Automatic transcription of historical weather observations

A report to the UK Met Office CCSP-China project
31 March 2017

Principal Investigator: Gilbert P. Compo[1,2]
Co-Principal Investigator: Prashant. D. Sardeshmukh[1,2]
Contributors: Chesley McColl[1,2], Lawrence Spencer[1,2], Raj Rajagopal[3], William F. West[3], Philip Brohan[4], Kevin Wood[5,6]

[1] University of Colorado, Cooperative Institute for Research in Environmental Sciences, Boulder, Colorado USA

[2] National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Physical Sciences Division, Boulder, Colorado USA

[3] WesTech eSolutions Inc., Longmont, Colorado USA

[4] UK Met Office, Hadley Center, Exeter UK

[5] University of Washington, Joint Institute for the Study of the Atmosphere and Ocean, Seattle, Washington USA

[6] National Oceanic and Atmospheric Administration, Pacific Marine Environmental Laboratory, Seattle, Washington USA

Global reconstructions of the Earth's weather and climate extending into the past for as long as possible are a key to understanding and predicting the varying risks of extreme and high impact weather. When such retrospective datasets are made using the modern scientific technique of data assimilation, combining numerical weather prediction model forecasts and observations, we obtain a long record of the state of the weather, the retrospective "analysis" or "reanalysis". The 20th Century Reanalysis (20CR, go.usa.gov/XTd) has generated the first such reanalysis dataset back to 1851 and is investigating going back 200 years. By using ensemble data assimilation, 20CR provides both an estimate of the state of weather and the uncertainty in that state. The standard deviation of the ensemble and its 56 individual members of weather estimates are important components of the dataset, allowing observational uncertainty information to be included in calculations of extreme event risk and high impact weather variability. To improve such risk estimation, reducing the main source of uncertainty: sparse historical observations, requires additional observations. One valuable source of untapped observations are those recorded in ship log books currently residing in paper form in national archives. Ocean observations of barometric pressure from these log books, as well as land observations from station records, provide information that the reanalysis system can use to represent the full three dimensional structure and time evolution of weather patterns, high impact storms, and extreme events, such as storms, floods, droughts, heat waves, cold spells, and hurricanes.

To improve the quantity and quality of weather observations, the Atmospheric Circulation Reconstructions over the Earth (www.met-acre.org) initiative, a partner of 20CR, in conjunction with Zooniverse.org and the UK Met Office, has undertaken the Oldweather.org citizen science project to recover and digitize millions of weather observations from ship log books. While successful, progress is slow. Crucial steps for using these observations in 20CR and other historical reanalysis datasets being imaged by ACRE partners are digitizing and quality controlling them and converting them to the standard format of ,e.g., International Marine Meteorological Archives so that these data can be ingested into international databases that form the input observations for 20CR such as the International Surface Pressure Databank (http://reanalyses.org/observations/international-surface-pressure-databank) and International Comprehensive Ocean Atmosphere Data Set (http://icoads.noaa.gov).

The importance of these observations is illustrated in **Figure 1**. The track of the *USS Jeannette* during an ill-fated Arctic research is shown in the left hand panels. In the right hand panels, the observations of barometric pressure taken aboard the Jeannette and transcribed by Oldwether.org volunteers are compared to estimates from the independent 20CR version 2c, which did not assimilate or include these observations. Where 20CRv2c has sufficient observations, such in the latitudes of 30N to 50N, there is good agreement. Where 20CRv2c, is observation poor, such in the Arctic, the agreement is poor.

The comparison highlights that to increase the skill of our global reanalyses, more observations are needed. These observations exist over the oceans and over land, but converting them from their current paper records to the scientific formats needed for study and predictive understanding remains a challenge. Not least among these challenges is the difficulty in getting the observations transcribed after the paper records are imaged. The current system of manual keying, whether by volunteers in Oldweather, or by paid professionals, is slow.

