of new data as it becomes available.

In particular there remains scope for further refinement of farm geocoding. Historical location data could be improved by revisiting farm address geocoding using the latest and best available sources (for example, the latest version of GNAF). In the longer-term, new farm property boundary information collected by the ABS could greatly improve the accuracy of farm geocoding, making it easier to link the FLAD with spatial datasets and satellite images.

While the current FLAD contains the vast majority of agricultural data collected by the ABS, there remain a few specialised variables which could be added to the dataset in future if a particular research demand required it, such as land management practice information (which has been included in ABS collections in some years).

4.1.2 BLADE integration

In the future, the integration of BLADE could be extended beyond FLAD to the larger ABS agricultural business 'frame'. This frame contains the full population of agricultural businesses (or at least the ABS's best approximation of it) and is used by the ABS to design (and weight) sample surveys. Linking the BLADE directly to the frame would greatly simplify the development of the approximate farm business register (detailed in appendix A) and should allow for more accurate population estimates.

The AgDIP will also benefit from general updates to the core BLADE datasets, including recent enhancements (such as better accounting for GST inclusive/ exclusive status in BAS data) which were not available in the version of BLADE used in this project.

4.1.3 Comparing FLAD / BLADE with ABARES farm survey data

While the FLAD contains information on a wide range of commodities it is subject to some key gaps relative to ABARES farm survey data (at least for the dairy and broadacre farming sectors). In particular, the FLAD contains limited information on livestock production, recording only closing livestock holdings, with no data on production of milk, beef, wool or lamb (as these commodities are collected by the ABS from processors rather than farmers). Further, BLADE financial data differs from ABARES survey data in reflecting financial outcomes for businesses which own farms (but which may also hold other business interests). In future, some of these gaps in the FLAD / BLADE data could potentially be addressed by integrating them with ABARES farm survey data and related models. However, attempts to integrate these datasets have been constrained by data privacy issues to date.

4.2 Models

4.2.1 Crop production

There exists scope to refine and extend the crop production modelling work started in this project. Refinements in methodology along with improvements in data (including more detailed weather data, and additional years of farm data as available) could see the performance of these models improve significantly over time.

These crop production models could have a range of applications. In particular, small area statistics generated in this study (for WA wheat) could be extended to cover all major crops across all of Australia. Further, research is required to test whether the approaches developed in this project, if applied more broadly, would satisfy all ABS confidentiality requirements.

In addition to these small area results, the model could also be applied to generate regular forecasts or 'nowcasts' (estimates for the current year in advance of survey collection lags) along with climate (seasonally) adjusted estimates.

4.2.2 Irrigation production and water use

The irrigation farm modelling undertaken in this project is limited in scope, and there exits potential to expand the coverage significantly to include a wider range of crops, regions and variables (including irrigated crop area). In future, farm level models of irrigated agriculture developed using the FLAD could be linked to ABARES economic model of the Southern Murray-Darling Basin Water market (Gupta et al. 2018). A linked water market and farm scale production model could have a range of applications in the analysis of water policy issues.

4.2.3 Farm financial outcomes

Given the time and resources available this project has only scratched surface in terms of analysis of BLADE financial data. Future research could for example extend the developed farm production models to include simulation of financial outcomes.

While this project has demonstrated correlation between BLADE financial data and FLAD data, establishing reliable farm–level relationships between the two will require more detailed analysis. In particular, there will be a need to account for sources of noise such as the effects of non-farming business activity, which can create a disconnect between total business financial outcomes (as recorded in BLADE) and farm production data (as recorded in FLAD).

These issues could be partially addressed through more detailed time series analysis (e.g., exploiting the panel data structure via fixed effects models) and/ or through the inclusion of external data sources. One possibility is the linking of ABARES *farmpredict* model (based on ABARES survey data) to the FLAD / BLADE data and models. This could for example enhance the capacity of the FLAD / BLADE datasets to analyse livestock farms.

4.3 Program evaluations and other research

The FLAD/ BLADE datasets could have a range of research applications. In particular, the ability to track individual farm outcomes consistently over time could support detailed program evaluation studies.

There exist a large number of state and commonwealth programs which target farm businesses including welfare programs (such as the Farm Household Allowance), drought programs (such

as the Future Drought Fund), water policy programs (such as those related to the Murray-Darling Basin Plan) and various environmental, land management and biosecurity programs. In future administrative data (detailing the interventions applied to individual farms) could be integrated to the FLAD / BLADE in order to estimate the causal 'treatment' effects of these programs on farms.

Similarly, the datasets could also be applied to examine the effects of specific farm management practices either by integrating external data or making use of ABS land management data.

4.4 Insurance and other commercial applications

The FLAD / BLADE database could have a range of practical applications within the rural finance sector, particularly in supporting drought insurance markets. The potential for these applications will depend greatly on data access / confidentiality limitations.

The simplest option would be the generation of small area statistics (such as those presented for WA wheat in this project) which could be made publically available. Consistent small area crop statistics could for example help support simple forms of crop insurance such as 'yield-area' insurance, which (while common in the U.S.) are currently not-feasible in Australia given the lack of consistent small region data.

