NTDS 043

Key:

I: Interviewer
R: Respondent

I: Okay. Great. So the way I usually start the interviews is just to ask you to introduce yourself and your background and your role at the Met Office with the purpose of talking about your involvement in MEDMI, how you got in, from what and stuff.

R: Absolutely. So my name is Rachel McInnes. I've worked at the Met Office for five or six years now. I work in Climate Science and specifically look at climate impacts, so the impact that the climate has on infrastructure or customers, people, any sort of impact, whether that's health or agriculture on people. Previously my background was in physics and astronomy and I have a PhD in astrophysics. The things that link I think the physics, the astrophysics and the climate science are the data. It's big data and it's data analysis. And these are the skills that I feel have joined all of my research together.

I: Right. Specifically what kind of skills? Do you want to open this up? It's very interesting.

R: So some of it is, when I say big data, I mean, it doesn't matter what the data are, but if you have a large, large volume of data you have some computational problems with how you are going to access it, process it, store it, share it. And whether that data has come from a telescope or whether that data has come from a climate model, you've got some of these big data problems about how you handle just vast terabytes of data. And then when I say data analysis, I'm talking about scientific programming, so what I would say was coding, so whether that's coding in a few different scientific languages. But even though the different disciplines have maybe a language or a few languages that they use, the general skills in how you code in one language are quite applicable about how you would code in another language.

I: Okay. What do you mean by coding here? Because I think usually most of the times I talk about coding data in medicine is like applying a terminology, like a cultural vocabulary to say a report, something to sort of map this is an instance of x.

R: Okay. I think that's quite different from what I mean, then. So what I mean is using a scientific programming language, so, for example, Python or R, IDL, MATLAB. One of these languages, which allows you to write lines of code, and that might be hundreds of lines of code where, if we take the example of a climate model, so data that's come out of a climate simulation, that will typically be gridded, so it will have a value in each… say, each grid square on the globe might be 1km and in that 1km you'll have a value for temperature, a value for rainfall, a value for humidity, you will have those values at different height levels as well in the atmosphere and for every different day whether your times dep is hourly or daily or monthly.

I: Yes.

R: So you've got this sort of multidimensional dataset. So when I say coding, I want to read in the files that are associated with this climate model. And that might be a number of files. There might be a file per day or a file per variable, whether it's temperature or rain. I want to read in those data and then I want to use the code to narrow down what it is I want to look at, so maybe I write a code that says read in this data and then I want you to select Europe from this global model and so I write down in the code the criteria that I want to select, whether that's latitude and longitude, to kind of select a small patch or whether I only want to know about data from 2000 until 2010, so I write down in my code I want to select this chunk in time and then maybe I'm only interested in humidity at a certain height above the surface, so I write in the code I want to select this height and then maybe I want to create some statistics over Europe for this timescale that I've selected and at this variable and at this height I've selected. Maybe I want to look at the mean as a very simple one to start with, I want to look at the average over this whole area, over this whole time, and for this one variable, then I maybe want to look at other statistics, quantiles, percentiles, standard deviations of this variable and then… So all of this I would code, I would write down in a script, in a code that tells me so that I can repeat this easily, it's traceable. The computer does the computation for me and I can share that code with someone else if they want to do something very similar. Or if I want to do something very similar but next time instead of humidity I want to look at precipitation, then it's very easy for me just to adjust one line in that code and I repeat the same analysis. Does that give you a sort of idea of what I mean?

I: Absolutely. Super interesting, yes. It's coding much more in the sense of computer science and programming.

R: Yes.

I: Analysis, yes.

R: I guess we often call it scientific programming because I think it is slightly different, the types of things we do than someone who is purely doing IT. And often we are not trained, we are not computer programmers [laughs], we've learnt on the job, and so I think we are much poorer at coding than software engineers or people who have been trained specifically in coding. But we use it as a tool and it's a really vital tool for us in our work.

I: I'm sure you probably have other advantages, like you really know the code well from a more sort of scientific point of view.

R: Yes. We can use it as a tool to make sure it does what we want it to do but maybe it's not the quickest code or it's not the most sophisticated. If someone looked at it purely from a code point of view, I suspect they would be able to find [laughs] problems with it or better ways it could be written. But we are happy it does what we want it to do.

I: Right. Yes. You mentioned this, there's no education around this, so how did you learn?