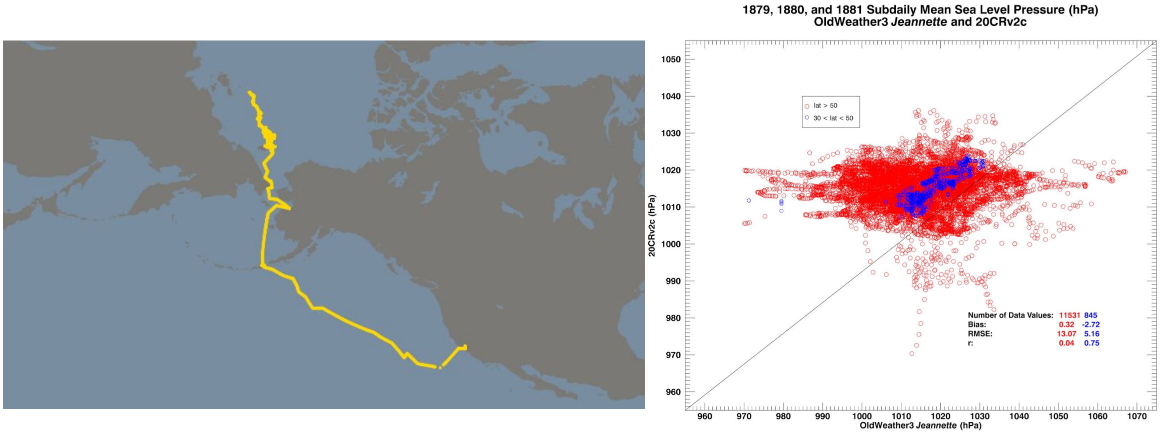


Figure 1. [left] Map of the track of the USS Jeannette from 1879 to 1881 (yellow dots), where it sank in the Arctic. [right] Scatterplot of barometric pressure observations from the USS Jeannette compared to the 20th Century Reanalysis version 2c, which is independent of the observations.

In this project with UK Met Office and University of Washington, the main effort has consisted of the University of Colorado Cooperative Institute for Research in Environmental Sciences with partners at NOAA Earth System Research Laboratory, the UK Met Office Hadley Centre, the University of Washington, and WesTech eSolutions investigating whether it is possible and feasible to accelerate the transcription and transition of ship log book and India daily weather observations to suitable formats for assimilation into 20CR and free inclusion in international databases.

The central question for the project was posed as "Suppose I have 1 million images (300dpi jpegs) containing tabular weather data in a mixture of formats, how do I get the data transcribed?"

1) Is there some turnkey software system that can do the job?
2) Can we build our own system from existing components?
3) Can we use software to do part of the job?
4) Do we still need to do everything manually?

A proof of concept software system has been developed by WesTech eSolutions to produce transcribed weather data from digital images of logbooks from US Government ships, and pages from the Indian Daily Weather Reports (IDWR). In the proof of concept, no manual intervention has been used to provide estimates of the current capabilities of machine-only autotranscription. From the effort, it is clear that the skill of the existing proprietary Parascript FormXtra recognition engine, with considerable effort in Form definition and thoughtful application of Templates, has demonstrated that not all tabular weather data needs to be recovered manually. But, fully-automated transcription, particularly of handwritten logbook observations, is not available, yet. The most printed document, the IDWR, had the most success in being transcribed with a fully-automated process.

Some images have been supplied from the Met Office and others have been taken from the public domain or under an appropriate open-access license in the United States.
The transcribed data have been provided to the Met Office for inclusion in future versions of ICOADS and similar climate datasets, as well as for ongoing study of the prospects of automated and combined automated/manual transcription.

Deliverables:

1. *Implemented transcription workflow*

Completed and described in the Appendix for 1) station records, 2) printed logbooks, and 3) handwritten logbooks.

2. *Automatic transcription of at least two types of documents*

Completed and exceeded with three types of documents tested: 1) printed station records from the IDWR, 2) Typed logbook pages, 3) handwritten logbook pages

3. *Data from at least 10 logbooks and one volume of IDWR converted into a mutually agreed digital text data format and available for research by 2017-03-31.*

Completed and exceeded. More than 500 pages of transcription results have been delivered to the UK Met Office via a Dropbox link.
https://www.dropbox.com/sh/6wb0t822slgi29g/AADr1fdyVkDBWffTtSgSWNcJa?dl=0

Results

The University of Colorado/CIRES subjectively analyzed the results of the autotranscription from 10 pages of each of the three document tables. The UK Met Office provided side-by-side comparisons of the transcription and the logbook images (see Figures 2 and 3). For CIRES, the primary focus was the success of transcribing the Barometric Pressure (or Sea Level Pressure), as this is the variable assimilated by 20CR. Tables are available detailing the results from the IDWR, the typed logbook of the Farragut,and the handwritten logbook of the USS Jeannette.