One step beyond this would be the application of statistical models to develop farm specific indexes, which could be used for index-based insurance products (as demonstrated in case study 4). In practice, this would require these models to be extracted from the ABS datalab environment. With this approach, insurers need to only collect a small amount of information from each policy holder (such as their location and land area). This information can then be input into the models to generate indexes (of crop yield or farm revenue) which could form the basis of insurance contracts.

A more ambitious approach would be to establish a version of the FLAD / BLADE database with identifiers (as opposed to the de-identified version used in this project), which could be queried to obtain detailed historical data for any farm in Australia. This might for example involve individual businesses requesting access to their personal data, which they could then share with financial providers (insurers, lenders etc.). Such an approach may not be feasible under current data sharing arrangements, but it could be of significant value to the industry, particularly in reducing the administrative costs faced by insurers (and in turn farmers) in collecting required historical data.

References

Goesch, T. Legg, P. and Donoghoe, M. (2020) Murray-Darling Basin water markets: trends and drivers 2002-03 to 2018-19 https://www.agriculture.gov.au/abares/research-topics/water/murray-darling-basin-trends-and-drivers

Gupta, M., Hughes, N., Wakerman-Powell, K, 2018 A model of water trade and irrigation activity in the southern Murray-Darling Basin, https://www.agriculture.gov.au/abares/research-topics/water/water-trade-irrigation-model-southern-mdb

Hatt, M, Heyhoe, E, & Whittle, L 2012, '"Options for insuring Australian agriculture"', *Australian Bureau of Agricultural and Resource Economics and Sciences*,, <https://www.agriculture.gov.au/sites/default/files/sitecollectiondocuments/ag-food/drought/ec/nrac/work-prog/abares-report/abares-report-insurance-options.pdf>.

Hertzler, G 2005, 'Prospects for insuring against drought in Australia', in, *From disaster response to risk management*, Springer, pp.127–138.

Hochman, Z, Gobbett, DL, & Horan, H 2017, 'Climate trends account for stalled wheat yields in Australia since 1990', Global Change Biology, vol. 23, no. 5, pp. 2071–2081.

Hughes, N, Lawson, K, & Valle, H 2017, *Farm performance and climate: climate adjusted productivity on broadacre cropping farms*, Australian Bureau of Agricultural and Resource Economics and Sciences.

Hughes, N. 2018 'Better data would help crack the drought insurance problem' , https://theconversation.com/better-data-would-help-crack-the-drought-insurance-problem-106154

Hughes, N., Burns, K., Soh, W., Lawson, K. 2020, *Measuring drought risk: The exposure and sensitivity of Australian farms to drought*, ABARES research report

Hughes, N, Soh, W., Boult, C., Lawson, K., Donoghoe, M., Valle, H. and Chancellor, W. 2019, 'farmpredict: A micro-simulation model of Australian farms', *ABARES working paper*.

Appendix A: Approximate farm register

As part of this project an approximate farm register was developed by combing the FLAD and the BLADE. The register attempts to fill a number of gaps in the FLAD/ BLADE data, including:

- BLADE data is (currently) only available from 2001-02 to 2016-17 (2000-01 and 2017-18 are not available).
- BLADE to FLAD linkage was only undertaken from 2005-06 (the year that the ABS agricultural census/surveys adopted the Australian Business Register - ABR).
- FLAD coverage varies from around 90% in ABS census years down to around 20% in survey years.

To fill these gaps a number of assumptions are applied:

- The BLADE sample is limited to units which have been linked to the FLAD at least once between (2005-06 and 2016-17).
- Only active BLADE units are included within each year (*X_AL_ST=1* and *BIRTH_DATE* < current year, and BAS *TURNOVER* data is not null).
- BLADE data for 2000-01 and 2017-18 is taken from 2001-02 and 2016-17 respectively.
- Where a linked FLAD unit is not available (i.e., it was not sampled by the ABS in that year) the nearest observation of that unit is taken as a replacement (e.g., missing FLAD units in 2006-07 could be replaced with data from the census year 2005-06).
- Farm units in FLAD that were never linked to a BLADE unit are added to the register.
- The register only includes records where either:
    - FLAD data is available for the current year or can be replaced with data from 1 or 2 years pre/post the current year
    - or a FLAD replacement is available 3 or more years post(pre) the current year but that unit is also observed in the FLAD pre(post) the current year.
- FLAD units with a *Z_usesti* code of 5 or greater (indicating 'dead' or imputed units) are excluded.

Table 11 provides a comparison of the FLAD/BLADE farm register with public ABS population estimates. Post 2005-06 (where the ABS and BLADE both use the ABR) the FLAD/BLADE register approximates the population relatively closely. Comparing the number of businesses is difficult given changes in ABS scope (ABR adoption in 2005-06 led to a large increase in units and the lower EVAO threshold in 2015-16 to a large decrease). These issues aside, the FLAD/BLADE register appears to contain around 5-10% fewer units than estimated in public figures. For many key variables (such as beef cattle numbers and wheat area) the FLAD/BLADE register tends to underestimate public estimates by a similar amount.