R: People tend to start learning at university, I would say, in my experience, certainly in physics. It's probably different in different disciplines. So I think my first introduction to it was in my undergraduate degree in a kind of Master's project, I started to use a programming language called MATLAB and this was just shown to us as a tool we could use. And maybe in some of my… yes, some kind of labs. So as well as practical experiments in the undergraduate labs, we did some coding, so we were introduced to it then. And then during my PhD was really when I had to become much more proficient at using it because that's how, in the field I was in, that's how you do all your analysis, all your work was done. And again, I used MATLAB for that and IDL, I think. So I think maybe you get a course, maybe you can go to a one-day or an afternoon and you get shown how to do things. I think mostly what you get is given someone else's code and a postdoc will help you or your supervisor will say, 'Adjust this,' so you get a piece of code that works, or hopefully works, and you learn to adapt that. Or you write very, very simple things from scratch using online tutorials, things like that. So there are courses. And during my PhD and since I've joined the Met Office I'm sure I have been on a few programming language specific courses but they tend to be a day or two days. I don't think you can learn how to code in a couple of days, they just give you a taster and then you know the type of thing you can do in that language and then go away and sort of try it out and learn from people around about you and online tutorials and things like that.

I: Right. Very interesting. So going to MEDMI, what are you doing with MEDMI?

R: So, I worked with Christophe, who I believe you've met, on a MEDMI pilot, so I was only involved in this pilot rather than the main MEDMI programme and our pilot was looking at pollen, so allergenic pollen that comes from plants, and people… I'm always thinking of the impacts [laughs]. So the impact of pollen is on people's health, they breathe in the pollen, and if they suffer from hay fever or asthma then it can aggravate these conditions and lead to a range of respiratory problems. So we were looking at pollen and linking the vegetation, so the plants that produce the pollen to health, or wanting to make tools in our pilot that would allow people to link the vegetation data with health data and we focused on a few key allergenic plant species that are found in the UK and that are commonly linked to health problems - hay fever and things. So we look at some weeds and we looked at grass. Grass is the main one that people in the UK are most allergic to.

I: Yes.

R: So I can tell you a bit more about my contribution to that.

I: Yes. And also the project, if you want. Yes.

R: Yes. I mean, it was a very short project. It was a few weeks that we did some work on this. It was quite a small project. It linked very closely with another project that I had been doing with Laura and other colleagues at the European Centre and also colleagues at Public Health England and London School of Hygiene and Tropical Medicine, which is called the Health Protection Research Unit, the HPRU, and in one part of that we've done a sort of two years so far, looking at allergenic pollen and mapping where these species are in the UK. So it linked very closely to that and it aimed to sort of take it just a little step further forward so that other researchers could try and link with our vegetation mapping data and kind of prove.. the pilot itself was a kind of proof of concept that you could just take this further. So it isn't a fully working finished… now you can do everything you would want to do with pollen and MEDMI, but we looked at how you would… what you would do for the next step and how you would link the vegetation mapping and the pollen to the health data. So what we did provide is maps. And by maps I mean gridded data over the UK that shows the abundance of different plant species. So I had different maps for each different species. I think for MEDMI we did six species, but you would need to double check that. Grass and five weeds, I think. So if we take grass as the example, then, at every location in this grid box in the UK it would tell you what percentage cover in that grid box was grass. So no grass it's 0%, so completely grass 100%, or somewhere in the middle, how much coverage in this grid box. And we did this for the different species.

I: Of grass.

R: Yes. I use the word species probably slightly incorrectly. So grass we just looked at as one collection of grass. So you are right, that's not one species, that's a group of species. But it's very difficult with pollen to distinguish the different grass species, so we look at grass together. But the five weeds, there were five different dock and nettle and ragweed and things like that. And again, they might technically not be species; some of them might be genera. I'm not good at biology [laughs]. I'm learning on this. So kind of vegetations types, we could call them, or different types of plant.

So we've got these maps. And I think one of the tools that Christophe did in the MEDMI… we worked together on the pilot. One of the tools that Christophe did means that someone who is now using MEDMI could access these maps, but if they are just looking at a specific location or they just want to know the coverage, maybe not in the format that I have produced it in, which would be gridded data on a particular resolution space or resolution, I think they can now go in as they can with the other variables and MEDMI and pull out the location they want or, for other variables, the time period they want. Ours has not got time data on it. So they should be able to pull out the coverage in a particular area or a particular resolution that they are interested in, which I hope would be more useful than…

I: Sorry, if I make you repeat.