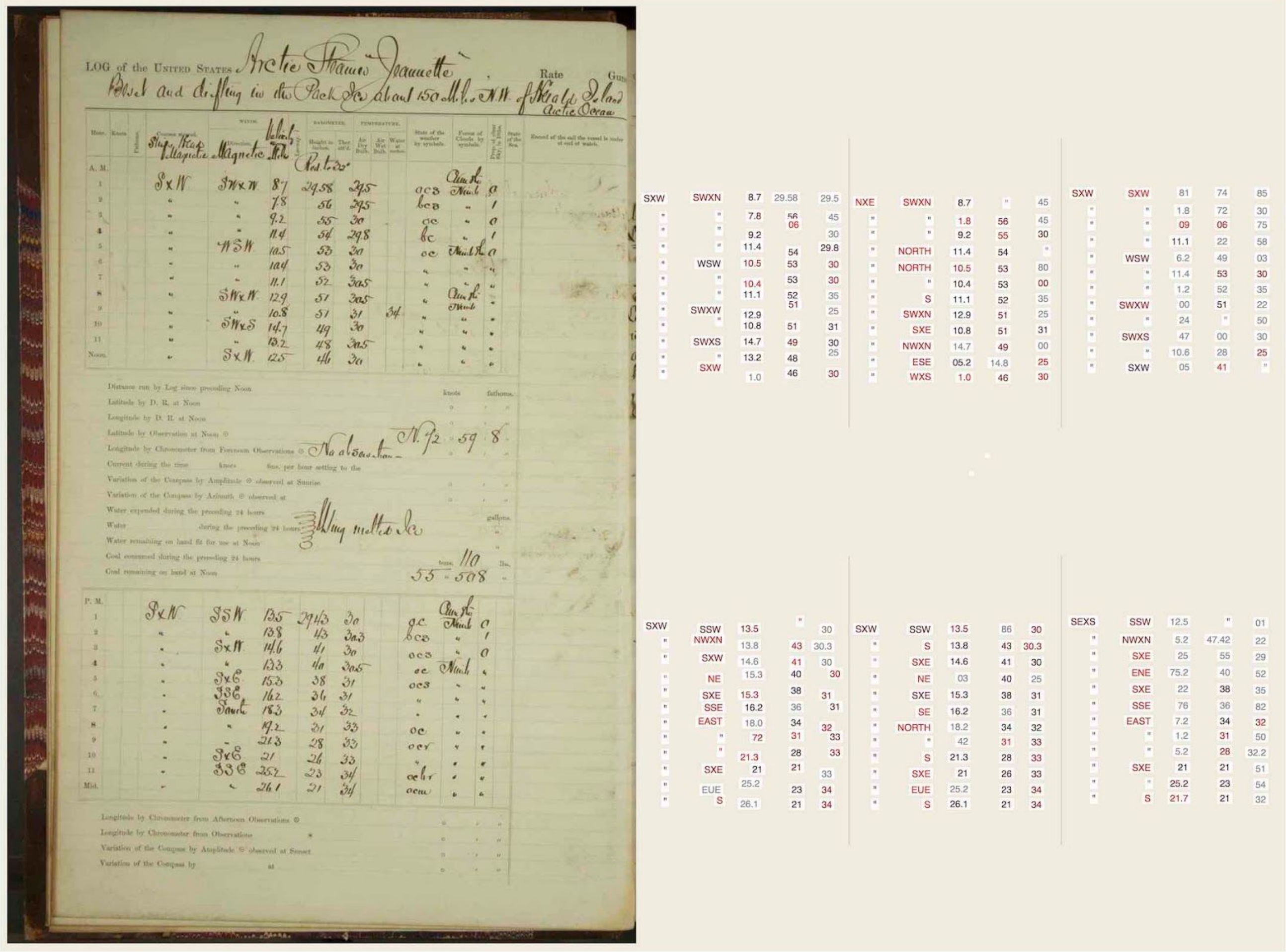
As an example of the results, for the recognition of Barometric pressure from the IDWR-1888, 62% of the values from the 10 random pages tests were perfect recognition. For the *Farragut*, 41% of its hourly values had perfect recognition.