From 2007-08 the register achieves a total land area similar to public estimates. The register tends to over estimate the population land area around 2005-06, likely due to the switch to the ABR. The register also underestimates land area in 2000-01 and 2001-02 due to missing BLADE data and lack of BLADE/ FLAD linkage pre 2005-06.

Table 11: Comparison of FLAD/BLADE farm register with ABS population estimates

	ABS public statistics		FLAD / BLADE Approximate farm register	
Year	Agricultural establishments (no.)	Total land area ('000 ha)	Farm/FLAD units (no.)	Total land area ('000 ha)
2000-01	140,516	455,714	114,923	404,263
2001-02	135,377	447,008	115,235	393,964
2002-03	132,983	440,162	115,351	395,694
2003-04	130,526	440,188	140,762	454,499
2004-05	129,934	445,149	146,317	479,235
2005-06	154,472	434,925	148,820	485,638
2006-07	150,403	425,449	144,507	481,314
2007-08	140,704	417,288	134,185	409,171
2008-09	134,996	409,029	122,054	388,424
2009-10	134,184	398,580	123,788	407,725
2010-11	135,447	409,673	125,077	418,111
2011-12	135,692	405,474	121,929	394,970
2012-13	128,917	396,615	117,520	392,025
2013-14	128,489	406,269	84,715	383,371
2014-15	123,091	384,558	81,936	386,620
2015-16	85,681	371,078	81,432	413,475
2016-17	88,073	393,797	76,532	386,642
2017-18	85,483	378,082	72,317	366,740

Appendix B: Statistical models

Crop farm model

Similar to *farmpredict* (Hughes et al. 2019) the crop production model is a statistical model linking multiple target (dependent variables) with multiple explanatory (feature) variables. Target variables include crop areas planted and yield, feature variables include climate conditions, prices, farm fixed inputs and other controls. These regression models are estimated from historical data using non-parametric regression methods (involving the *xgboost* method, see Hughes et al. 2019).

Formally, the crop production model takes the form

$$D_{jit}\ ,\dot{A}_{jit}\ ,\dot{Q}_{jit} = \boldsymbol{F}(\boldsymbol{K}_{it}\ \boldsymbol{C}_{it}\ \boldsymbol{Z}_{it}\ \boldsymbol{P}_{it})$$

$$\textit{targets} \qquad\qquad \textit{features}$$

Where:

A_{jit} Area of crop *j* for farm *i* in year *t*

D_{jit} Crop classification, = 1 if $A_{jit} > 0$

Q_{jit} Quantity of crop *j* produced

$\dot{Q}_{jit}$ Crop yield for crop *j*, $= \frac{Q_{jit}}{A_{jit}}$

$\dot{A}_{jit}$ Proportion of cropping land planted to crop *j*, $= \frac{Q_{jit}}{Z_{it}^{LC}}$

Z_{it}^{LC} Farm land area available for cropping on farm *i*

$\boldsymbol{Z}_{it}$ Vector of control variables

$\boldsymbol{C}_{it}$ Vector of climate variables

$\boldsymbol{P}_{it}$ Vector of commodity prices

$\boldsymbol{K}_{it}$ Vector of fixed inputs (including livestock no., tree numbers, land area)

Eight crop types *j* are included in the model *Wheat, Barley, Oats, Sorghum, Triticale, Maize* and *Canola* and *Other cereals*.

Predictions of $D_{jit}\ \dot{A}_{jit}\ \dot{Q}_{jit}$ generated from the above statistical model can then be used to simulate farm crop production and value as follows

$\hat{A}_{jit} = Z_{it}^{LC} D_{jit} \hat{\dot{A}}_{jit}$

$\hat{Q}_{jit} = \hat{A}_{jit} \hat{\dot{Q}}_{jit}$

$\hat{V}_{jit} = \hat{Q}_{jit} \,.\, P_{jit}$

$$\hat{V}_{it} = \sum_{j} \hat{V}_{jit}$$

Where:

P_{jit} Price of crop *j* produced

V_{jit} Value crop *j* produced, = $P_{jit}. Q_{jit}$

Here the value of production for each farm unit V_{it} includes the value (price times quantity) of the seven modelled crops, as well as the value of any other agricultural commodities produced on the farm (although production of all other commodities is assumed exogenous / as observed in the data).

Note while the model estimation is limited to farms actually observed in FLAD, the results $\hat{V}_{it}, \hat{Q}_{jit}, \hat{A}_{jit}$ can be generated for the full farm population (or at least for all farms in the FLAD/BLADE based register).