R: No, no.

I: So it's about pullout, the coverage, a data for selections of the data that are different from the selections that you use to produce a map. Is that correct?

R: Kind of. So this wouldn't help with your recording, but if I draw this.

I: Yes.

R: So if I have a patch of the UK, if we imagine this is the South West, and then I have gridded the UK. A lot of my data that I produce is gridded because I can't provide the exact value every single, you know, tiny resolution, so we provide gridded data. But I'm aware that that's not how health data often is or how other researchers might want their data. So if we imagine this is Exeter here. Now, depending on the size of this grid box, it might be 1km by 1km in size, and in that square kilometre my map might say the value here is 58%

coverage of grass and there will be another map that tells you about dock and another map that tells you about ragweed or the other things. But if we just imagine that this is a grass one. And the neighbouring grid box here might have a value of 60 and the one down here might have a value of 2%, whatever. So this, to me, is how I have made this map and how I've been able to be happy with the data that I provide. But someone else might want to know what's the value at a point. So not in the whole grid box, not in the 1km surrounding area, they might have data at a point, so either that's an observation site – they've physically got at one specific location something that's measuring rainfall or something that's measuring, probably not a health impact but another environmental variable, and that's being produced at just one point.

I: Right, yes.

R: And then I guess the other thing with health is that often… Oh here is Neil. Hi Neil.

I: Hello.

N: Hi there.

I: Niccolo.

N: I think I've met you before, haven't I, or is that my imagination?

I: No. I'm not sure. Yes.

N: Okay.

R: We are just running. We started a little bit late. So do you know how much longer, just so that Neil is not hanging around?

I: Yes, maybe 20 minutes or something. It depends also on you, how much time [overspeaking]?

R: Yes. I need to go shortly after 12. But yes, 20 minutes.

N: Okay. Well I can come back.

R: Is that okay, Neil?

N: Yes. At about five past, ten past twelve?

R: Yes, that would be super. Thanks, Neil. Sorry, we've been running a bit late.

And someone looking at health data might have health data for a hospital and the hospital might be in Exeter but the catchment for the hospital could be much larger, so they might have number of asthma admissions per day. And so I guess what I mean about people being able to access the data how they want, with MEDMI, I think the way it works is that they could use this area and they can ask for what is the average grass coverage in that area or they can pull out the data just in this area or they can ask at a specific point what's the best value at that point? So they don't need to want the data on a 1km by 1km gridded… they don't need it in that format or at the sort of spatial resolution, they might want it in a different way to match with their data. And often with health data it's not nice neat grids, it's… we call them polygons, but I don't know what the health people call them. It's a shape, it's a funny shape, and it might outline a postcode area or it might outline the area that a GP practice covers or something like that. But hopefully MEDMI will allow them to pull out the values in the area they are interested in and not worry that I've originally made it on a nice neat grid, they can hopefully put in their requirements and the MEDMI itself will kind of spit it out.

I: Right. So is it a way to get basically behind the link on the grid and sort of get to the data in the way the data has been generated?

R: It doesn't help them kind of go behind the data because…

I: Okay. So this point data would then be derived data, it's not original data?

R: Yes. That's correct.

I: Okay.

R: But if the point is… or if they ask for what's the grid here and it overlaps all my grid squares, it should do the working for you and say, well it's got some contribution from here and some contribution from here and some from here and some from here and some from here and I will give you this one number.

I: Fantastic.

R: Does that make sense?

I: Yes.

R: And that's all that researcher wants to know. They don't want to get my grid and have to manually work it out for themselves. So I think MEDMI should…

I: Right, of course. It could be also extremely challenging for if you don't have the right skills.

R: Exactly. And if you are not used to gridded data or big… I mean, I'm showing you 1km, but some of my maps were 25m by 25m. So this is quite small areas. So if you were looking at a big area like this, suddenly you've got quite a lot of grid squares to deal with, and if that's not what you are used to doing and you just want the answer, then hopefully some of the tools in MEDMI will allow you to get a sensible answer without you having to know how to get it.