For the *Jeannette*, as illustrated in **Figure 2**, for the early logbook pages of the voyage, the recognition of the first row of the data table is particularly important, as the first two digits of the pressure value are found there, while subsequent rows have only the two digits after the decimal. This was much more successful for the AM table of data, which appears on the top of the page, than the PM table, which appears at the bottom. No first row pressure of the PM table was successfully recognized, while the first row pressure of the AM table was perfectly recognized on 3 of the 5 pages examined in this configuration, and was plausible in the other 2. The subsequent hourly values with the last two digits were recognized perfectly in 91% of the 55 values from AM tables and 85% of the 55 values from the PM tables. In the later part of the voyage, the complete value was recorded 3 to 4 times per day. This was more of challenge, as illustrated in **Figure 3**, only 3 out of 32 values were recognized perfectly. Two of those are shown in the figure.

As readily apparent from Figures 2 and 3, and also representative of the success in the *Farragut* case, wind speed (force) was well-captured. For the *Farragut* test pages, 61% of the wind force values were recognized perfectly. For the *Jeannette*, 56% of the values on the CIRES test images were perfectly recognized.

Well over half of the values targeted in the specification document (see Appendix) were transcribed perfectly. While recognition of date and position for the *Jeannette* was a challenge, 70% of the Farragut dates and 60% of the Farragut latitudes were recognized perfectly in the CIRES study pages.

In conclusion, the purely automated transcription results in a confidence-assessed recognition of key meteorological variables. The initial results suggest better than 50% recognition of most values, even from the handwritten *Jeannette*, with some values, such as the final two digits of pressure readings at better than 80%.



LOG of the UNITED STATES Arctic Steamer Jeannette Rate Gun

Beset and drifting in the Pack Ice about 150 Miles N.W. of Herald Island Arctic Ocean

Hour	Courses steered Ship Head Magnetic	Wind Direction Magnetic	Velocity Miles	Barometer Height in inches (Red. to 32°)	Air Dry Bulb	Water at surface	State of the weather by symbols	Forms of Clouds by symbols	Prop. of clear sky
A. M.									
1	S×W	SW×W	8.7	29.58	29.5		ocs	Cum. Str. Nimb.	0
2	"	"	7.8	56	29.5		bcs	"	1
3	"	"	9.2	55	30		oc	"	0
4	"	"	11.4	54	29.8		bc	"	1
5	"	WSW	10.5	53	30		oc	Nimb. Str.	0
6	"	"	10.4	53	30		"	"	"
7	"	"	11.1	52	30.5		"	"	"
8	"	SW×W	12.9	51	30.5		"	Cum. Str. Nimb.	"
9	"	"	10.8	51	31	34	"	"	"
10	"	SW×S	14.7	49	30		"	"	"
11	"	"	13.2	48	30.5		"	"	"
Noon	"	S×W	12.5	46	30		"	"	"

Distance run by Log since preceding Noon knots fathoms

Latitude by D. R. at Noon

Longitude by D. R. at Noon

Latitude by Observation at Noon ☉ N. 72 59 8

Longitude by Chronometer from Forenoon Observations ☉ No observation

Current during the time knots fms. per hour setting to the

Variation of the Compass by Amplitude ☉ observed at Sunrise

Variation of the Compass by Azimuth ☉ observed at

Water expended during the preceding 24 hours gallons

Water during the preceding 24 hours Using melted Ice

Water remaining on hand fit for use at Noon

Coal consumed during the preceding 24 hours tons 110 lbs.

Coal remaining on hand at Noon 55 508

Hour	Courses steered	Wind Direction	Velocity	Barometer	Air Dry Bulb	Water at surface	State of the weather	Forms of Clouds	Prop. of clear sky
P. M.									
1	S×W	SSW	13.5	29.43	30		oc	Cum. Str. Nimb.	0
2	"	"	13.8	43	30.3		bcs	"	1
3	"	S×W	14.6	41	30		ocs	"	0
4	"	"	13.3	40	30.5		oc	Nimb.	"
5	"	S×E	15.3	38	31		ocs	"	"
6	"	SSE	16.2	36	31		"	"	"
7	"	South	18.3	34	32		"	"	"
8	"	"	19.2	31	33		oc	"	"
9	"	"	21.3	28	33		ocr	"	"
10	"	S×E	21	26	33		"	"	"
11	"	SSE	25.2	23	34		ocbr	"	"
Mid.	"	"	26.1	21	34		ocm	"	"