Table 12: Crop farm model regression results

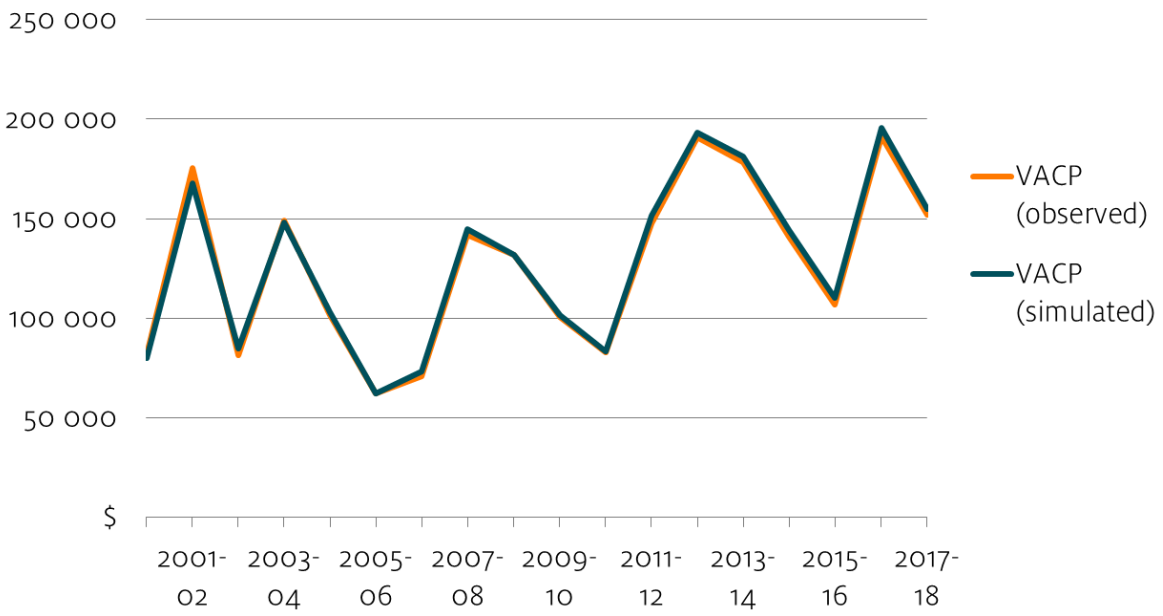
	N	AUC		R-squared		RMAE	
Target		OLS	XGB	OLS	XGB	OLS	XGB
A_barley_dot	110,690			0.10	0.41	0.37	0.31
A_canola_dot	55,112			0.20	0.45	0.35	0.31
A_maize_dot	4,633			0.13	0.43	0.36	0.30
A_oats_dot	61,702			0.22	0.58	0.33	0.25
A_sorghum_dot	17,725			0.13	0.37	0.39	0.34
A_triticale_dot	15,084			0.22	0.51	0.31	0.27
A_wheat_dot	151,302			0.15	0.40	0.33	0.27
D_barley	268,275	0.81	0.91				
D_canola	268,273	0.82	0.93				
D_sorghum	268,273	0.96	0.98				
D_wheat	268,275	0.89	0.95				
Q_barley_dot	111,810			0.20	0.39	0.38	0.31
Q_canola_dot	55,280			0.23	0.41	0.35	0.29
Q_maize_dot	4,492			0.34	0.46	0.40	0.36
Q_oats_dot	61,976			0.20	0.33	0.48	0.41
Q_other_cereals_dot	7,346			0.28	0.43	0.79	0.63
Q_sorghum_dot	17,783			0.19	0.34	0.38	0.34
Q_triticale_dot	15,281			0.27	0.39	0.38	0.33
Q_wheat_dot	156,887			0.27	0.46	0.33	0.26

RMAE – Relative Mean Absolute Error, AUC – Area under the curve

Table 13: Crop farm model validation results (R^2)

Variable	Farm	Region	National
A_barley	0.64	0.97	0.99
A_canola	0.62	0.96	1.00
A_maize	0.58	0.85	0.98
A_oats	0.52	0.92	0.96
A_sorghum	0.55	0.96	0.97
A_triticale	0.63	0.93	1.00
A_wheat	0.83	0.99	0.99
Q_barley	0.60	0.96	0.99
Q_canola	0.61	0.96	1.00
Q_maize	0.56	0.84	0.96
Q_oats	0.46	0.90	0.97
Q_sorghum	0.57	0.96	0.99
Q_triticale	0.53	0.87	0.99
Q_wheat	0.75	0.98	0.99
V_barley	0.60	0.96	0.99
V_canola	0.61	0.96	1.00
V_maize	0.58	0.84	0.97
V_oats	0.44	0.89	0.98
V_sorghum	0.59	0.96	0.99
V_triticale	0.52	0.85	0.99
V_wheat	0.75	0.98	0.99
V_total_endog	0.80	0.98	0.99

Figure 22: Average annual value of farm crop production (*V_total_endog*) actual vs predicted



Irrigation farm model

The irrigation model predicts the water use and yield (production) for key irrigated crops.

$$\dot{Q}_{jit}\, \dot{W}_{kit} = \boldsymbol{F}(\boldsymbol{K}_{it}\, \boldsymbol{C}_{it}\, \boldsymbol{Z}_{it}\, \boldsymbol{P}_{it})$$

Where:

W_{kit}	Water use for irrigation activity k
$\dot{W}_{kit}$	Water application rate for irrigation activity $k = \frac{W_{kit}}{A^{irrig}{}_{kit}}$
$\dot{Q}_{jit}$	Yield for commodity j
Q_{kit}	Production for commodity j

The model currently predicts water application rates for four irrigation activities *Fruit and nuts*, *Grapes*, *Cotton*, *Rice* and production (yield) for a selection of key irrigated commodities (*Cotton*, *Rice*, *Grapes*, *Almonds* and *Oranges*). This model currently takes crop areas planted as exogenous. Crop yields are defined as production relative to area planted (except for almonds and oranges) which are defined relative to the number of trees. Similar to the crop production model the estimation uses a non-parametric (*xgboost*) algorithm adapted from ABARES *farmpredict* model.