I: Yes. A trivial question about this, I was wondering sort of this grass coverage and this weeds coverage, how is that calculated? Do you use satellite images?

R: Yes. It's not a trivial question at all. That's been a big part of our research, has been making these, and it's not simple to make. So you can get lots of… they tend to be called land cover maps and they often use satellite data or what's called remote sensing data and they are essentially maps and they are gridded, like I've kind of described to you. This square of land, this parcel of land is forest or this patch is a river and this patch is urban, this patch is suburb and so these land cover maps split the UK into land types, and for different products, different land cover maps, there will be different classifications for what they are. So we have used those, but they don't tell you… so probably the example of trees or something is easier to understand. The land cover maps will tell me where forests are. Some of them might tell me where different types of forests are, so are they conifer forests or broadleaf forests. You know, do they mostly have pine trees or just the… we call them broadleaf, but oaks and birch trees and things like that. And certainly I use Forestry Commission data as well which will tell me very detailed information on where the forests are and whether they tend to be these conifer trees or the broadleaf trees. But none of them will tell me where are the oak trees, where are the birch trees, which is what I want. I want to be able to say not just that's a forest, I want to say how many of the trees are allergenic, and oak might be one of them. So as well as using these maps that tell me where land types are, the type of vegetation I'm looking for, there's different ways where I use other information, which is sometimes from the Forestry Commission, if it's trees. Or for the weeds I use the panel of experts that knew about the type of habitats that the weeds would thrive in.

For the grass, we just looked at grass as a whole, so grass was more simple, we weren't trying to pick out specific species of grass. So that was easier to use, these land cover maps, but then I used a combination of these different types of information, so some about the species, the proportion of different species that would be found in different regions, so the proportion of, say, pine in a conifer forest in the South West of England, maybe that's – now I'm just guessing here – but maybe 60% of the conifer forests are pine. But in Northern Scotland probably the species that exist will be quite different, so maybe pine only makes up 15% of the conifer forests and actually there's large or another fir that makes up the most. So these species proportions were original. So I've combined these different types of data together to produce what we are kind of calling the species map or the vegetation maps at the end.

I: Okay. So you have associated sort of another kind of linkage, isn't it, linking characteristics or certain types of forest with these location maps that tell you sort of… in order to basically enhance [overspeaking].

R: Yes, I think so. I think so. And because we provide them on a gridded basis, I feel more confident about the result that's in it. It wouldn't be accurate for me to say this specific tree in this location is oak. I don't know that without going to look at that tree, but I can say in a 1km grid box of these trees in this area I believe there is 2% coverage of oak here in this bigger area. And that, to me, I can be more confident about that. Scientifically, it's a more robust result, so I'm sort of more confident in being able to say that, although there will obviously be associated kind of areas in the region about how these estimates have been made.

I: Yes. And all this work was pre-existing to MEDMI, obviously. It's something that you have worked on for a while.

R: Yes. Mostly through this Health Protection Research Unit that I mentioned.

I: Okay, yes.

R: Yes.

I: It's for the HPRU?

R: Yes.

I: Okay. So it's relatively new anyway.

R: Yes.

I: What I was asking was it's not an asset of the Met Office that has been existing for five years before or something?

R: No. The mapping is work that's being done… I've just completed under the HPRU, but that's been two years in us doing that.

I: Okay.