Longitude by Chronometer from Afternoon Observations ☉

Longitude by Chronometer from Observations

Variation of the Compass by Amplitude ☉ observed at Sunset

Variation of the Compass by at

SXW	SWXN	8.7	29.58	29.5
"	"	7.8	56 06	45
"	"	9.2		30
"	"	11.4	54	29.8
"	WSW	10.5	53	30
"	"	10.4	53	30
"	"	11.1	52 51	35
"	SWXW	12.9		25
"	"	10.8	51	31
"	SWXS	14.7	49	30 25
"	"	13.2	48	
"	SXW	1.0	46	30

NXE	SWXN	8.7	"	45
"	"	1.8	56	45
"	"	9.2	55	30
"	NORTH	11.4	54	"
"	NORTH	10.5	53	80
"	"	10.4	53	00
"	S	11.1	52	35
"	SWXN	12.9	51	25
"	SXE	10.8	51	31
"	NWXN	14.7	49	00
"	ESE	05.2	14.8	25
"	WXS	1.0	46	30

SXW	SXW	81	74	85
"	"	1.8	72	30
"	"	09	06	75
"	"	11.1	22	58
"	WSW	6.2	49	03
"	"	11.4	53	30
"	"	1.2	52	35
"	SWXW	00	51	22
"	"	24	"	50
"	SWXS	47	00	30
"	"	10.6	28	25
"	SXW	05	41	"

SXW	SSW	13.5	"	30
"	NWXN	13.8	43	30.3
"	SXW	14.6	41	30
"	NE	15.3	40	30
"	SXE	15.3	38	31
"	SSE	16.2	36	31
"	EAST	18.0	34	32
"	"	72	31	33
"	"	21.3	28	33
"	SXE	21	21	33
"	EUE	25.2	23	34
"	S	26.1	21	34

SXW	SSW	13.5	86	30
"	S	13.8	43	30.3
"	SXE	14.6	41	30
"	NE	03	40	25
"	SXE	15.3	38	31
"	SE	16.2	36	31
"	NORTH	18.2	34	32
"	"	42	31	33
"	S	21.3	28	33
"	SXE	21	26	33
"	EUE	25.2	23	34
"	S	26.1	21	34

SEXS	SSW	12.5	"	01
"	NWXN	5.2	47.42	22
"	SXE	25	55	29
"	ENE	75.2	40	52
"	SXE	22	38	35
"	SSE	76	36	82
"	EAST	7.2	34	32
"	"	1.2	31	50
"	"	5.2	28	32.2
"	SXE	21	21	51
"	"	25.2	23	54
"	S	21.7	21	32

Figure 2. (left) Jeannette logbook page with hourly observations from the early part of the voyage. (right) WesTech transcription. The first set of columns shows the Best result, and the subsequent sets show alternatives based on two of the grid cell alignments tested automatically. The colors show the confidence in the recovery, with lighter reds indicating less confidence.