This model is limited to irrigation farms within the Murray-Darling Basin. Annual average prices for water allocations in MDB catchment areas are included as an additional feature variable. The irrigation farm model could be extended to cover a much broader range of crops. Further it could be extended to make land areas for non-perennial crops (such as rice and cotton) endogenous (as in the above crop production model).

Table 14: Irrigation farm model regression results

	Method: xgbboost		Method: ols	
Target	**R^2**	**RMAE**	**R^2**	**RMAE**
Q_almonds_dot	0.39	0.38	-0.03	0.43
Q_cotton_dot	0.13	0.16	-0.04	0.18
Q_grapes_dot	0.50	0.29	0.01	0.37
Q_oranges_dot	0.24	0.29	0.05	0.34
Q_rice_dot	0.31	0.11	0.16	0.12
W_cotton_dot	0.44	0.21	0.35	0.24
W_fruit_and_nuts_dot	0.55	0.33	0.35	0.44
W_grapes_dot	0.71	0.25	0.58	0.34
W_rice_dot	0.25	0.13	0.14	0.15

Table 15: Irrigation farm model validation results

	R-squared	
Target	**farm-level**	**across all farms**
Q_almonds	0.96	0.99
Q_cotton	0.88	0.99
Q_grapes	0.85	0.99
Q_oranges	0.93	0.98
Q_rice	0.96	0.99
W_cotton	0.94	0.99
W_fruit_and_nuts	0.59	0.96
W_grapes	0.76	0.99
W_rice	0.95	0.99

Figure 23: Average annual water application rate for rice actual vs predicted

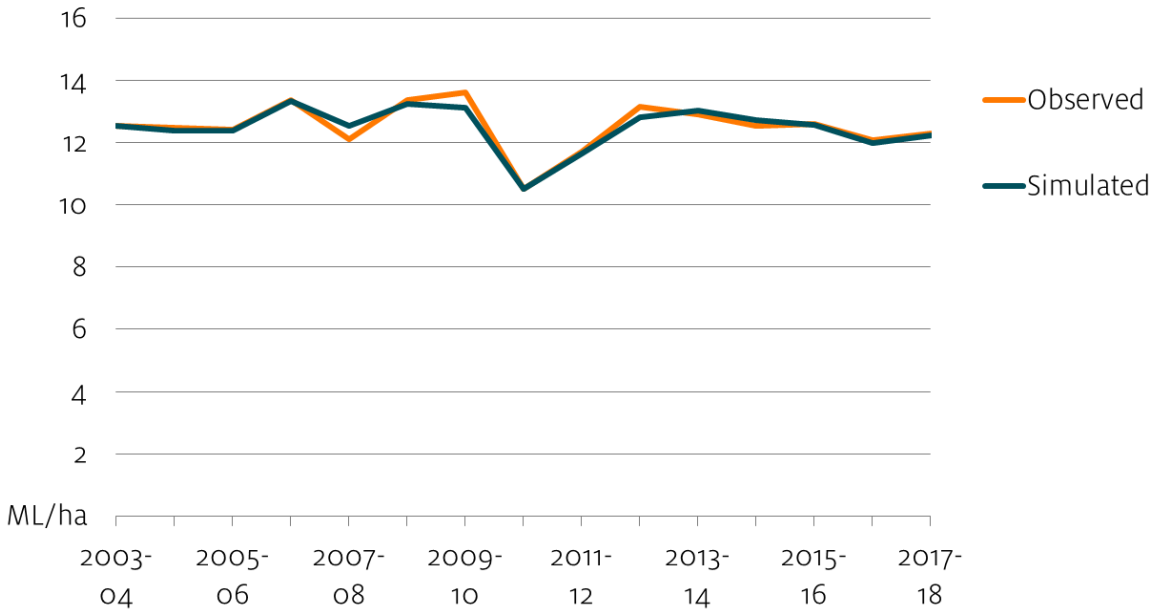
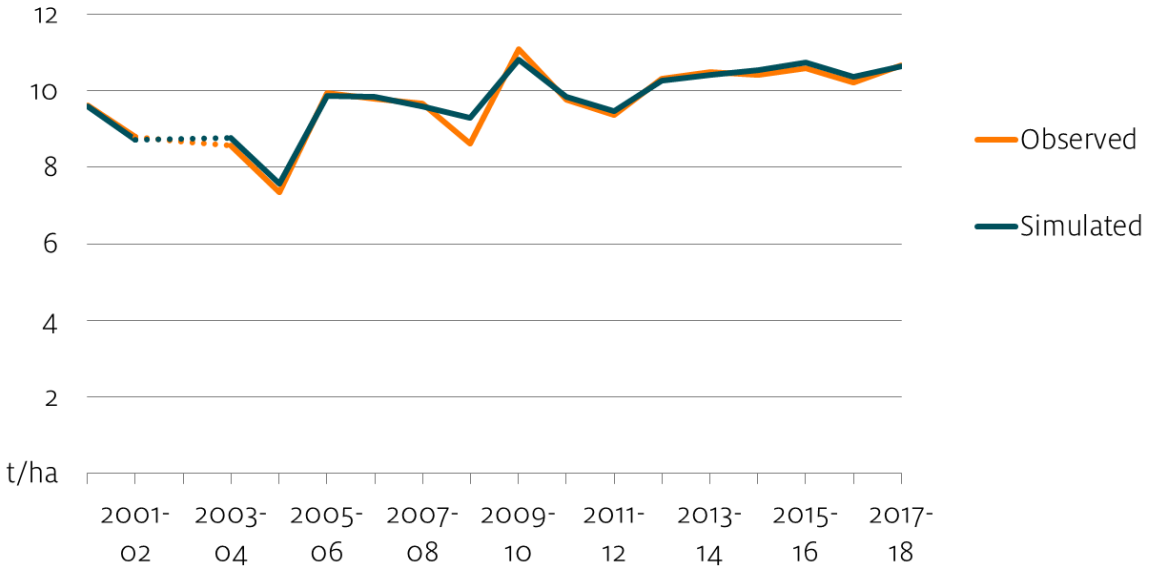