R: One thing I'll just quickly explain. So when you are thinking about the health impacts of pollen, so you've now got an idea of where the grass mostly are in a patch of the UK that you are interested in and you are looking at your health impacts, whether that's hospital admissions for asthma or GP consultations for hay fever, whatever your health impact is that you are looking at, and you want to associate it with pollen, is there one particular species of pollen that's causing this for most people or a combination of… or are all the pollens equally allergenic? So you want to assess that as a sort of health researcher. But if you think of the actual plants with their pollen and they are releasing their pollen, when the plant flowers and the pollen is released by the wind, of course the pollen doesn't stay where the plant is, the pollen is carried in the air by the wind. And so with this MEDMI pilot project what we looked at, as well as having the maps to say this is where the pollen starts, we looked at how could we do a very simple model of where the pollen will end up and where will the concentrations of pollen then be highest, where will these pollen go? And because the MEDMI is linking with the environmental data, so the meteorological data as well, the wind, the wind speeds, the wind direction on different days and also… I'd say there's hundreds of environmental factors that affect pollen and when it's released and where it goes, but the big, big drivers are wind speed and wind direction because they are just very light particles and they'll be carried with the wind but also rainfall because the minute it rains the pollen is brought to the ground and you can't breathe if it's wet and on the ground, so it needs to be in the air for you or I to breathe it into our respiratory tract. So we looked at, as a sort of proof of concept, how would we, in MEDMI, code, yes, write a programme that would take at a particular location using the species maps, using a simple model of when the pollen might be released. Because that's another thing of course, we know that grass doesn't flower all year round, a tree doesn't flower all year round, so we would have some simple model of when the pollen would be released and then a model of where that pollen is most likely to go and that would be an estimate. Obviously, it's not going to travel in nice neat bundles, but the predominant amount of the pollen, where is it going to end up, and we believe that there will a big enough signal coming from where most of the pollen goes that you would see that, I suspect, in the health data as well. So it was kind of a proof of concept. We haven't completed all the codes. We started to look at it. We started to look at how you would code it up, what kind of… if your pollen was… if you looked at pollen that started here, from this space here and you knew that the wind direction was in this direction on a particular day and you knew it was going at 20m per second, you've got a direction and you've got a speed, where does this pollen go? So we looked at what shape would this… we call it a plume.

I: Okay. A plume?

R: A plume, like a plume of smoke. So Christophe and I in the pilot looked at what shape is this plume and that plume will be three dimensional, the pollen will rise and fall and where will most… maybe most of the pollen will be in the centre of that. So we kind of looked at what mathematical model we can use to define this plume, how does the wind speed and the wind direction… so you can tell that I've aligned the plume with the direction but how does the… if this wind was going faster, how does the shape of this change? And so that's what Christophe and I looked at, what would be a simple model? And we have very complex models here that do this with our weather simulations, we have what's called a dispersion model which will do very highly accurate modelling of the atmosphere and all the winds and things, but that's not what we need for MEDMI, that's too complicated, it's very computationally expensive to run. So what we are looking for here is a simple model - can we describe where most of the pollen will end up so that we know where the highest impact of people being able to breathe in this pollen would be and then, of course, because this is gridded data, you want to repeat this for every point and then end up with a new map. And the new map will not be where are the plants; the new map will be where is the pollen in the air and at a height that you or I could breathe it in. So does that make sense?

I: Yes, absolutely.

R: Okay. So that's what we looked at. But as the kind of proof of concept we tested some things out, we tested some different models that changed the sort of shape of this and we looked at how you could… what you would need to take into account and how you would take that forward. But the benefit of MEDMI is that the wind data and the rainfall data are already in MEDMI.

I: Yes, the pilot dataset.

R: Exactly. So the researchers wanting to use this already have all the data they need. What they need is, I guess, a tool, a model that will join it up for them.

I: So working on the plume, the shape and stuff, so where does all this information or model about the shape… sort of is that something you produce a doc from this or for this, or is this part of the library of shapes and calculations that you can take from elsewhere?

R: Somewhere in between, I think. So there's not a specific if you are doing pollen, this is how you want to model it with a plume, but equally there's a lot of studies that have done similar things or have looked at plumes of other particles which are similar to pollen and so we could start with those and then look to adapt them.

I: So like pollutant particles? Or what kind of particles [overspeaking]?

R: I think so. Or sometimes it's just very theoretical, so it's just people talking about particle dispersion for a particle of size this and buoyancy this.

Sometimes it's very theoretical and it makes too many oversimplifications that are not very applicable in your world or sometimes it's for a very specific other case, someone looking at the dispersion of something else. So I looked in the literature, in the published literature, to read about this and also spoke with experts here on dispersion. So yes, they are looking at much more… Oh I think this is room is needed; someone is trying to come in.

I: Do we have a minute to close?

R: I think we've got one minute, yes, and then we can just sit on the balcony till Neil comes. I can't remember what I was saying. So I also spoke to the experts that do the much more complex detailed dispersion modelling about the problem that I was doing for MEDMI. We've got lots of people that do atmospheric dispersion in the Met Office, so I also spoke to them to get their advice on what we would want to look at it. But a bit of it from what they advise would be trial and error, we'd be trying it out, testing it, seeing what results that produced, is that realistic and maybe before you use the model you would need to test that it worked, so you would need to use real data of pollen monitoring, so observations of pollen, compare them with their maps - so we run our model, we think you would observe this much pollen at this location and then compare it to what actually happened, compare and then you could refine your model further.