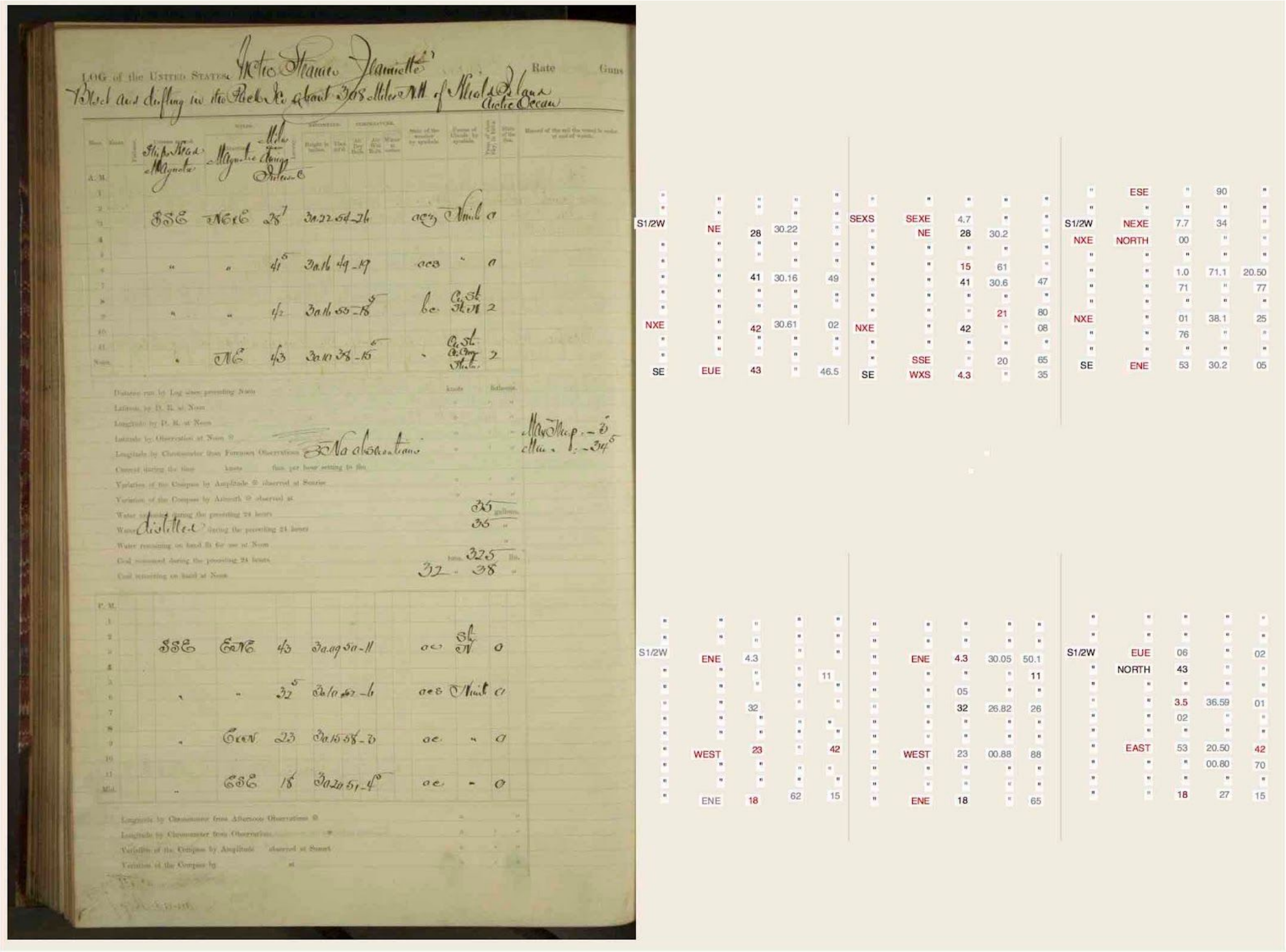


Figure 3. (left) Jeannette logbook page with hourly observations from the early part of the voyage. (right) WesTech transcription. Colors and format as in Fig. 2.

Appendix: Specification document, cost estimates, and technical workflow

At the start of the project, a specification document was drawn up to help guide the work plan between the participating groups. This document and the technical details and status of the WesTech experience are included below.

A. Specification document

In the auto-digitization project with University of Colorado, NOAA, UK Met Office, and University of Washington, we hope to achieve the following: we are looking forward to working with WesTech to help determine what are the effective workflows for auto-digitization. We are happy to help on areas below in which we have some developing expertise and trials.

Overarching question: ‘Suppose I have 1 million images (300dpi jpegs) containing tabular weather data in a mixture of formats, how do I get the data transcribed?’

1) Is there some turnkey software system that can do the job?

    2) Can we build our own system from existing components?
    3) Can we use software to do part of the job?
    4) Do we still need to do everything manually?

I. Target by document types:

1) Printed tables (e.g. Indian Daily Weather Report, IDWR)
2) Typescript logs (e.g. Farragut, Pennsylvania)
3) Manuscript logs (e.g. Jeannette, Rodgers)

II. Target parameters:

1) Pressure
2) Temperature
3) Wind speed
4) Wind direction
5) Date/time
6) Ship location

III. Task components:

1) Document analysis: Can we find the numbers - given a page image, can we find the location of the pressure reading supplied from one or more of
    a) WesTech analysis
    b) Manual gridding (U. of Washington, Kevin Wood)
    c) OpenCV image analysis (Oldweather volunteer, Uli)

2) Handwriting recognition: Can we read numbers, e.g., given a sub-image containing just a pressure reading, with what accuracy can the software transcribe it?