Figure 24: Average annual yield for rice actual vs predicted



Note: Data for rice area irrigated was not collected in 2002-03.

Appendix C: Case study assumptions

Scenario assumptions

Crop farm model

Results for case studies 1, 2, 3 and 4 are based on scenario results from the crop farm model. Three model scenarios were simulated:

- *Baseline (sample data):* this scenario takes actual historical climate data, prices and farm characteristics as defined in the model training data. The scenario provides the model's best (cross-validated / out-of-sample) predictions of historical farm outcomes.
- *Baseline (population data)*: the baseline scenario is also applied to the larger population data set (approximate farm register). Here the model provides predictions of historical farm outcomes both for sampled and non-sampled farms ('copy' farms).
- *Climate scenario (population data):* in this scenario commodity prices and farm characteristics are held constant (at historical values), and climate conditions are simulated using the historical climate sequence (2000-01 to 2017-18). This climate simulation is performed for each observation in the population data (18 farm/ price years by 18 climate years).

Irrigation farm model

Model results for case study 5 are based on a climate scenario generated from the irrigation farm model:

- *Climate scenario (population data):* in this scenario output prices and farm characteristics are held constant (at historical values), and climate conditions and water market prices are simulated using the historical climate sequence (2000-01 to 2017-18). This climate simulation is performed for each observation in the population data (18 farm/ price years by 18 climate years). An assumption of 'full maturity' is applied to all tree crops (all non-bearing trees are assumed to reach bearing age).

Case study 1: Trends in Australian crop production

Results for this case study are based on the crop farm model *climate scenario*. Here we define $\tilde{Q}_{jitc}$, $\tilde{A}_{jitc}$, $\tilde{V}_{jitc}$ as the simulated results for crop production, area and value for crop *j* farm *i* in year *t* under year *c* climate conditions.

Climate adjusted values are then defined as:

$$\bar{Q}_{jit} = \frac{1}{18}\sum_{c} \tilde{Q}_{jitc}$$

While non-climate adjusted values are defined simply as $\tilde{Q}_{it} = \tilde{Q}_{itt}$. Regional and national crop yields are defined as:

$$\bar{Q}_{jt} = \frac{\sum_i \bar{Q}_{jit}}{\sum_i \bar{A}_{jit}}$$

To generate aggregate (population) results, state level scaling factors are applied for each crop. These scaling factors account for underestimation of the population in the approximate farm register (see Appendix A). The scaling factors ensure that under the *baseline (population)* scenario, results for crop area and production match published ABS state totals. Further detail on the scaling factors used for WA wheat is provided below.

Case study 2: Small area statistics for WA wheat

Results from the crop production model *baseline (population) scenario* are used to generate estimates of wheat area and production by *SA2Ag* region in Western Australia. State level results for area planted and yield are presented in Figure 25 and Figure 26 below. These charts compare the modelled values against published figures (pre-scaling). These values are then scaled to match the published WA totals.

As would be expected the model predicted areas underestimate the published totals due to the approximate farm register underestimating the farm population (see Appendix A). For reasons discussed in Appendix A this underestimation is stronger in the earliest and latest years. Model yield estimates are however extremely close to published estimates at the state level.