I: Yes. You have done this, or is this something you will do?

R: No. We just kind of produce the… we are kind of calling it a proof of concept but showing this is what you would now need to do. We've done this much, this is where we think you would need to go next and these are the steps you would need to sort of test your model. Does that make sense? I'd love to finish it. I'd love to do all the work, but that's a much more bigger project.

I: Yes, absolutely. You want to move?

R: Shall we sit outside? Yes, thanks.

<u>NTDS 043 2</u>

I: Okay. And so this was a pilot project. Was this something that you started lately, or what point in time of MEDMI is this?

R: We did this - if I remember correctly - October/November 2015.

I: Okay. Recent.

R: Yes.

I: So how does this work relate to the other sort of work? Because it's very complex and there's lots of work, programming and stuff. So I would like to ask you more about how does it relate to all the other kinds of things that happen in the Met. For example, you said obviously we write scripts and the scripts can be reused by myself or shared. So I was wondering more about the larger practices in the Met Office, if there are libraries or groups where people share and stuff.

R: Yes. I think there is quite a range of work that goes on in the Met Office. If you speak to someone in different bits of the office I think it will be quite different. I can only talk about the science programmes, so not the forecasting. So if I just talk about science because that's what I know about. I suspect there will be slightly different practices in other bits of the office. So within the science, there's weather science and climate science and then there's the more applied impacts, so your customer facing kind of application of the science. There's kind of core research that will go on for a long time and then there's more customer facing… more like consultancy projects where you are doing a particular project for a particular customer who is interested in a specific problem. I think we all share the same skills in analysing data and looking at data. Yes, the datasets will be different, but we still need to be able to look at often observational data, modelled climate data that have come out of our models and then when we are talking about the applied work as well, also some skills in understanding what it is that the customer actually is interested in, what specifically about this project applies to them and making sure that the thing that you are doing for them, the research that you are doing for them is really what they need. Does that make sense?

I: Yes.

R: There's a number of tools we use for… we are calling it quality assurance here in terms of our coding. So there's always been a great deal of reviewing and checking and peer reviewing in terms of reports, any written work and research that goes out. Certainly, in terms of the science, there will be, for a big project, you know, a science review board and experts that are always reviewing for the robustness of the science and of the reports that go out but we now have also quality assurance practices that we are using across science that look at the actual code that we are writing to make sure that that has been tested, it's been reviewed the same way you would review a report, you have someone review the code, that the code is backed up somewhere, you've got your two microphones there to record, you want to make sure that the code and all the data are backed up. So sometimes that's a case of saving it twice in terms of data. But in terms of the code, often… we'll just be one minute, if that's alright.

N: Okay. That's fine.

R: So often the code, if we want to share it with someone, we might want to share previous versions of it, so we use software that let's us version control. I don't know if you've heard of version control.

I: Yes.

R: So we version control our code and the software across the office that we use that's supported to help scientists learn to do that, so you've got that, well, as well as if I make a mistake and I break my code one afternoon [laughs], I can go back to the previous version that worked. So it helps with human error. But I think also, fortunately, it provides like an audit trail of the research, how it's evolved, if anyone queries, 'I've published this paper and it has a graph in it,' and you want to know where it's come from, I can show you exactly where that's come from, how it made that graph and someone can repeat it and they can go and get that exact bit of code and reproduce it. So that's one example of the quality assurance, but there's lots of practices in the office to sort of test and review and document code.

I: And these are sort of obviously across projects.

R: People have different ways of implementing them, but the theory is the same. Maybe you use a slightly different bit of software to do your version controlling, but yes.

I: Okay. One question about the MEDMI, how does this fall in this division of the science programmes?

R: Where does it fit?

I: Was that more like a customer facing…?

N: Did you want to just come to my desk?

R: Yes, sorry. I do need to go actually now, so probably if you could just…

N: Oh okay.

R: Is it okay if we just wrap up? Maybe you could email me if you have a specific question. Is that alright?

I: Yes.

R: Sorry.

N: Yes.

I: I'll let you go. You need to go. Thanks a lot.

(End of recording)