What do we need to report on:
How successful was either of the two task components?
What solution technologies did we test?
What other possibilities are there we could consider in future?

Our ambition is to find a workflow that transcribes all the listed parameters II above, from each page, with 95% accuracy (human check of sample output agrees this often), for a cost of less than 1c/image, with an achievable throughput of at least 10,000 images/day. [Note, **ambition** - not requirement for this project - the idea is that if we achieve this we can stop worrying about how to do it, workflows that fail to reach this target may also be useful].

For each document type I, we are looking to WesTech's expertise to help determine what is the most effective workflow. How does it compare with the ambition, and if it falls short, what might we do in the future to improve it.

The output of the project should be a report [due 31 March 2017], answering the question above; with quantitative rates for accuracy, speed, and cost per parameter & document type; and accompanied by evidence and examples.

Test images of (I):

https://www.dropbox.com/sh/b9qyk2ozhc8dizm/AABzXHmY3vvN-VAb2gcfL4cBa

B. Cost Estimates prepared by WesTech eSolutions

Ambition –

Machine Accuracy – our expectations, based on the Proof of Concept, are that the raw recognition accuracy has been determined. Some of the techniques used to improve the recognition accuracy will be including vocabularies of known values. We also believe by using volunteer keyers the accuracy will definitely approach 90% or better.

Software Cost –

The following assumes that NOAA, CIRES, or some other third party provides the processing environment. Additional fees will be required, if WesTech provides a hosted solution.

Extracting data from 1M images at a rate of 10K/day it will take 100 days to process. The assumption here is that the environment will support the required processing, memory and throughput. The software is not limited [can process more than 10K/day] to a set throughput. But, the limitation will be based on the availability of the volunteer keyers. The cost per image of the software will be at a rate of $0.054/image or approximately $54K/year including maintenance. Considering that NOAA, CIRES, JISAO, or the UK Met Office may have tens of Millions of documents of this type, it’s not inconceivable to process at least 2M documents/year at a rate of $0.051 or approximately $101K/year.

Professional Services –

As a guide the average additional Professional Services costs to add additional documents is estimated below:

Image Preprocessing as required Avg. – 14 hours/document type - $2,800

Custom Scripting as required Avg. – 15 hours/document type - $3,000

Create a new Form Definition with 2 templates Avg. – 30 hours - $9,000

To add new templates to the current Form Definitions Avg. – 2.5 hours - $500/Template

C. Technical details of the Workflow from the Proof of Concept prepared by WesTech eSolutions.

Pre-Processing:

The primary images conversion (using Kofax Express) from the existing (color / large resolution / size) format to Parascript's FormXtra recognition engine acceptable (Black & White / reduced resolution / size) format was achieved by processing those images through multiple iteration, to find the appropriate / allowable format. Each Document type was processed using various pre-processing techniques to achieve the best (and suitable) settings. When we reached the optimum settings, all the images in that Document type was processed in batches for maximum efficiency and speed. Now moving forward, any & all of those Document types could apply the same pre-process settings / technique and processed in sizable batches.

The hours spent for pre-processing the following Document Types…

Jeannette & Rodgers :	15 hours
Farragut & Pennsylvania:	10 hours
IDWR-1888,1910,1931:	18 hours

Form Definition / Template building:

Once the images were pre-processed and ready for recognition, they were brought into Form Definition studio for image (template) registration and field (metadata) recognition. Each Document type / group was build on it's own Form Definition project, for greater data accuracy. This is the most time consuming aspect of this whole process, to achieve independent results from the Document (image) set. Each Project (Jeannette, Rodgers etc.,) had multiple templates for registration, as not all of the scanned images were identical in size / shape / exact orientation. All the templates were deskewed / despeckled and primed from final registration, using appropriate Anchor zones. Once the templates were ready, all the required fields (Pressure / Temperature etc.,) were identified in the templates and zoned in, using Parascript's Table field matrix format. Additional Field types (Alpha, Numeric, date etc.,) were introduced on these table matrix to provide supplemental help to the recognition engine, to make right decision. Alpha data types (Direction) were provided with Vocabulary coverage and Numeric data types was supported with value range (for whole numbers), which helped the engine recognize the values with greater accuracy.