Figure 25: WA total wheat area model predicted (pre-scaling) vs ABS published estimate

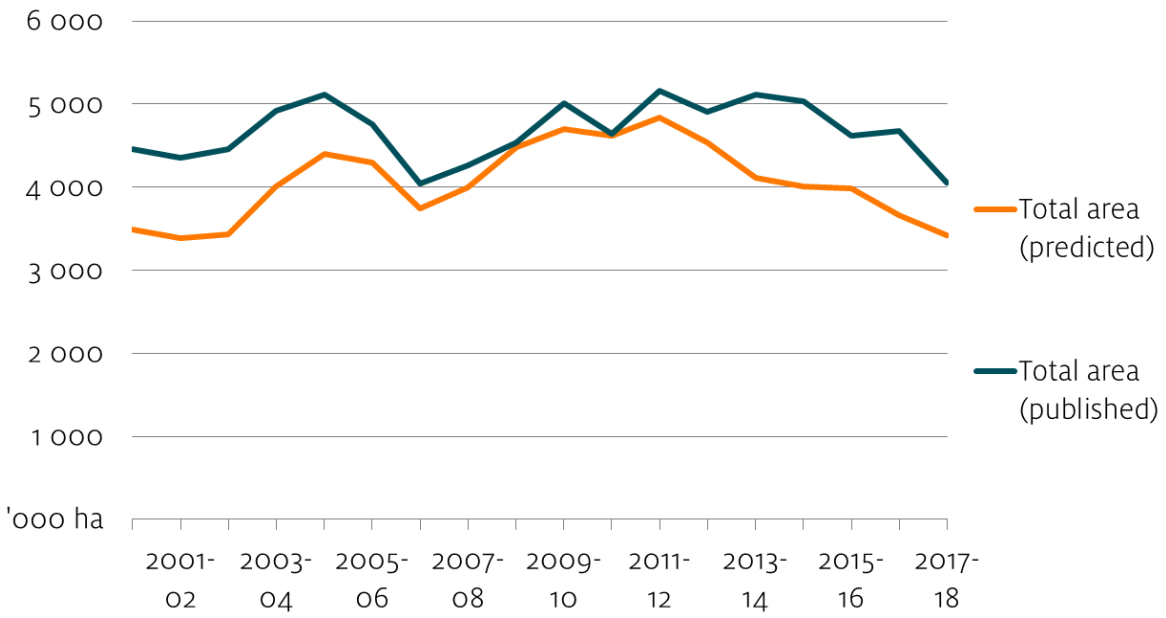


Figure 26: WA wheat yield model predicted (pre-scaling) vs ABS published estimate

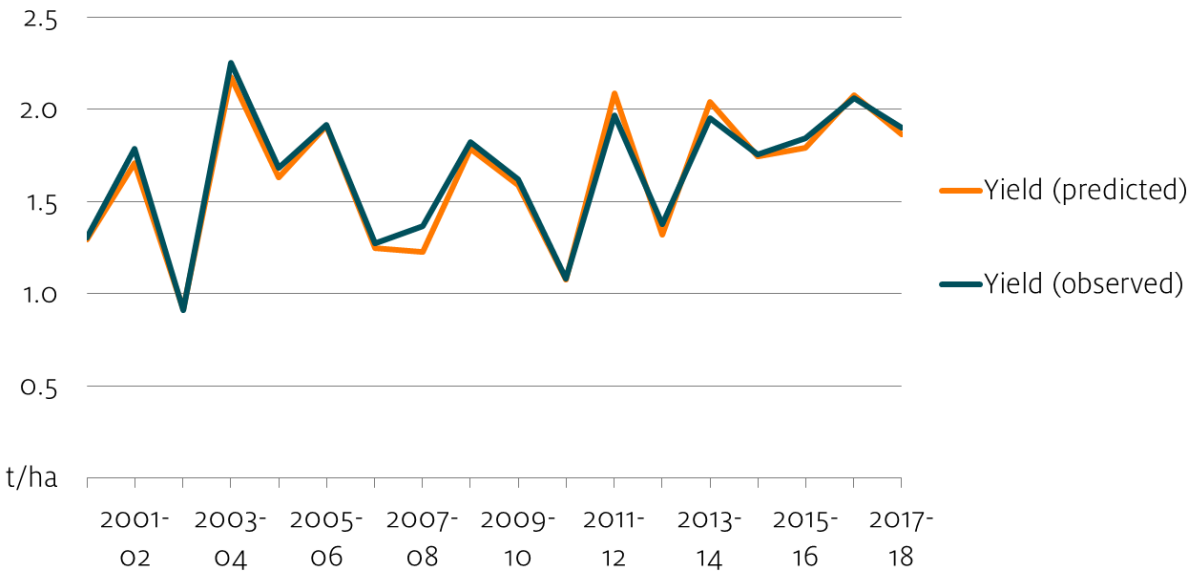


Table 16: WA Wheat area and production model predicted (pre-scaling) vs ABS published estimate

