1.a

$$V_{opt}^{(t)}(s) = \max_{a \in Actions(s)} \sum_{s'} T(s,a,s')\left[Reward(s,a,s') + \gamma V_{opt}^{(t-1)}(s')\right]$$

$V_{opt}^{(0)}(-2) = 0$; $V_{opt}^{(0)}(-1) = 0$; $V_{opt}^{(0)}(0) = 0$;

$V_{opt}^{(0)}(1) = 0$; $V_{opt}^{(0)}(2) = 0$

Iteration 1:

$a = -1 \Rightarrow 80\%\,(20+0) + 20\%\,(-5+0) = 15$

$a = 1 \Rightarrow 70\%\,(20+0) + 30\%\,(-5+0) = 12.5$

$a = -1 \Rightarrow 80\%\,(-5+0) + 20\%\,(-5+0) = -5$

$a = 1 \Rightarrow 70\%\,(-5+0) + 30\%\,(-5+0) = -5$

$a = -1 \Rightarrow 80\%\,(-5+0) + 20\%\,(100+0) = 16$

$a = 1 \Rightarrow 70\%\,(-5+0) + 30\%\,(100+0) = 26.5$

$V_{opt}^{(1)}(-2) = 0$; $V_{opt}^{(1)}(-1) = 15$; $V_{opt}^{(1)}(0) = -5$;

$V_{opt}^{(1)}(1) = 26.5$; $V_{opt}^{(1)}(2) = 0$.

Iteration 2:

$a=1 \Rightarrow 70\%\,(20+0)+30\%\,(-5+(-5)) = 11$

$a=-1 \Rightarrow 80\%\,(20+0)+20\%\,(-5+(-5)) = 14$

$a=-1 \Rightarrow 80\%\,(-5+15)+20\%\,(-5+26.5) = 12.3$

$a=1 \Rightarrow 70\%\,(-5+15)+30\%\,(-5+26.5) = 13.45$

$a=-1 \Rightarrow 80\%\,(-5+(-5))+20\%\,(0+100) = 12$

$a=1 \Rightarrow 70\%\,(-5+(-5))+30\%\,(0+100) = 23$

$V_{opt}^{(2)}(-2) = 0$; $V_{opt}^{(2)}(-1) = 14$; $V_{opt}^{(2)}(0) = 13,45$;

$V_{opt}^{(2)}(1) = 23$; $V_{opt}^{(2)}(2) = 0$;

1.b.

The optimal policy $\pi_{opt} = \begin{cases} 0 \text{ if } isEnd(s) \\ \max_{a \in Actions} Q_{opt}(s,a) \end{cases}$

Resulting policy for non terminal states:

$\pi_{opt}(-1) = -1$; $\pi_{opt}(0) = 1$; $\pi_{opt}(1) = 1$.

2.a

We have define a new MDP with states $S' = S \cup \{o\}$, where o is a new state. Let's assume that the probability reaching that state from any other state is $1-\lambda$. The probability of not reaching this o state being in any other state is λ. So we have to adjust our transition probabilities by λ. It means that our transition probabilities $T'(s,a,s') = \lambda T(s,a,s')$ We also need to adjust rewards to $\underbrace{\text{rewards}(s,a,s')}_{\lambda}$, if we want to have the same optimal value. Thus having an MDP solver that can only handle with discount of 1, while solving for another discount factor $\gamma < 1$ we need to adjust probabilities and rewards like:

$$T'(s,a,s') = \begin{cases} (1-\lambda) & \text{if } s' = 0 \\ \lambda T(s,a,s') & \text{otherwise} \end{cases}$$

$$R'(s,a,s') = \begin{cases} 0 & \text{if } s = 0 \\ \frac{1}{\lambda} R(s,a,s') & \text{otherwise} \end{cases}$$

4 b.

On the small test case Q-learning shows great results. We've got only 2 states different out of 21 possible. We can assume that it can learn properly when to take, peek, quit. On the other hand the large test wasn't so successful, we have got 278 different out of 2705 states. It couldn't be able to learn properly when to quit, take or peek because it is quite easy to be over 40 with small multiplicity and big amount of cards.

4d

FixedAlgorithm gets its policy from original MDP with average reward of 6.78. But when we using Q-learning, average reward is 9.45. Thus it has better results because it can easily adapts to the new problem, adjusting parameters to new circumstances. While FixedAlgorithm will follow that fixed policy and couldn't adjust to the new problem.

5.a.

According to our MDP simulation we've got different results. For 40 year time horizon economically preferable would be holding off on investment and rearrange cities budget on other spheres, as ratio of investment to wait states is very low (0.0015). For 100 year period it would be opposite situation as our simulation shows almost 50% ratio on investment policy due to much higher probability of sea level rising.

5.6.

As to 50 years economic agenda our recommendations would be to invest in infrastructure. Taking into consideration not only simulated predictions on sea level rising, which is highly likely to occure in longer term period but also vulnerability of future generations their wellbeing depends on environmental decisions we are making today. Like with human health better to preserve then to cure.

5.6

Both strategies recommend to invest in infrastructure, with almost 40% ratio based on discounted model and 100% based on non-discounted. Discounted model which take into account devastating weather events shows us that our efforts will be less paid off due to certain devastation of the infrastructure, but we will receive a reward after each time step. On the other hand non-discounted model shows that our investments will succed.

5. d

Our original NDP predicted economic benefits despite negative flooding cost. As it still is a model it can't predict everything. But it can adjusted with discount factor which will allow our model to be more precise in its predictions in different scenarios.