As the table structure in the images were not defined (absolute straight lines for column/rows), we had to introduce multiple table structures, overlaying on top of each other, but with only a slight alteration of their coordinates (one with the best possible layout and two others - one slightly above the primary table structure and second slightly below the primary) to get maximum recognition area for the individual cells. This approach was deployed to see if any cell data was outside the original (primary) table structure. The goal was to see if any of these three table (cells) would catch the required data, then use some special custom scripting to find which of these (3) cells had the absolute demarcation of the objective data (to be recognized). By this, there were 3 probable values for each cells and we picked the one with highest (engine) confidence.

The hours spent for template building for the following Document Types…

Jeannette :	25 hours
Farragut:	20 hours
IDWR-1888:	30 hours

Custom Scripting:

Once these multi-table matrix data were recognized, they needed to be validated for highest confidence and pick the best candidate picked for release. Also, the release has multi formatted - CSV (comma separated), Raw Text and an XML format (with all the metadata provided - position, confidence, all available recognition candidates). With the IDWR Document type, as the images were scanned (captured) in a book format (right and left pages), they had center of page indentation (the data structure curved in the center). When these images were brought into Form Definition Solution, they were not aligning due to the curvature in the middle. We had to split the image in the middle (additional pre-processing) as left and right pages, so that when we ingest them, they can deskew and align as horizontally as possible. This helped in building the table matrix with reasonable cell recognition. We had to use additional custom scripting to validate the width of the whole page (both Left & right) and slice them up appropriately.

The hours spent for custom scripting the following Document Types…

Jeannette :	10 hours
Farragut:	10 hours
IDWR-1888:	25 hours

Data Accuracy & Field confidence Threshold:

Every field (cell) data will be accompanied with an engine recognized confidence value. This will vary by the data clarity, placement within a cell area as well as the character formation (handprint data are the most difficult to recognize). This value could be in a range from 0 to 100, where 0 is with the lowest confidence (the engine is not sure the recognized data is accurate) and 100 being the best result. We need to arrive at a threshold for each field, where we (NOAA/CIRES & WesTech) decide the best threshold number for the best result. Any field confidence value below that threshold will be presented to a Keyer, who can visually see the field (in a web browser based keying tool) and make appropriate judgement call and key the respective values. Higher the threshold, possibility of more fields going to the keyer and high accuracy in data release. Lower will be the other way around. To arrive at an optimum threshold, we need to process a lot of sample data and compare it with the 'truth data' (which is the real data, maybe eyeball verification or someone manually key in the real data). This would help us attain the best threshold rate for the best accuracy.

Environment:

All the images could be processed in the WesTech Cloud environment or a NOAA /CIRES specified environment. The total volume process per day depends on the CPU throughput and memory. In a nutshell, we can process images in multiple batches, and send any fields under certain threshold to a Keying queue, where they will be available for Volunteer Keyers (from anywhere in the world - as they are using a web based keying tool) to key these fields at their own schedule. There is no Software limitation for processing the volume, but only Physical. If we have enough Keyers to help, our estimate is to process 1M images in a month's time.

Conclusion:

When moving to a full production cycle, all the pre-processing and the Form Definitions already created will be used for all the subsequent images (batches), greatly reducing any repetitive work. Occasionally, we might have to introduce additional templates (due to new, unseen format of images), tune the table matrix format (cell orientation) accordingly. Additional, if there are any specific release format required, we might have to do some custom scripting.

Estimate:

Pre-processing:

All the new images which we need to extract data have to be pre-processed (using existing settings in Kofax Express) in batch mode. This could take approximately 1 hour for about 10K images. As these are set in batch mode, we can process them as scheduled task, in non-business hours.

New template (each):

As mentioned above, if we come across any new image formats, which we are yet to see, we need manual intervention to add them into the Form Definition templates, and an approximate hours to process them (including custom scripting)

Jeannette / Rodgers Document type:	2 hours
Farragut / Pennsylvania Document type:	2 hours
IDWR-1888,1910,1931 Document type:	3 hours

Environment: To Be Determined

****

There is no requirement for additional product or development once the recognition engine is tuned, the confidence values and thresholds are validated. After that point the overall accuracy is in hands of manual keyers.