	Predicted		Observed	
Year	Total area	Total production	Total area	Total production
2000-01	3,489,905	4,503,421	4,459,525	5,812,908
2001-02	3,383,068	5,782,498	4,349,860	7,759,869
2002-03	3,427,251	3,132,086	4,457,670	4,046,941
2003-04	4,008,765	8,703,741	4,916,999	11,070,004
2004-05	4,403,850	7,177,594	5,118,332	8,619,022
2005-06	4,295,312	8,200,411	4,752,742	9,088,093
2006-07	3,741,795	4,662,643	4,037,048	5,134,318
2007-08	3,998,579	4,908,437	4,258,162	5,820,230
2008-09	4,478,125	7,986,956	4,541,947	8,273,975
2009-10	4,701,028	7,468,221	5,005,949	8,114,121
2010-11	4,612,686	4,963,856	4,639,518	5,004,615
2011-12	4,841,765	10,106,921	5,155,761	10,145,107
2012-13	4,535,520	5,980,988	4,909,209	6,744,055
2013-14	4,111,816	8,377,517	5,114,891	9,976,941
2014-15	4,008,994	7,000,155	5,038,134	8,824,410
2015-16	3,984,207	7,133,893	4,615,761	8,510,577
2016-17	3,661,949	7,602,153	4,677,774	9,644,881
2017-18	3,417,682	6,366,477	4,056,574	7,698,552

Case study 3: Effects of drought on cropping farms

Results in this case study are derived firstly from the crop production model *climate scenario.* In this case study the focus is on total crop production value (across all broadacre crops): $\tilde{V}_{jit} = \sum_j \tilde{V}_{jit}$. The analysis is limited to cropping specialist farms (ANZSIC code 149).

For each farm observation (farm *i* in year *t*) the climate years *c* are ranked from best to worst to define the range of percentile scores. Here we define $\tilde{V}_{it}^{p}$ as the *p*th climate percentile value of farm production (for farm *i* in year *t*). These farm level percentile values can be aggregated to regional or national level: $\tilde{V}_{t}^{p} = \sum_i \tilde{V}_{it}^{p}$.

Figure 12 then shows average farm value of production for each percentile *p* (for farm year *t* = 2018): $\dot{V}_t^p = \tilde{V}_t^p / \sum_i Z_{it}^{LC}$. Figure 13 shows the percentage difference between the $\dot{V}_t^{50}$ and $\dot{V}_t^0$ for each region.

This case study also compares these model results with observed BLADE data. When merging BLADE data any 'one-to-many' and 'many-to-one' FLAD/BLADE links are addressed through aggregation. For example, if a BLADE unit has multiple linked farm units *V* estimates for that BLADE unit are totalled across all linked FLAD units.

Three BLADE variables are considered: business turnover (from the BAS data), business income (revenue) (from the BIT data.) and business profit (also from BIT data). The BIT variables combine data for different business types (company, trust, partnership, individual) into a single variable. Non-primary production income and profit is excluded for those business types where this data is available (all except companies).

Case study 4: Index-based cropping farm drought insurance

Case study 4 considers a hypothetical insurance product, where payouts are tied to an index of crop production (the simulated value of crop production, $\tilde{V}_{itt}$). Under this insurance product farmers receive payouts when simulated production is below the 20th percentile threshold $\tilde{V}_{it}^{20}$ (note that this threshold is specific to the farm / price year *t*). Insurance pay-outs for farm *i* in period *t* are defined as:

$$B_{it} = max\left(\tilde{V}_{it}^{20} - \tilde{V}_{itt}, 0\right)$$

Expected payouts for farm *i* in period *t* $E[B_{it}]$ are defined as:

$$E[B_{it}] = \frac{1}{18}\sum_c max\left(\tilde{V}_{it}^{20} - \tilde{V}_{itc}, 0\right)$$

In practice, insurance premiums would involve some margin above $E[B_{it}]$, for simplicity here we assume insurance premiums are just set to $E[B_{it}]$ for each farm.

This insurance product is assumed to be held by cropping specialist and mixed-cropping livestock farms (ANZSIC 149) in the approximate farm register (population data) with at least 5 per cent of total land set-up for cropping. Total pay-outs by year (nationally and by state) are then defined as the sum of farm level payouts, multiplied by scaling factors (to account for underestimation of the farm population as described previously).

Note that these simulations make use of a limited climate sequence (2000-01 to 2017-18). In the presence of climate change (particularly where there are trends towards more frequent / severe droughts) this approach will under-estimate insurance payouts relative to a more realistic scenario where pay-out thresholds are defined using a longer historical climate sequence.

Case study 5: Water productivity in the Murray-Darling Basin

Case study 5 results are based on the irrigation farm model *climate scenario.* Here, $\tilde{Q}_{jitc}$, $\tilde{W}_{jitc}$ are defined as the simulated results for production and water use for irrigated crop *j* farm *i* in year *t* under year *c* climate conditions (and water prices). Climate adjusted values and regional yields are defined as in case study 1, while climate percentiles are defined in case study 3.

Water productivity is then defined as:

$$\bar{K}_{jt} = \frac{\sum_i \bar{Q}_{jit}}{\sum_i \bar{W}_{jit}}$$

Results for each irrigated commodity (rice, cotton, grapes, oranges, almonds) are generated only for farms with that commodity (rice and grapes area greater than 0; and orange or almonds and total irrigated horticultural area greater than 0).

Water use for specific horticultural activities (oranges, almonds) is not collected by the ABS. In this case study, water use for these activities is approximated by assuming a farm's total water use for fruits and nuts, and limiting the sample to only those farms where at least half of the farm's total trees are dedicated to that specific horticultural activity (oranges, almonds).